Andreas Philipp Hassler, BSc

RISK FACTOR DISCOVERY AND MODEL DEVELOPMENT FOR FRAILTY PREDICTION

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Assoc. Prof. Andreas Holzinger, PhD, MSc, MPh, BEng, CEng, DipEd, MBCS

Institute of Interactive Systems and Data Science

Assoc. Prof. Ernestina Menasalvas Ruiz, PhD, BEng
Universidad Politecnica de Madrid

Graz, May 2017

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Master's Thesis

RISK FACTOR DISCOVERY AND MODEL DEVELOPMENT FOR FRAILTY PREDICTION

Andreas Philipp Hassler, BSc

INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE,
GRAZ UNIVERSITY OF TECHNOLOGY

Supervisor: Assoc. Prof. Andreas Holzinger, PhD, MSc, MPh, BEng, CEng, DipEd, MBCS

Co-Supervisor: Assoc. Prof. Ernestina Menasalvas Ruiz, PhD, BEng

Graz, May 2017

This page intentionally left blank

Masterarbeit

(Diese Arbeit ist in englischer Sprache verfasst)

FINDUNG VON RISIKOFAKTOREN UND ENTWICKLUNG EINES MODELLS FÜR DIE VORHERSAGE DES FRAILTY-SYNDROMS

Andreas Philipp Hassler, BSc

INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE ,
TECHNISCHE UNIVERSITÄT GRAZ

Betreuer: Univ.-Doz. Mag. phil. Mag. rer. nat. Dr. phil. Ing. Andreas Holzinger

Co-Betreuer: Univ.-Doz. Dr. Ernestina Menasalvas Ruiz

Graz, Mai 2017

This page intentionally left blank

Abstract

As people today live longer, there are more elderly people struggling with age related diseases. Therefore, healthy ageing becomes an important topic. This presents a challenging task towards establishing new approaches for maintaining health at a higher age. Such approaches would be beneficial on the one hand for the affected individuals themselves and on the other for avoiding a rapid increase in health and care costs.

A representative syndrome for the age related deterioration of the general condition of the patient is frailty. This syndrome is associated with a high risk for falls, disability, hospitalization and mortality (Fried et al., 2001).

In the Toledo Study for Healthy Aging (TSHA), medical data of adults with ages over 64 was collected. The data contains physical examination results, blood results and interview answers. For retrieving the latter, questions regarding health status, psychological status and cognitive status were asked.

Using predictive data mining given the data of this study makes it possible to derive a clinical decision support system, which provides the doctor with information on the probable clinical outcome of the patient. This vital information can be used to react promptly and avert likely adverse events. Also, potential frailty risk factors can be derived using sophisticated feature selection methods.

In this work, which is framed in an EIT-HEALTH financed EU project called FACET, a methodology for building a predictive model and retrieving potential predictors for the frailty syndrome has been presented. Further, the beneficial collaboration of the data scientist and the medical doctors, resulting in a better performing predictive model has been shown. Moreover, the importance of the data preprocessing has been demonstrated. Especially, the significance of dealing with missing values.

Nevertheless, in future work the findings have to be further analyzed and validated in bigger cohorts, with the objective of realizing a model, which can finally be deployed in the health care system.

Keywords

HEALTH, DATA MINING, MACHINE LEARNING, PREDICTIVE MODELLING, RISK FACTOR DISCOVERY, DATA PREPROCESSING, MISSING VALUE IMPUTATION

ÖSTAT Klassifikation

Information Systems (102015)

Machine Learning (102019)

Medical Informatics (102020)

ACM Klassifikation

Information systems: Data mining

Computing methodologies: Machine learning

Applied computing: Health informatics

This page intentionally left blank

Kurzfassung

Das zunehmende Älterwerden der Gesellschaft führt dazu, dass immer mehr Menschen unter altersbedingten Erkrankungen leiden. Aus diesem Grunde stellt gesundes Altern heutzutage ein topaktuelles Thema dar.

Dies birgt nun die Herausforderung, neue Ansätze zur Erhaltung der Gesundheit im höheren Alter zu finden. Solche würden einerseits den betroffenen Individuen, andererseits aber auch dem Gesundheits- und Pflegesystem zu Gute kommen.

Ein repräsentatives Krankheitsbild für den altersbezogenen, gesundheitlichen Verfall von Patienten stellt das Frailty-Syndrom dar. Dieses wird mit einem erhöhten Risiko für Stürze, Invalidität, Hospitalisierung und Mortalität assoziiert (Fried et al., 2001).

In der Toledo Study für Healthy Aging (TSHA) wurden medizinische Daten von Erwachsenen mit über 64 Jahren gesammelt. Diese Daten beinhalten Resultate der ärztlichen Untersuchungen, Blutwerte und Antworten von Befragungen. Für den Erhalt der Letztgenannten wurden Fragen bezüglich des Gesundheitsstatus, des psychologischen Zustandes und solche zur Testung der kognitiven Leistungsfähigkeit gestellt. Unter der Verwendung von prädiktivem Datamining und den zur Verfügung stehenden Daten, können klinische Entscheidungsunterstützungssysteme generiert werden, welche dem praktizierenden Arzt Informationen zum wahrscheinlichen klinischen Ausgang des Patienten bereitstellen. Diese Informationen können dabei helfen, schnell und abwendend in einen unerwünschten möglichen gesundheitlichen Verlauf einzugreifen. Zusätzlich können potentielle Risikofaktoren mithilfe von ausgeklügelten Feature Selection Methoden ermittelt werden.

In dieser Diplomarbeit, welche im Rahmen eines EIT-HEALTH finanzierten EU-Projektes namens FACET verfasst wurde, wird eine Methodologie für die Erstellung eines prädiktiven Modells und für das Auffinden von potentiellen Risikofaktoren für das Frailty-Syndrom vorgestellt. Außerdem wird die vorteilbringende Kollaboration von Ärzten und dem Datenwissenschaftler, welche sich in der Verbesserung des prädiktiven Modells widerspiegelt, aufgezeigt. Des Weiteren wird die außerordentliche Wichtigkeit der Datenvorbehandlung demonstriert, im Speziellen der Umgang mit fehlenden Werten.

Auf den hier dargelegten Ergebnissen aufbauend können in zukünftigen Arbeiten gefundene Einblicke weiter analysiert und in größeren Kohorten validiert werden. Dies mit dem Ziel, ein Modell zu realisieren, dass schließlich und endlich im Gesundheitssystem zum Einsatz kommt.

Schlüsselwörter

GESUNDHEIT, MASCHINELLES LERNEN, VORHERSAGEMODELLE, RISIKOFAKTOR FINDUNG, DATENVORVERARBEITUNG, IMPUTATION FEHLENDER WERTE

ÖSTAT Klassifikation

Information Systems (102015)

Machine Learning (102019)

Medical Informatics (102020)

ACM Klassifikation

Information systems: Data mining

Computing methodologies: Machine learning

Applied computing: Health informatics

This page intentionally left blank

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, May 16th 2017

Andreas Philipp Hassler, BSc

This page intentionally left blank

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor Professor Ernestina Menasalvas of the CTB at Universidad Politécnica de Madrid, for the continuous support and for having made this thesis possible in the first place.
Further, I am very grateful to my other thesis advisor Professor Andreas Holzinger of the ISDS at Technical University of Graz, for mentoring and encouraging me.
My sincere thanks also goes to my fellow labmates at the MIDAS Laboratory: Johnny, Cesar, Gerardo and especially to Ángel Garcia for his help and his insightful comments.
Also, I want to thank all my friends for their great support, especially Carolina for helping me correct this thesis.
Last but not the least, I would like to thank my family: my mother Reinhild and my father Josef-Hubert for always encouraging and supporting me throughout writing this thesis and my life in general, my little sister Sophia-Alexandra for keeping my mood up, my grandma Anna for providing the best Kärntner Kasnudeln in the world and also the rest of the family I have not mentioned by name.

Andreas Philipp Hassler, BSc

Graz, May 16th 2017

This page intentionally left blank

Table of Contents

1. Introduction and Motivation for Research

1.1 Introduction

Demographic predictions for the 21st century (2009 EU Ageing Report) show a new scenario characterized by a modest increase in life expectancy, but a significantly greater burden of disability, which will increase the demand for health and care costs and challenge the sustainability of the system. Both the ageing of the population and the growth of the population are driving the increase in Disability Adjusted Life Years (i.e. DALYs) due to the burden of non-communicable diseases in older ages, associated with an increase in years lived with disability. According to the last Global Burden of Disease (2010), disability is the main consequence of the concurrence of the ageing process, lifestyles and health conditions.(Murray et al., 2013) According to the report Ageing 2009 from the European Union (EU) Commission, the number of people aged 65+, in Europe, will almost double over the next 50 years, from 85 million in 2008 to 151 million in 2060. This is a great challenge for establishing new approaches with more efficient targets for public health and for older people. Hence, the aim is the increase of the life expectancy free of disability and therefore preventing and/or delaying the onset of dependence. This will favor optimization of opportunities for health, participation and security in order to improve quality of life as people age. That is active and healthy aging.(Committee et al., 2009)

In the field of today's data science there is a wide variety of new and sophisticated computational methods and also tools for building predictive models and performing enhanced data analysis. This collection of methods also offers a vast

variety of applications in the field of medicine and has already become an essential instrument. Hence, predictive data mining is for example intensively used in the research of molecular biology nowadays. The analysis of high-throughput data coming from mass-spectrometers or from DNA-micro-arrays serves as an example for this. In clinical medicine these methods are used to offer support in tasks such as decision making based on the patient's data. This covers the spectrum of diagnostic, therapeutic and monitoring tasks. Previous collected patient data can be used to build a predictive model which provides a prediction for the clinical outcome. Clinicians can act on this information and promptly react to possible or likely adverse events. (Bellazzi and Zupan, 2008)

Such an adverse event is for example the onset of the frailty syndrome, which according to Fried et al. (2001) is defined as follows:

> Frailty is considered highly prevalent in old age and to confer high risk for falls, disability, hospitalization, and mortality. Frailty has been considered synonymous with disability, comorbidity, and other characteristics, but it is recognized that it may have a biologic basis and be a distinct clinical syndrome. A standardized definition has not yet been established.

Data analytics can of course also be applied to analyze retrospective clinical data of the ageing population which can be crudely separated into healthy and frail people. This, in order to help to find early predictors for frailty, which in turn would enable the creation of policies for early prevention and adequate early on treatment of the frailty syndrome.

Furthermore, this would undoubtedly have a high beneficial impact on society. Sure enough this undertaking, in order to be fruitful, requires extensive medical records of elderly patients.

The Toledo Study for Healthy Aging (TSHA) began in 2006 and includes older adults selected by random sampling from the Toledo census, with ages over 64 years. Briefly, the TSHA is a population prospective cohort study aimed at studying the determinants and consequences of frailty in institutionalized and community-dwelling individuals older than 64 years living in the province of Toledo, Spain. Data was collected in three ways. Firstly, six psychologists conducted computer-assisted

interviews, performed face to face. Secondly, three nurses did a physical examination and performed some clinical and performance tests at the subject´s home. Finally, the participants went to their health center to provide a blood sample while fasting. (Garcia-Garcia et al., 2011)

FACET, which is short for **F**r**a**ilty **C**are and w**e**ll func**t**ion, is an EIT-HEALTH financed project, focused on the development of a platform and new methodologies to prevent the frailty syndrome. FACET focuses on the 'quality' of the years to be lived. The aim of that project is to develop a tool to integrate and query human phenotypic data in order to early detect frailty. In general, the early detection of impeding disease is complex. Therefore, a clear algorithm and clinical-friendly screening tools for detection of frailty and disability are lacking. There is a gap between living longer and living healthy. The development of early detection tools will permit intervention to prevent or delay the onset of frailty (and prevent further disability).

One of the main components of the FACET project is the data analysis layer, as it is responsible for providing the platform with the intelligence and the knowledge on which future decisions and policies are based.

In fact, the Universidad Politécnica de Madrid is responsible for this layer and this project thesis is framed inside the work of the development for the data analysis layer. As the FACET project has a more extensive goal, this thesis should be considered as an early development stage for the data analysis layer.

1.2 Objectives

The main aim of this thesis is to demonstrate that data science applied to medical data of elderly, partly frail people can help to obtain a predictive model for the frailty syndrome. This model could prolong noteworthily the healthy and independent living of the older European population, enhancing the functional autonomy by early detecting the risk of becoming frail.

Fulfilling this aforementioned goal means achieving the following scientific goals:

1. Generation of a classification model which is able to predict the risk of frailty

in patients.

(a) Developing a methodology for pre-processing the data.

(b) Developing a methodology for handling missing data.

(c) Identification of risk factors which can be used as predictors ("biomarkers").

(d) Learning satisfactorily accurate models for frailty prediction.

2. Theoretical Background

2.1 Data Analytics

The proliferation, ubiquity and in-creasing power of computer technology has dramatically increased the data collection, the storage, and the manipulation ability. This, in turn, has created a new need for automatic data analysis, classification, and understanding.

In today's world there is an excess of data. It is accumulating with high speed and no end is in sight. The necessities for storing it are quite inexpensive and enable postponing of decisions about their actual use and purpose. Potential useful information remains hidden and is rarely exploited. So the overall idea of data mining is to find these hidden patterns in available electronic records with the help of computational tools. Data, which is analyzed in a considered and careful manner could arise to be a valuable resource. Therefore, it is not surprising that data mining is a fast growing interdisciplinary subject, as its principle is to turn a large quantity of information into useful knowledge. Quite often it is used as a synonym for knowledge discovery in databases (KDD).

Hand et al. (2001) uses the following as his working definition:

> Data mining is the analysis of often large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

However, according to Han et al. (2011) some see data mining just as a step in the knowledge discovery process. This process, described by Fayyad et al. (1996), is shown in figure 2.1. Here the iterative steps composing the KDD process are illustrated.

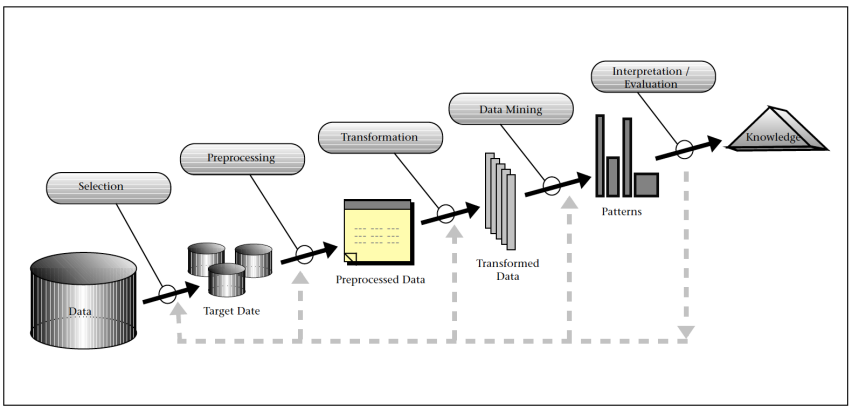


Figure 2.1: The steps and parts of the KDD process, which lead the way from data to knowledge.(Fayyad et al., 1996)

According to Fayyad et al. the process can be outlined as follows. At first one wants to obtain an understanding of the domain in question and to derive a goal for the KDD process. The next step is to build a suitable data set, which potentially contains the knowledge one wants to retrieve. The third step contains cleaning and preprocessing of the data. Tasks therefore include dealing with noise, missing values and time-sequence information. The following step is the projection and reduction of the data. This contains the tasks of finding suitable features with regard to the data mining goal and the reduction of dimensionality of the data set. The fifth step has to do with determining, which particular data mining method should be applied according to the goal. Some examples among other things are classification, regression and clustering. Then follows exploratory analysis and the selection of a model and a hypothesis. This includes selecting a data mining algorithm, which can be used for the search of patterns. The seventh step represents the actual search of patterns. The next step consists of the interpretation of the results or the found patterns. Here, one should consider revisiting the previous steps if needed. The ninth and last step is dealing with the obtained knowledge. Eventually integrating it into another system and make use of it or delivering it to a third party. The whole process can have iterations and loops between the steps.

2.2 Data Visualization

The vast amount of data, which was generated in the last decades, demands sophisticated data visualization and data analysis tools. Because of that many different

techniques have evolved with the aim to help the data explorer to get insights and raise his involvement in the data mining process. To ensure high quality data mining it is of highest importance to include the human in the data exploration process and combine the advantages of the computational power and human resources and expertise. In the task of visualization especially the perceptual capacity of humans is very useful. In order to ensure that this capacity can be exploited, the (probably multidimensional) data has to be transformed and presented in a lower dimensional form in order to be interpretable. When there is no exact goal and not much previous knowledge about the data, visualization techniques appear to be notably useful. Through the data exploration process also new hypotheses can be phrased, which can be validated also by visualization techniques themselves or machine learning algorithms and statistics. The main advantages of visual data exploration over automatic learning for one is, that it is able to work with inhomogeneous and noisy data sets. For the other, that it is more intuitive for the user and it has no requirements in terms of complex mathematical and statistical understanding. That is probably why visual data exploration often tends to provide better results where common automatic algorithms perform badly. This is why visualization methods nowadays are very thought-after. (Keim, 2002)
Keim proposes to classify visual data exploration into following three classes: *data type to be visualized*, *visualization technique* and the *interaction and distortion technique*.

2.3 Pre-Processing

The importance of preparing the data before starting to model a hopefully sophisticated predictive model, is highly underestimated. The majority of the time in a data mining project is spent on analysing and accordingly treating the data in order to obtain a suitable data set for the learning algorithms. A rule of thumb is that a data engineer spends 80% of his time preparing the data. A question that arises is, why prepare the data? According to Pyle (1999), one aspect of data pre-processing is that it prepares also the miner himself, which of course leads to the development of much better models. Further, appearing errors in the data are potentially harmful for the built model. Moreover, many algorithms cannot work with incomplete

data. Also the pre-processing may make the data easier to "digest" for the different used tools in the data mining pipeline. As there is obviously a strong need for data preparation techniques, find below some commonly used ones.

2.3.1 Discretization

There are some clustering and classification methods that can only work with nominal features and are not able to process numeric variables. So therefore they have to be discretized into a smaller number of different ranges. Also algorithms which are indeed able to work with numeric features could behave in a non-satisfactorily way, as many statistical methods assume, that numeric attributes are "well-built" in terms of distribution (optimally normal distributed).(Witten et al., 2016) Additionally, discretization accelerates the induction process and bears the potential to lead to much simpler and more accurate classification models. Also the risk of over-fitting is reduced, this by narrowing the space of hypotheses candidates that the learning scheme can search through, thereby possibly avoiding finding a very complex hypothesis which fits the data too well. Resulting models based on discretized data therefore also appear to be less complex. (Frank and Witten, 1999)
There are many supervised and unsupervised methods. Some work globally and others locally. A very common unsupervised method, as described by Witten et al. (2016) is to divide the range into a predetermined number of equal intervals. Fine distinctions could be easily destroyed by doing this. Further, this so called equal-width binning fails to distribute the data evenly. Some of the bins may contain no instances and others many. Therefore, using intervals of different sizes while making sure every bin contains the same number of observations, could be a better approach and is called equal-frequency binning.

2.3.2 Outlier Treatment

According to Pyle (1999) an outlier is a "single, or very low frequency occurrence of the value of a variable that is far away from the bulk of the values of the variable" (p. 73). He says that the first question that arises is, if it seems to be a mistake. The effect of an outlier, with regard to the final modeling result, could be big. The outlier could introduce an extreme distortion to the feature's statistics. Techniques for the

treatment of outliers are divided into two different sections, one for the treatment of univariate data and the other for multivariate data (Cousineau and Chartier, 2010). In the univariate domain the values of a feature itself are compared and a decision on "outlierness" is done. In the multivariate domain all features of an observation are considered and compared to the others in a multidimensional space. This possibly results in defining whole observations as outliers.

An example for a multivariate outlier identification technique is calculating the from Breunig et al. (2000) derived local outlier factor (LOF). It is a local measure and gives the degree of "isolation" of an observation. Here the density of k neighbours is compared to the density of the observation itself and the derived measure is the LOF. Outside a certain range, the observation is considered as outlier.

Laurikkala et al. (2000) studied the informal box plot identification of outliers in real-world medical data. Here they used box plots in order to detect univariate outliers directly. Further, they also used Mahalanobis distances to identify multivariate outliers. They found that removing these outliers increased the classification accuracy (they used discriminant analysis functions and the nearest neighbour method), while they noted a reduction of the predictive ability of the used methods. They further claim that statistic assessment usually acts on the assumption that there are well-behaving distributions. The main part of test statistics are created to identify single univariate outliers using a normal distribution (Barnett and Lewis, 1998). On the basis of this, appearing extreme values are declared as possible outliers. In clinical or medical data this is seldom the case, usually the data tends to be somehow skewed or definitely non-normal. The use of test statistics would need certain statistical parameters like distribution-type, transformations and even estimates of the distribution parameters. For large medical data sets, the execution of these preparation tasks would be very hard and work-intensive and definitely not applicable for practical use.

Box plots are a way of displaying the five-number summary (lower extreme, lower quartile, median, upper quartile, upper extreme) (Seigel, 1988). As Laurikkala et al. (2000) state, both skewed and symmetric data can be explored by using them. They also seem to be quite useful to find values, which do not appear frequently in categorical data. The definition of the thresholds for lower and upper outliers is defined in the following manner:

- $threshold_{lower} = quartile_{lower} - step$
- $threshold_{upper} = quartile_{upper} + step.$

The inter-quartile range times 1.5 is considered as *step*. The inter-quartile range is defined as $upperquartile - lowerquartile$ and contains 50 percent of the data (Laurikkala et al., 2000). A certain value x is considered a lower/upper outlier if it exceeds the lower/upper threshold. Laurikkala et al. conclude that there are mainly two motivations for the purpose of identifying the outliers. The first one is that outliers represent suspicious data, which should be removed before executing learning algorithms. The second one is that found outliers could contain important knowledge, which could be somehow valuable for domain-experts, in terms of gaining additional insight into the data. They further claim that the removal of outliers might help the descriptive analysis but may harm predictive accuracy for unseen values. All in all the effects caused by the treatment of outliers can lead to very different effects, strongly depending on the present data set.

2.3.3 Imputation Techniques

Already in the 70s the statisticians became aware of the fact that omitting observations with missing data, in order to receive a "complete-case" data set is inopportune. In the earlier days the missing data was usually replaced by the mean or the mode of the existing values for the feature. This approach became somewhat outdated because of its non-conformance. In order to achieve a valid subsequent statistical inference, there is the need to insert an adequate amount of randomness into the imputations and further, for the incorporation of that uncertainty when calculating standard errors and confidence intervals for interesting features. (Royston et al., 2004) There is a wide range of techniques for estimating the values of the missing values. There are methods which could yield more information than others, but they tend to be computationally costly. Other techniques are powerful under certain conditions, but they tend to introduce bias under different conditions. Estimation techniques which aim to produce mathematically optimal estimates appear to be very complex and they vary depending on the type of data they are applied to. These methods of high complexity are too time-consuming for big data, this also in regard to modern computer systems. Especially if time is of the essence, as in

certain business applications, these methods should be avoided. Highest priority lays on doing as little "damage" as possible. During imputation, there is a certain probability that out-of-range values may appear, which haven't been observed in the data. This is because not all values of the population may be covered, which strongly depends on the the sample size. Generally speaking, one is interested in finding a suitable estimator, which is able to make a satisfactory guess about the missing value. The perfect variant would be an unbiased estimator, one which does not interfere with the general characteristics of the variable. According to Pyle (1999) following statement stands:

> "Statistically, an unbiased estimator produces an estimate whose "expected" value is the value that would be estimated from the population."

Let's take for example the observations $10, 20, 30, NA, 50$. Here "NA" stands for "not available" and represents a missing entry. An unbiased estimate, which would produce the least amount of "damage" to the data, should be found. But the least amount of "damage" is not clearly defined. In regard to the mean an unbiased estimate would be 27.5. For an unbiased standard deviation the imputation should be about 46.59. So in order to not bias a certain statistical aspect, another one is harmed. Therefore, a decision in this regard has to be made. Also very important is to know which inter- and intra-relations of the variable should be preserved. There is not only the within-variable relationship but further also the between-variable-relationship. The latter one describes in which way the variable of interest changes depending on the behaviour of another one. The modeling tool of choice should definitely be able to preserve all this relations when imputing new values.
Regarding the decision which intra-variable measure is of higher importance, Pyle (1999) claims that the standard deviation contains by far more information because it reflects the variability of the variable in comparison to the mean, which is only a measure of central tendency. The standard deviation therefore delivers a measure for the distribution itself and provides because of that a more suitable estimate. Back to an even more important aspect: the inter-variable relations. In order to keep them, one sees that simply imputing static values obtained by statistical within-variable measures is not the way to go. Put more accurately, it would be a drastic distortion of the existing between-variable relationships. Especially, when the missing values

are not missing at random, a replacement with the same value in all missing places will introduce a strong bias. Optimally, all existing variables, whether they are strongly or weakly related, should be taken into account for the imputation, given that they contain values for the observation in question. So actually a prediction model should be built to predict the missing values, where not the accuracy is of highest importance but rather the creation of an estimate that least distorts the actually present values. Therefore, the main purpose of the replacement of missing data with certain imputations is not the use of these values themselves, but to enable the learning machine to work with the information that is contained in the other variables' values that are present. By simply not replacing the missings, the whole observation would be discarded and therefore valuable information may be lost as well. On the other hand, by replacing the missing values the introduction of bias and distortion is a possible outcome. Taken into account that imperfect multiple linear estimation produces far less bias than any method using constant values, the former clearly should be preferred. An example for such a method is the multiple linear regression technique. Regression methods after all are inherently mathematical and tend to be very susceptible for missing values themselves.
Pyle (1999) concludes that replacing missing values appears to be a very important step in the data pre-processing in order to make use of all the information that is contained in the data. Where high importance lays on the preservation of the feature relationship as well as on the original distribution of the feature. Also the introduction of new artificial patterns should be avoided. Further, Pyle states, that these introduced patterns could be "discovered" by the data mining analyst and even may somehow appear meaningful. Thus, sensible techniques are required which maintain even the weakest existing patterns.

Types Of Missing Data

The risk of introducing a bias due to the missing data depends on the underlying reasons for the missingness. According to Little and Rubin (2014) these reasons can be classified as:

- Missing Completely At Random (MCAR)

  This is the case when between the missing and the observed values are no

systematic differences. For example missing values because of a breakdown of the measuring device. Here the probability of missingness is identical for all observations. So the missingness neither depends on the feature itself nor another one.

- Missing At Random (MAR)

  A term introduced by Rubin (1976). When the systemic difference between observed and missing data is completely explainable by differences in observed data ones speaks of MAR. In other words, the probability that a value is missing depends only on features in the data set. For example women tend to not state their weight with a higher probability than man. So the missingness of values of the feature "weight" depend on the feature "gender".

- Missing Not At Random (MNAR)

  If the missingness depends on the feature itself or on unobserved properties which are not covered by the data, one speaks of MNAR. For example, overweight people tend to withhold information about their weight with a higher probability. Or people of certain religions or cults tend to not give blood and therefore certain blood features are missing.

Methods

Regressions Linear regression only focuses on two variables and is therefore more clearly in an explanatory sense. The basic assumption is a linear relationship between the two variables, with which the changing of one variable can be explained by the other. According to Pyle the linear regression technique involves discovering the joint variability of these two features. The obtained knowledge is then used to determine which value matches to the other available one. The term joint variability represents a measure of the way how one feature varies depending on the variation of the other feature. Due to the linearity, the relationship between the two variables can be expressed by a linear equation (2.1), which gives a straight line when visualized:

$$y = \alpha + \beta x \tag{2.1}$$

For every known value x a value y can be calculated.

Multiple linear regression works quite similar, the only difference is that it considers more features' joint distributions to extract an estimate for an missing value. Using more than one feature, which is contributing to the joint variability, is somehow using more evidence and leads therefore generally to a better estimation of the missing value. In linear algebra notation, multiple regression can be expressed as shown in formula 2.2. The missing value for the observation i for the feature y can be calculated using all other features x, where n marks the total number of the features without the feature in question (y).

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + ... + \beta_n x_{i,n} \tag{2.2}$$

As Pyle states, in practice linear relationships between the features can be found quite often. Even when the relationships is of a non-linear nature, an estimate created under linearity-assumption, tends often to be adequate. The portion of bias introduced by doing this appears to be often under the noise level. The finding of linear relationships is compared to non-linear relationships fast and easy. So Pyle says that linear techniques run very quick even when the number of dimensions is high. He further claims that the amount of introduced distortion and also the relation between speed and flexibility is putting linear techniques in favor over non-linear ones. However, when the relationship is obviously and extremely non-linear and the modeler has knowledge about that, a special replacement method should be used instead.

Multiple Imputation

Multiple imputation (MI) techniques appear to be very useful for general-purpose treatment of features with missing values in multivariate analysis. Unlike single imputation, MI generates a defined number m of imputed data sets. So every missing value is replaced by m different imputed values. The uncertainty of imputation is considered. Statistics for each imputed data set are estimated and then combined into a single estimate. The criticized downside of single imputation, ignoring the uncertainty and the resulting bias, is avoided with MI, as a correct execution can lead to a good estimate of the "real" respectively probable values. (Zhang, 2016) The basic concept was first introduced by Rubin (1977) and consists according to

Allison (2000) of the following steps:

- Imputation of the missing values by using an appropriate model that incorporates random variation.
- This is done M times (where M is usually between 3 and 5), so that there are M complete data sets produced.
- Analysis on the different obtained data sets is done by using standard complete-data methods.
- Averaging of the values of the parameter estimates across the M samples to obtain a single-point estimate.
- Calculation of the standard errors by (i) building the average of the squared standard errors of the M estimates, (ii) calculating the variance of the M parameter estimates across samples and (iii) combining the two quantities using a formula.

A promising method presents MICE, which stands for multivariate imputation by chained equations and was described by Van Buuren et al. (1999). The MICE algorithm is a Markov chain Monte Carlo (MCMC) technique. In case that the conditionals are compatible, the algorithm works as a Gibbs sampler, a Bayesian simulation method, which samples from the conditional distributions in order to obtain samples from the joint distribution. Conventually, a derivation from the joint probability distribution is done to obtain the full conditional distributions. In MICE, however these conditional distributions are controlled by the user and the joint probability distribution itself is only known implicitly, and further may not exist. The last mentioned part is quite unfavourable seen in terms of theory. Yet in practice this has not let to unsatisfying results. Convergence to a stationary distribution is only reached when the Markov chain satisfies three certain characteristics. Firstly, irreducibility, which means that the chain is capable of reaching all interesting areas of the state space. Secondly, aperiodicity, meaning that there is no oscillation between states and lastly, recurrence, which signifies that all interesting parts can be reached infinite times, and this at least from all starting points. The first mentioned criterion usually does not represent a problem for the MICE algorithm. The second one is a possible issue, when imputation models are inconsistent.

Further, non-recurrence could be problematic, expressing itself by non-stationary or explosive behaviour. Van Buuren (2012) states though, that in his experience as long as the imputation model parameters are estimated from the data, appearing non-recurrence is mild or absent.

Missing Data In EHRs

Especially in the field of medicine where electronic health records (EHRs) for the collection of patient data are used, missing data has a high prevalence. The different, often unknown, causes of missing data could introduce a bias. (Beaulieu-Jones et al., 2016) Also the observations may be missing sporadically. Depending on the different features, a complete-case data set may only contain half the data (Royston et al., 2004). So in order to built sophisticated models with small data sets, the contained information should be used as good as possible. Therefore, there is clearly a need for a suitable imputation of missing values.

2.3.4 Dimensionality Reduction

The already mentioned problem arising from the massive accumulation of data is asking for special processing tools. One such tool is the dimensionality reduction. The goal is to reduce the dimensions without loosing important information. It is often used as pre-processing step. Dimensionality reduction is one of the techniques which are used to remove noisy (or irrelevant) and redundant features. The dimensionality reduction techniques can be divided into feature extraction and feature selection. The primer ones are used to project the features in a new and lower-dimensional feature space, where the new built features commonly are combinations of the original features. Examples therefor are the Principal Component Analysis (PCA), the Linear Discriminant Analysis (LDA) and the Canonical Correlation Analysis (CCA). The other technique, feature selection, aims to select a suitable subset of features, which minimizes redundancy and maximizes the relevance to the target variable (examples: Gain, Relief, Lasso and Fisher Score). Both mentioned technique categories, feature extraction and feature selection, are very capable in terms of improving the performance of the classification model, lowering the computational costs, lowering the needed memory storage and further for building

improved models regarding generalization. Feature extraction makes it very hard to relate the new derived feature to the original ones, they further do not contain physical meaning. Feature selection on the other side keeps the original features and therefore the underlying original physical or *"real-world"* meaning. This makes the selection of features superior over the extraction. This because the readability and interpret-ability are much better. (Aggarwal, 2014)

Feature Extraction Techniques

Principal Component Analysis (PCA) One of the traditional tools for dimensionality reduction is the Principal Component Analysis (PCA). It projects the data in a space with fewer dimensions by creating new axis, which keep the maximum of the initial data variance. A big disadvantage is that this tool is linear, non-linear dependencies or relations between the features could get lost. If in the next step the linear pre-processed data is used with nonlinear data analytic tools, this should be considered bad practice. One possibility of using PCA for non-linear projections is to apply it locally in restricted sub-spaces. Conceptually, joining local linear models leads to a global non-linear one. However, this carries the big disadvantage of being non-continuous.
An interesting application which bypasses the mentioned disadvantage is Kernel PCA. Here the data is at first transformed into a space with more dimensions. Having the sophisticated developed kernel methods up one's sleeve, the data is at first transformed into a space with more dimensions. This can lead to fruitful results. Contradictory seems the initial transformation in a higher space and then the reduction of the same, but in some cases this can be quite useful. (Verleysen and François, 2005)

Linear Discriminant Analysis (LDA) This method is well suited for the application to cases where the within-class frequencies are not equal and their performances have been investigated in test data which was generated randomly. LDA maximizes the ratio of between-class variance to within-class variance in any given data set, thereby providing maximal separability and projecting the data into a lower-dimensional space. The overall goal is to decrease the variation within the classes and to maximize the separation between the classes. Here, in comparison to

PCA, the location of the original data sets is not changed but more class separability is provided. (Balakrishnama and Ganapathiraju, 1998) LDA is a well-known scheme for the reduction of dimensions and feature extraction. Fields of applications are for example image retrieval, microarray data classification, face recognition and also speech recognition. (Ye et al., 2004) (Balakrishnama and Ganapathiraju, 1998)

Canonical Correlation Analysis (CCA) This statistical method is used to investigate the relationship among two or more variable sets. It represents the multivariate form of the general linear model, which holds the presumption that all analyses are correlational. (Thompson, 2005) The correlation coefficients can be directly calculated from the data sets and also from the reduced/lower-dimensional representations like co-variance matrices.(Weenink, 2003)

Feature Selection Techniques Methods for feature selection became popular in the late 90's. Then when it was still an advantage in data understanding when the number of variables was not all too high. Which was also good in terms of reducing training time and improving the prediction performance in order to help to deal with the curse of dimensionality.

In Blum and Langley (1997) one can find a extensive review of methods for feature selection. Tasks of data analytics in the realm of gene and protein expression, chemistry or text classification have elevated the importance of feature selection extremely, not only because of the high number of features in data sets nowadays, but also because in some cases there are not many observations to work with. An ample review and comparison of feature selection methods can be found in Guyon and Elisseeff (2003)

At the beginning, before executing the classification algorithm, feature selection is done. Many times the data collection is done by individuals who are no experts in the respective domain. This leads to an accumulation of irrelevant features, which in turn leads to the building of poorly performing models and the needless use of computational resources. This due to the insufficient relation to the target feature/label one is interested in. These non-related features actually lower the accuracy of the predictive model (Kohavi and John, 1997a) and they lead to over-fitting. Especially when there are a small number of observations, these features can have a high neg-

ative impact on the result. One feature alone may not worsen the model much, but a multiplicity of them can have an observable adverse effect. The resulting model could therefore have a poor generalization. This is why selecting suitable features is that important and why there should be found a small (possibly minimal) feature set, which leads to the best result in the classification task.(Aggarwal, 2014) This here elucidated problem, also called *minimal-optimal problem* (Nilsson et al., 2007), has already extensively investigated and lead to the development of plenty solutions. Another very important problem, which should not be underestimated, is the overall identification of all-relevant features. This can be of particular interest when it is not simply the goal to implement a high precision classifier ("black-box principle"), but to better understand underlying mechanisms in the data.(Nilsson et al., 2007)

Depending on the aim, which depends on the labeling of the trainings set (labeled or not) the algorithms can be divided into supervised, unsupervised and semi-supervised feature selection algorithms. The supervised techniques can further be divided into filter, wrapper and embedded models.

- **Filter Models**
  They select subsets of features as a pre-processing step and use a certain performance criterion on them to perform an evaluation of their suitability for the classification. There is no dependency on the specific algorithm which is used. In some cases they compete with wrappers as being more efficient. The quantification of the relevance of the feature to the process of classification is done by different measures (examples given: Gini Index, Entropy, Fihser's Index). In the filter model there is a separation of feature selection and classifier learning, therefore the bias of a machine learning algorithm does not interact with the bias of a feature selection algorithm. It is depending on general characteristics of the data, like certain measures (distance, correlation, consistency, dependency and information). (Aggarwal, 2014)

- **Wrapper Models**
  The wrapper methods popularized by Kohavi and John (1997b) assess subsets of features according to their suitability for predicting the target variable using a search algorithm to search through the space of possible features and do an evaluation on each subset by executing a model on the subset. In these

methods the induction learning machine algorithm is taken as a black-box to score subsets of features with regard to their predictive power. The induction algorithm itself is used as part of the evaluation function. Here, the feature selection process is sensitive to the used classification algorithm. This method takes into account that different algorithms may work better with different features. (Aggarwal, 2014) Given that the number of features in the data set is not all too high, the complete feature set can be thoroughly searched through. Wrapper models tend to be computationally expensive and they are therefore criticized as being a "brute-force"-method. In general efficient search strategies are desirable. Also it has been found that coarse search strategies may lower the risk of over-fitting (see Reunanen (2003)).

- **Embedded Models**

  The embedded techniques were supposed to minimize the shortcomings of the aforementioned models. They work, like the filter models, as well with the help of statistical criteria in order to select suitable features with a given cardinality. Further, like the wrapper model, they consider classification accuracy with the goal to maximize it via selection of the most suitable subset of features. The great advantage of the embedded models is that they are comparable in terms of accuracy to the wrapper models as well as in terms of efficiency to the filter methods.(Aggarwal, 2014) As mentioned, they implement the same concept as in the wrapper model, but work by optimizing a two-part objective function with a goodness-of-fit term and a penalty for a large number of variables. Here the feature selection is done as part of the training process and is in general specific to the learning algorithm. The fitting of the model and the selection of the features is done at the same time, which makes them far more efficient. The available data seems to be better used, because there is no need of splitting it into a training and validation set. Further, they do reach a result faster. This because they avoid retraining a predictor from scratch for every feature subset which is under investigation. (Guyon and Elisseeff, 2003)

An example for the use of embedded techniques is the random forest algorithm (Breiman, 2001). The from Genuer et al. proposed two steps are: (i) preliminary elimination and ranking and (ii) the variable selection itself. In

step (i) the random forest scores of importance are computed and variables of small importance are discarded. The m remaining variables are then ordered according to their importance. They main objectives of (ii) are: on the one hand to find variables which are strongly correlated with the target variable (for the purpose of *interpretation*) and on the other hand, to find a small set of variables which are sufficient enough for a good *prediction* of the target variable.

- **Interpretation:** RF models are constructed where the k first variables ($k = 1...m$) are used. The variables which are involved in the model with the smallest OOB (out of bag) error are then chosen/selected.
- **Prediction:** Here an ascending sequence of RF models is created by step-wise invoking and testing the variables. The start point is the list of ordered variables from the previous step. At the end, the variables which are part of the last model are finally selected.

Unsupervised feature selection represents a search problem without any class labels. They use clustering quality measures, but in high-dimensional data additional constraints should be used as well. Without them finding suitable features is very unlikely. Concluding, there are supervised feature selection techniques, which assess the relevance of the feature to the target variable, therefore needing a sufficient number of labeled data. Moreover, there are unsupervised techniques working with unlabeled data, where the determination of the relevance is very hard. Often one has to work with high-dimensional data with only a few labels. Here the combination of both feature selection techniques can be very useful. This is the so called semi-supervised feature selection, which uses both labeled and unlabeled data in order to find suitable features.
The generalization of feature selection is feature weighting. In feature selection techniques the feature receives a binary weight. Zero means not selected and one means selected. This is extended in feature weighting, where a weight usually in the interval $[-1, 1]$ or $[0, 1]$ is assigned to each feature.
According to Aggarwal (2014), the selection of the features can in general be crudely divided into the four steps:

- The generation of a candidate subset.

- The evaluation of the subset according to an evaluation criterion.
- The determination of the best subset, regarding the evaluation criterion. It is found, when the stopping criterion is met.
- The validation of the chosen subset with a validation set or by using domain knowledge.

After using the feature selection methods, irrelevant and redundant features should have been successfully removed. In classification problems this should leave the features which are highly associated with the target concept or variable. Now, by exclusively using the chosen subset, the running time will be lower and the generalization of the model will be much better. According to Aggarwal, following criteria for feature selection for classification regarding the (possibly minimal) chosen subset do stand:

- Accuracy of the classification based on the selected feature subset does not significantly decrease, compared to the classification accuracy using the complete feature set.
- The distribution of the resulting class (contains only values of the chosen features) should be as close as possible to the original class distribution, given the complete feature set.

2.4 Modelling

One speaks of learning when a new input leads to an enhancement of the performance of a system in the future. Like animals and humans are able to learn from new experiences, also "machines" have inherited this ability. In computational terms speaking, a new data input can change the code of an algorithm in a way so that it will perform in an altered manner in future interactions. Many machine learning techniques are derived from the scientific field of psychology.
Machine learning tasks are associated with artificial intelligence (AI). Some examples therefore are diagnosis, prediction, recognition and planning. (Nilsson, 1996)

Abu-Mostafa et al. (2012) nicely demonstrated in his book the (machine) learning process which is shown in figure 2.2. Samples of the input values $x \in \mathcal{X}$ and the

output values $y \in \mathcal{Y}$ are used to approximate the target function $f : \mathcal{X} \to \mathcal{Y}$. The samples are given as paired records of input and output values (x_i, y_i). The chosen final hypothesis is called $g : \mathcal{X} \to \mathcal{Y}$, where $g \in \mathcal{H}$ and the hypothesis set $\mathcal{H} = \{h\}$. In the example given by Abu-Mostafa et al. (2012) the goal is to derive a credit approval function. Given are the historical records of the customers and a hypothesis set to chose from. The learning algorithm $\mathcal{A}$ uses the available hypotheses out of the hypothesis set $\mathcal{H}$ and tries to find the "best" fitting one. These two, $\mathcal{A}$ and $\mathcal{H}$ together, build the so called *learning model*.

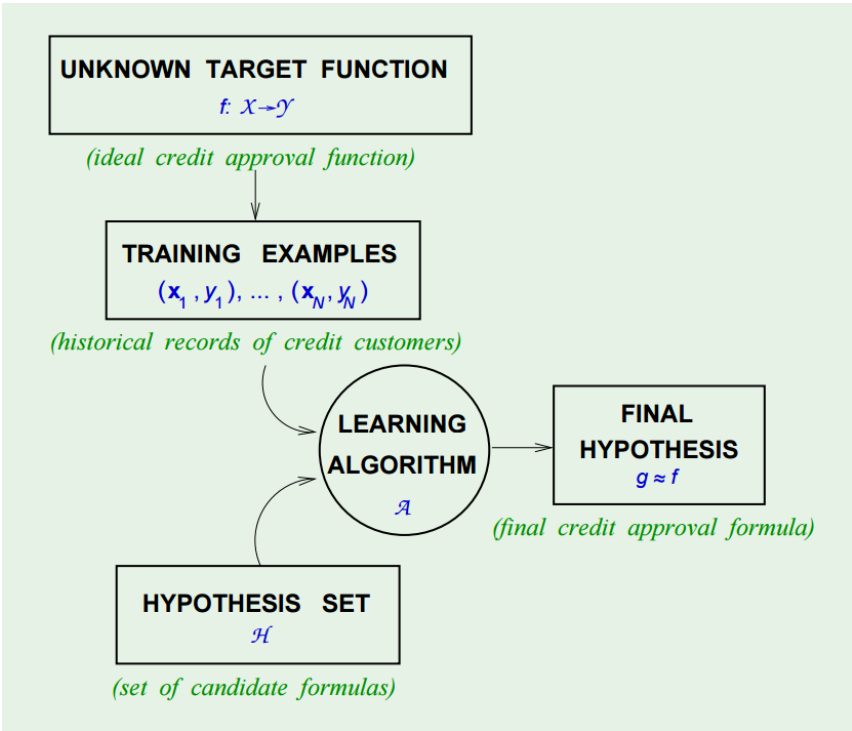


Figure 2.2: This graphic shows the necessary steps for learning the final hypothesis g, which tries to approximate the "true" hypothesis f. Using records (x_i, y_i), the learning algorithm $\mathcal{A}$ and the hypothesis set $\mathcal{H}$. Abu-Mostafa et al. (2012)

Machine learning is a very broad domain and has many subbranches. Therefore, it is not that easy to clearly separate the paradigms and concepts because of the overlap.

2.4.1 Paradigms

Referring to Holzinger (2016b) the learning paradigms can be crudely divided in the following manner.

Supervised learning

If the expected output for the test data is explicitly given, someone speaks of a supervised learning setup. As data a collection of (x,y) pairs is given. Learning methods from this paradigm are the most widely used ones. (Jordan and Mitchell, 2015)
An example would be hand-written digit recognition, where the test data is a collection of images of hand-written digits (x-data) with a corresponding label (y-data) which contains the actual digit as a numeric value. It is called supervised because it seems that someone - the supervisor - already has denominated them correctly by assigning the correct output digit. (Abu-Mostafa et al., 2012) Here the task is to learn the right label for every input. Assigning every future sample (for example the image of an unknown hand-written digit) correctly to a finite number of discrete classes is the final goal. Therefore, we speak of a *classification* problem. (Bishop, 2006) If we want to have a continuous output, it is also possible to use supervised learning methods. In this case the y-data contains continuous values instead of discrete labels. Here on speaks of *regression*.

Unsupervised learning

In this learning scenario we do not have any output information, only the input data is given. Here there is a different aim, one does not want do assign the data to a label or a numeric value like before, but rather see if it contains some structure and is therefore separable. (Abu-Mostafa et al., 2012) Thus, this is not a class prediction as it was the case with supervised learning but class discovery. (Ramaswamy and Golub, 2002) The goal is to divide the data set into groups of similar data, which can be done by using different similarity measures. (Zanin et al., 2016)

Semisupervised Learning

It uses labeled and unlabeled data to perform supervised and unsupervised learning tasks. In inductive semi-supervised learning the learner has both labeled and unlabeled data and tries to learn a predictor f. The main aim is to find a predictor which performs better than the one which was just devised from the labelled data alone.

Another sub-field here is transductive learning, where the same setting stands as before. The main goal here is to make predictions on the unlabelled training data, where one has no intention of generalizing to unseen test data. (Zhu, 2011)

Reinforcement Learning

In this learning setup the output is given partially. We are only given some output data and furthermore a grade or a measure which tells us how "good" or "bad" the assigned output is. This kind of learning can for example be useful if the task is to learn a game. Different actions lead to different outcomes and the goal is to find the best action which maximizes the obtained reward. (Bishop, 2006)(Abu-Mostafa et al., 2012) Therefore, this branch of machine learning can be seen as one that benefits from experience, which was attained through interaction with the surrounding and the resulting feedback to evaluate prior behavior. This evaluation leads then to an improvement of the posterior behavior of the system. While it is more autonomous than supervised machine learning, it is not able to learn from interactions on its own. Often it is unfeasible to obtain samples that are representative and correct for all situations. (Holzinger, 2016a)

Active Learning

It belongs to semi-supervised machine learning. The basic principle of active learning (AL) is, that the machine learning algorithm itself is able to create the data query. This can lead to an higher accuracy while using fewer training labels (y-data). The queries are answered by a so called *oracle*, which could be for example a human who assigns labels to the unlabeled data-instances of the query. This machine learning technique finds its application in situations where there are many observations, which can be easily accessed, but where the labeling process is tied with high costs or time consumption. (Holzinger, 2016a)

Preference Learning

In preference learning (PL) the main aim is to create a predictive preference model based on empirical data, which contains specific preferences of a user or a collective

of users. Methods for preference mining can be used to create a personalized recommendation system based on the information which is available on the user. In the beginning preference learnings' central task was learning to rank. It can be regarded as a natural link between machine learning and decision support.(Holzinger, 2016a)

Interactive Machine Learning

The previous three described paradigms build up the basis for interactive machine learning (iML). According to Holzinger following definition stands:
"We define iML-approaches as algorithms that can interact with both computational agents and human agents and can optimize their learning behavior through these interactions."
So the main aim here is, to include a domain-expert as an agent in the knowledge discovery process. The machine learning algorithm together with the expert can achieve fruitful results, which could not have been accomplished by each alone. This domain-expert can be considered as the "human-in-the-loop".
Concluding, the combined use of human-computer interaction (HCI) and knowledge discovery and data mining (KDD), where human and machine intelligence are working together, can be used to attain novel insights into data. (Holzinger, 2016a)

2.4.2 Methods

Even given that the main focus of this thesis is more on classification models, also clustering methods are briefly reviewed as it will be shown how these models can be used in the data understanding stage to get some insights into the data.

Classification Models

Linear Regression Is a method used in statistics for explaining the behavior of one target variable depending on one or more independent variables, which can also be called predictors. Its main application is the determination of the mean value of the target variable. The error of this prediction is normally distributed. Therefore, the underlying assumption seems to be that the target variable as well as the error are normally distributed. Nevertheless, the model appears to be very robust in case

of violations of these presumptions. (Hilbe, 2009) The resulting output variable $y(x, w)$ is a linear combination of the input variables $x = (x_1, ..., x_D)^T$ and the parameters $w = w_0, ..., w_D$. The formula can be seen in 2.3. The total number of parameters is M and $\phi_j(x)$ represents the basis functions.

$$y(x, w) = w_0 + w_1 x_1 + ... + w_D x_D = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x) \tag{2.3}$$

In order to use linear regression as a classification technique, a threshold is used to assign the resulting value to a class.

Logistic Regression The overall principle of this linear model is more or less the same as in the linear regression model. More precisely, linear regression represents a generalisation of it. The main difference between them is that the result in the logistic regression model is binary or dichotomous (Hosmer Jr and Lemeshow, 2004). Here a logistic sigmoid function is used on the features (inputs). This model seems to be the better choice in the case of binary responses. The posterior probability of the class c, given the observation x, is denoted in 2.4

$$p(c|x) = y(x) = \sigma(w^T x) \tag{2.4}$$

Support Vector Machines SVMs are binary linear classifiers that model concepts by creating hyperplanes in a multidimensional space and can be used for classification and regression. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class as this minimizes the error. The axes of this space are given by the features available in the data set, whose values should always have a numerical form. Records are mapped into this space, and the best linear separation between them is then calculated. (Cortes and Vapnik, 1995)

Decision Trees A decision tree consists of nodes and leafs. The nodes can be considered as tests and lead to a splitting of the input space. This splitting is based on a specific feature and leads to a certain root-to-leaf path. The resulting leaf represents a category or a label. At each node such a test is performed, the outcome

is exclusive and follows strictly the input pattern.
Depending on which features are used, one can speak of *multivariate* - the tests are performed on some features of the input data at once - or *univariate* - the tests are applied on one of the features - tests.
If all the tests on the nodes have two possible outcomes, one speaks of a *binary decision tree.* (Nilsson, 1996)

Random Forests This model presents a ensemble method which uses a combination of decision trees. Each of these trees was grown from a randomized vector sampled in an independent way and they all show the same distribution in the forest. In case of classification, each of these trees votes for a class and at the end the most popular one is chosen as the result class.(Breiman, 2001) (Louppe, 2014)
Definition of Breiman (2001):

> "A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, ...\}$ where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x."

Random forests perform calibrated as well as uncalibrated very well on medical data. They seem to operate very "save" and show a very good overall performance on different data sets. (Caruana and Niculescu-Mizil, 2006)

Especially in problems where a large number of variables are given, like in medical problems, each containing very little information, the classification accuracy has shown to improve from growing an ensemble of trees and letting them vote for the most popular class. Random forests Breiman (2001) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Each tree in random forest is grown as follows:

- Sample with replacement the number of cases in the training set at random. This sample will be the training set for growing the tree.
- Given M input variables, select randomly at each node $m << M$ variables and choose the best to split the node.

- Grow the tree without pruning.

The greatest advantage of Random forests is that they do not over-fit. Further, they are known to outperform most of the known algorithms in terms of accuracy and also, as earlier on mentioned, in terms of stability.

Logistic Model Trees Two prominent classification methods are here combined in order to merge their advantages and at the same time to attenuate their disadvantages. One of them is the linear logistic regression model and the other one is the tree induction model. The first one is on one hand known to be stable in the process of model fitting - therefore showing low variance - but shows on the other hand a potentially high bias. The second one often shows high variance and a low bias and is therefore working more "freely" and hence more capable of capturing nonlinear patterns, yet more prone to over-fitting. Logistic model trees lead to a higher average accuracy than C4.5, logistic regression, model trees and seem to be competitive with boosted trees. (Landwehr et al., 2005)

Naive Bayes Classifiers A Naive Bayes classifier presents a quite simple probabilistic classifier based on the application of Bayes's theorem. This with the assumption of strong (naive) independence between the features. According to Aggarwal (2014) it is very well suited for applications where there are many dimensions. He further claims, that notwithstanding its simplicity the achieved classification performance is quite comparable to more complex, sophisticated models such as neural networks and decision tree based classifiers. The naive bayes classifier also impresses with a high accuracy and a high velocity when applied to huge data sets. A common and good application of this algorithm is document classification, medical diagnosis and computer performance management (Aggarwal, 2014).
Bayes' theorem shown in 2.5 consists of an output, the posterior probability $p(c|x)$ which describes the probability of the class c given the observation x. Further, as input serves the likelihood function $p(x|c)$, which denotes the probability of the observation x given the class c. Before observing the data, the assumptions about class c are captured in form of a prior probability function $p(c)$. The probability of the

value x ($p(x)$) denotes the evidence.

$$p(c|x) = \frac{p(x|c) \cdot p(c)}{p(x)} \tag{2.5}$$

Less formal, put into words, the relationship can be represented as in 2.6.

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{2.6}$$

Artificial Neural Networks(ANN) This method is inspired by the structural aspects of biological neural networks. ANNs are represented by a set of connected nodes in which each connection has a weight associated with it. The network learns the classification function by adjusting the node weights. The simplest kind of neural network is the single layer perceptron Rosenblatt (1958), which has two important drawbacks: i) perceptron-like methods are binary, in the case of multi-class problems the whole classification problem must be split to multiple binary sub-problems, ii) single layer perceptrons are only capable of learning linearly separable functions, and thus are not suitable for the kind of problems usually found in real KDD applications. The back-propagation algorithm Werbos (1974) used in conjunction with an optimization method such as gradient descent were proposed to avoid those problems. The method calculates the gradient of a loss function with respect to all the weights in the network. The gradient is fed to the optimization method which in turn uses it to update the weights of the nodes in the network trying to minimize the loss function. All the basics to build neural networks can be found in Hagan et al. (1996) and in Zurada (1992).

2.5 Clustering

Clustering, which is an unsupervised learning approach, is the division of data into groups or clusters that contain similar records (according to some chosen similarity measure) and separation of dissimilar records into different clusters. According to Kaufman and Rousseeuw (1990) clustering is defined as follows: *partition a given data set in groups, called clusters, so that the points belonging to a cluster are more similar to each other than the rest of the items belonging to other clusters* In Jain et

al. (1999) a taxonomy of clustering techniques is presented, and further an overview of its fundamental concepts and methods. Moreover, it describes several successful applications of clustering such as image segmentation or object and character recognition. However, it is not easy to classify clustering algorithms as the categories very often overlap. According to the survey that can be found in Berkhin (2002) the following kind of clustering algorithms can be distinguished:

- Based on hierarchies. The hierarchical clustering combines instances of the data set forming successive clusters in a tree form that is called dendrogram. Thus in the lower level of the tree there is a unique cluster for instances, and the upper levels are clusters of the nodes bellow. Here it can be distinguished between agglomerative clustering and divisive clustering, depending on the criteria for the group nodes.

- Partitions based. The clustering methods based on partitions divide the data set into different disjoint subsets. The operation involves assigning points to different clusters, whose number is initially set, improving clusters in each iteration until a heuristic defined previously finds the optimal division. For example, the k-means Hartigan and Wong (1979a) algorithm belongs to this category of methods.

- Density based. In the previous mentioned algorithms the similarity measure for points to be assigned to a certain cluster is a distance measure. However in density based algorithms, clusters are not based on distance but on density measures. For example, the DBSCAN Ester et al. (1996) algorithm belongs to this kind of clustering techniques.

More mathematically speaking, the goal of clustering is to divide n data points in a d-dimensional space $\mathbb{R}^d$ into K clusters. In other words to group physical or abstract objects into classes with high similarity. Overall, it is desired to maximize the intra-cluster similarity while minimizing the inter-cluster similarity. The methodologies are following the maxim *divide et impera* (lat. for divide and conquer) in order to pave the way for further processing of the data.
Over the time many different approaches have been developed to obtain that certain objective. Further, also many different similarity measures have evolved which, depending on the data, lead to success in the partitioning task.(Chen et al., 1996)

2.5.1 k-means

Formally speaking, this algorithm divides M instances in N dimensions into K clusters following the minimum of the within-cluster sum of squares (WCSS). Because it is not virtual that the result shows the minimal sum of squares against all partitions, only the *local* optimum is sought-after. This is achieved when the assignment of any point to another cluster does not result in a reduction of the WCSS. (Hartigan and Wong, 1979b)

Determining the cluster number k

Like mentioned before, k-means clustering algorithms themselves can't figure out the optimal cluster number. Further, the right number of clusters is in the most cases not apparent. Often, the number of clusters are chosen *ad hoc* on the base of prior knowledge, presumptions and practice. High dimensionality complicates the task of finding an adequate cluster number even more, also when the data appears in well separated clusters. (Hamerly and Elkan, 2004)

Often the Akaike information criterion (AIC, Akaike (1974)) and the Bayesian information criterion (BIC, Schwarz et al. (1978)) are used to determine which number of clusters seems to be the best. One always tries to obtain the minimum AIC respectively BIC value and then the best number of clusters k is found. Another often in practice used measure is the WCSS. It is plotted and according to the "elbow criterion" the best number of clusters is chosen. When the from the cluster/WCSS xy-plot depicted function is flattening, similar clusters are divided, therefore the "elbow" is selected as optimum. In practice, often more than one "elbow" can be found and then it depends strongly on the clustering goal, which "elbow" finally is chosen. (Ketchen Jr and Shook, 1996)

2.6 Evaluation and Validation

There is no optimal or best machine learning model for all data problems. This is because of the "no free lunch" theorem, which according to (Wolpert and Macready, 1997) states, that if there is an algorithm which performs well on a certain class of

problems, it necessarily "pays" for that with degraded performance in other problem classes. Therefore, the solutions of the different built models have to be compared. Generally, after the modeling one wants to get a better estimation of the true risk of the prediction of the built predictive model. A very common approach is to split the available data. One part represents the training set, which is used for the modeling and the other represents the test set, which is used for the process of validation. That is evaluating the success of the prediction model on unknown data. This is done with certain evaluation measures. As an example measure serves the error rate for classification problems. A rather old approach was to use the whole data for modeling as well as for the testing. This resulted in way too optimistic estimates of the out of sample error (Aggarwal, 2014). Today often 80% of the data is used for the training data and the remaining 20% for the validation data. Sometimes one doesn't have enough data to split it up like that. For this case a simple solution was derived, the so called cross validation (CV). The training data is split into K different, generally equal-sized folds. Then, for each fold k the model is trained on all the folds but the k'th. This is repeated for all K folds, where $k = 1...K$. The error averaged over all the folds is then computed. Often used values for K are 5 and 10. The choice of K represents a trade-off between the bias and the variance. Choosing a low K leads to more biased classifications. For high K values, there arises a stronger dependence on the training data, because of the increasing similarity of the training sets. Very commonly used is the 10-CV and in order to obtain reliable results it is repeated 10 times (10x10-CV). A special case is given when the number of folds K equals the data size, this is called Leave-one-out-CV. Another approach would be using all possible subsets of size P by leaving each time one of the P subsets out of the trainings phase, this is called leave-P-out-CV. Here more combinations are possible but it is computationally very costly. Other methods which use more combinations of training instances are repeated learning-testing methods, they are called Monte-Carlo-CV methods. A subset of the data is randomly chosen and used as training data, the rest as test data. This process is repeated multiple times.

There is also a sampling method with replacement and it is called the bootstrap method. Here, the instances for the training set are chosen with replacement, so the same observation can appear more than once in the training set. The probability in an ideally infinite sample space for an instance not to be picked would be 36.8% and

to be picked 63.2 %. Therefore it is called 0.632 bootstrap as this factor is applied to correct the probably too optimistic estimate of the performance. (Aggarwal, 2014) (Witten et al., 2016)

2.6.1 Model evaluation

It has to be distinguished between:

- *Metrics for Performance Evaluation*
  Here one needs to evaluate the performance of a model in such a way that the estimate is reliable.

- *Methods for Model Comparison*
  The problem that arises, lies in the comparison of the relative performance among competing models especially in the case where the size of the data sets can make the difference in accuracy not statistically significant. Consequently, for comparison issues a confidence interval should be established for accuracy.

Metrics for Performance Evaluation

Here, the focus lies rather on the predictive capability of a model than on other metrics such as the time required to build models or their scale-ability. The performance of a model is linked to the number of errors it produces. In this aspect it should be distinguished between the training error and the generalization error. The training error represents the number of occurring errors of the model in the training set while the generalization error is the error the model will have in records not previously seen. A good classification model should not only fit the training set well, but also accurately classify unseen records. When a model behaves very satisfactorily on the training set, it can be possible that it behaves badly in the unseen records. This certain situation is called over-fitting and should be avoided. As the chances of over-fitting increase with the complexity of the built model, normally the Occam's razor Rasmussen and Ghahramani (2001) principle is applied. Therefore, in the presence of two models with the same generalization error, normally the simpler one is chosen.

2.6.2 Accuracy Related Measures

As the goal of this thesis is to develop a discrete classifier, only measures which can be used for this case are listed here. The confusion matrix of the prediction results is the basis for the following measures (2.3).

		Predicted Class	
		1	0
Actual Class	1	TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
	0	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

Figure 2.3: The outcome of a 2-class prediction can be represented in a confusion matrix, where the reality is compared with the prediction.

Accuracy And Error Rate

The accuracy is defined as the ratio of correctly classified instances. In formula 2.7 it is shown how this measure is calculated.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

As can be seen in formula 2.8 the error rate is calculated by building the sum of the indicator variable values I, which equals 1 if the predicted label $\hat{y}_i$ is not equal to the true label y_i. The variable n represents the number of all observations.

$$errorrate = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2.8}$$

By knowing one of those two, either accuracy or error rate, the other one can be calculated easily. The relationship between them is presented in formula 2.9

$$accuracy = 1 - errorrate \tag{2.9}$$

Recall

Also known as sensitivity or True Positive Rate (TPR). It is the ratio of the number of true positives compared to all the really positive observations (2.10)

$$recall = \frac{TP}{TP + FN} \tag{2.10}$$

Precision

The precision is the number of true positives compared to the true and false positives (2.11).

$$precision = \frac{TP}{TP + FP} \tag{2.11}$$

Specificity

It is also called True Negative Rate (TNR) and is defined as the ratio of correctly as negative classified observations to all the really negative observations (2.12).

$$specificity = \frac{TN}{FP + TN} \tag{2.12}$$

Falarm

Also known as False Positive Rate (FPR) is the ratio of false positives to all the really negative observations (2.13).

$$recall = \frac{FP}{FP + TN} \tag{2.13}$$

F_1-Score

Represents the harmonic mean of the precision and the recall. It is calculated as shown in 2.14.

$$F_1 = \frac{precision \cdot recall}{precision + recall} \tag{2.14}$$

AUC

The area under the Receiver Operating Characteristic (ROC) curve can also be used as a measure of classifier performance and is in short called AUC. The ROC curve is drawn with the sensitivity on the ordinate and 1-specificity on the abcissa. It is used to determine a suitable operating point with regard to the trade-off between sensitivity (benefits) and specificity (costs). When for a classifier its parameters (e.g. threshold) are varied, different points for the ROC curve can be obtained.

In case of binary classification the ROC curve represents a trapezoid built by the points $(0, 0)$, $(1 - specificity, sensitivity)$, $(1, 0)$ and $(1, 1)$. The area of this trapezoid is the AUC.

3. Related Work

The main focus of this thesis lies on demonstrating how a data mining approach for health data analytics, more specifically for frailty data, can be helpful.

Consequently, it will be reviewed in what follows, the existing work of the literature related to: i) data mining in the medical domain, ii) CRISP-DM in the medical domain, iii) challenges of EHR analysis, iv) common analysis techniques in the health domain v) frailty.

3.1 Data Mining In The Medical Domain

In Bellazzi and Zupan (2008) a review can be found, in which current issues and guidelines in predictive data mining in clinical medicine are discussed. The authors conclude that predictive data mining is becoming an important instrument for the scientific community and clinical practitioners in the field of medicine. Bellazzi et al. further state that the main issues regarding these methods should be understood and the application of standardized procedures for their deployment should be made obligatory. The combination of clinical, molecular and genomic data has provided a new push to the field, but apart from that also a new group of problems which need to be addressed promptly.

The fact that clinical data collections enable data mining in order to perform retrospective analysis, which may provide new opportunities to better understand clinical processes, is stated in Bellazzi et al. (2011). Further, the molecular data holds the potential to offer insights on single patients, therefore changing decision-making strategies. Thus, it seems predictive data mining will be a strong ally for the transformation of medicine from population-based to personalized practice. Bellazzi et al. (2011) concludes that for this purpose the use of methods that are able to work

with temporal data is key, as well as the development of new data mining tools, which are able to combine data and knowledge in a framework. Thereby derived clinical models should be massively statistically evaluated.

In Prokosch et al. (2009) an overview of the various approaches for reusing the electronic medical records (EMRs) for clinical research is presented and further, published concepts and possible solutions are illustrated. The three following challenges were presented: establishing comprehensive clinical data warehouses, establishing professional IT infrastructure applications supporting clinical trial data capture and the integration of medical record systems and clinical trial databases(Prokosch et al., 2009). He especially points out the need for the integration of data repositories in clinical research projects which are deployed while the documentation of routinely done clinical care is done. Prokosch et al. further states that regulatory requirements, data privacy issues and data standards still remain an issue in this field today.

Haux (2010) stated that medical informatics as a discipline is still young and forms, being a cross-sectional discipline, the basis for medicine and health care. Therefore, there prevails a high responsibility for the people who are working in the field of medical informatics, in terms of improving the current health care system. Further, this imposes the mission for practicing innovative research in the different related fields. Haux further states that health care is continuously changing because the underlying science and practice of health are also in a continuous transformation. The field of medical informatics upholds an important role for this manner and is strongly affected by these changes.

(Ramakrishnan et al., 2010) noted that initial efforts in the area of mining in electronic health records (EHRs) are not likely to lead to serious pioneering insights, but there are a lot of opportunities in terms of improvement of delivery, efficiency and effectiveness of health care. At the moment the research is focused on health system integration, reducing medical errors, and providing reliable support to medical stuff. A vast amount of opportunities lies in the data mining and computer-aided decision making. However, Ramakrishnan et al. further state that we should be cautious not adopting the EHRs too fast, as they potentially delay the urgently needed standardization of data.

Jaspers et al. (2011) synthesised the literature on clinical decision support sys-

tems' (CDSSs) impact on health care practitioner's performance and patient outcomes. The authors analysed high-quality systemic reviews on CDSSs in hospitals. They found evidence in more than half of the studies that CDSSs significantly impact practitioner's performance. In more than one quarter of the studies, evidence was reported that CDSSs impacted patient outcomes in a positive way. Jaspers et al. conclude that only few studies were able to present benefits on patient outcomes and that this might be based on too small sample sizes or too short periods of time to reveal important effects. There exists no significant evidence that CDSSs improve the performance of the health care providers regarding the ordering of drugs and preventive care reminder systems. They further state that this could be explained by the lack of available patient data which the CDSS would require at the time the clinician is about to make a decision.

In the systematic review done by Bright et al. (2012), they evaluate the effect of CDSSs on clinical outcomes, workload, efficiency, patient satisfaction, costs and provider use and implementation. Investigators screened reports and identified 148 randomized controlled trials. Bright et al. conclude that both, locally and commercially developed CDSSs have shown to improve health care process measures but they found no sufficient evidence for clinical, economic, workload, and efficiency outcomes.

Data mining in electronic health records (EHRs) has a high potential for revealing not known disease correlations. Nevertheless, today there are many obstacles, such as ethical, legal and technical issues (Jensen et al., 2012). Despite the high potential, which EHRs could embrace in the data mining process, possibly resulting in a high performance predictive model, there are some limiting factors because of the data. Most provided databases are disorganized and the provided formats are often incompatible. This makes the data harder available for researchers. Further, Jensen et al. states that phenotypic manifestations are often not sufficiently covered in the data, because broad disease categories are used. Therefore, there is clearly the need to include detailed phenotypes which better cover the underlying comorbidities.

In the literature survey done by Yoo et al. (2012) the authors say that data mining in healthcare and biomedicine is still a relatively new concept which emerged in the middle of the 90s and provides novel and also deep insights. Further, it can potentially facilitate understanding of enormous biomedical data sets. This

work contains an introduction in how data mining technologies have been used for various purposes (prediction of health insurance fraud, under-diagnosed-patients, health care costs, disease prognosis/diagnosis, the length of stay in the hospital, detection of patterns in order to discover relationships between health conditions and disease and relationships between diseases and between drugs). The authors conclude that the requirement of parameter configuration of the mining algorithms and the quality of the patient data still remains a problem. Yoo et al. further state that an ideal data mining package should be more intelligent and be able to support data pre-processing and selection and should also fully automate the knowledge discovery process.

Recently, in Holzinger et al. (2014) authors state in their review that we are at the beginning of the era of data intensive life sciences, which brings many problems but also many potential research directions. They see the challenge in building a sophisticated framework which allows domain experts to interact with their data sets, without the need for prior training in mathematics or computational sciences. Holzinger et al. further suggest as solution for problem solving, to combine the individual advantages of humans and computers in order to obtain better results.

Ohno-Machado et al. (2015) claim that not much is said about the readiness of EHRs for data analyses. The existence of this data is often equated with standardized high-quality data, which can be used for fruitful data analyses leading to the discovery of "gold nuggets", which represent patterns of interest. However, the difficulties of preparing such data are still enormous. Ohno-Machado et al. further stated that bringing together data from different health systems is also difficult. The authors conclude that a lot has to be done regarding EHRs before they can be used for sophisticated analyses and decision-support applications.

In Goldstein et al. (2016) an evaluation on the current state of EHR based risk prediction modelling is presented, which was done via a systematic review of clinical prediction studies using EHR data. They searched PubMed for relevant articles in the years 2009 to 2014. In total 107 articles were identified. The found studies were very large in general with a median sample size of 26100. The authors claim, that the studies did not make full use of the EHR data as they in general did not make use of longitudinal information and integrated relative few predictors (median = 27 features). Not even half of the studies were multicenter and only 26 of them

performed validation across sites. Appearing biases in the data were usually not fully addressed, especially missing data or loss to follow-up. The average c-statistics of the outcomes are: mortality (0.84), clinical prediction (0.83), hospitalization (0.71), and service utilization (0.71). The authors concluded that EHRs present many challenges as well as opportunities for prediction modelling and that there is a great potential for improving the design of such studies.

3.2 CRISP-DM in the Medical Domain

The standard data mining process, which seems to have a high overall acceptance, seems to be CRISP-DM (Bellazzi and Zupan, 2008). It should not be seen as a precise guideline on what specific techniques to use, but more as a tool that gives an overall structure. The used methods themselves strongly depend on the specific problem domain. Predictive data mining with medical data is such a domain. As the aim in this domain is to build a model that is stable and reliable, important questions have to be answered which possibly can be done via data mining (Bellazzi and Zupan, 2008). According to Bellazzi and Zupan the following questions are of importance in the "business and data understanding" phase:

1. Are the available features sufficient in terms of predictiveness, so that a high performance model can be build?

2. Which features are the most predictive? Which of them have to be included in the predictive model?

3. What kind of relationship is there between feature and target variable?

4. Can there be a relationship or combination of interest found between the features? Is it possible to derive new features, possibly more predictive ones, from the original variables?

So the author further suggests, in order to evaluate question *1.*, to define some measures of success. Therefore, the statistics for evaluation have to be chosen before proceeding further. This could also be beneficial in terms of receiving a less biased evaluation of the results. For the remaining questions certain techniques like

feature ranking, selection and constructive induction can be helpful to find the most important features and could also help in the task of forming new ones (feature extraction).

Bellazzi and Zupan (2008) further states that following questions should be clear and answered, prior to the actual data mining:

1. Transparency of the model: should the model be interpretable for users? For example the generation of a set of rules.

2. Offering explanation: should the prediction model offer explanations for decisions?

3. Probabilities of outcomes and confidence intervals: can and should they be provided?

4. Domain-expert knowledge: is it available and can it be integrated into the models?

Regarding these questions, it seems to be the current practice that there is a "black box". The user often has no insight into the decision making process and also no information about the certainty of the decision. An example therefore are neural networks, their inner working is hard to understand because they work in a quite complex manner. Therefore, much simpler techniques like the naive Bayes classifier should be considered, as they behave also very well and additionally provide explanations and insights for the decision making process. Also logistic regression seems to be a simple and powerful tool, which further should be considered as "baseline" for the comparison with other models. (Bellazzi and Zupan, 2008)

Niaksu (2015) deals with the barriers of the practical application of CRISP-DM in the medical domain: technology, interdisciplinary communication, ethics and protection of patient data. He also focuses on well-known problems of medical data (inaccuracy, fragmentation). He therefore derived the CRISP-MED-DM model, which addresses the challenges and issues of the CRISP-DM reference model in the medical domain and introduces 38 new generic tasks as extension of the model.

3.3 Challenges of EHR Analysis

In this subsection we will review some of the work of literature in which data mining has been applied to obtain patterns in different medical diseases and where based on that, predictive models were built.

3.3.1 Cancer

Delen et al. (2005) compared 3 different machine learning models regarding their ability to predict breast cancer survival. They used a data set with more than 400000 cases and 72 features. With regard to the obtained models themselves, despite their high accuracy (ANN: 93.6%, C5: 91.2% and LR: 89.2%), Delen et al. (2005) states that they should be looked at with caution. As they may be valuable tools, the following has to be considered in the model development:

1. All clinical relevant variables should be included.
2. Testing on an independent sample should be done.
3. It has to be understandable for medical professionals.

Botsis et al. (2010) discuss issues when working with EHRs. They worked with EHR data of a cohort of pancreatic cancer patients collected over 10 years. Botsis et al. report that incompleteness was the main problem regarding data quality, followed by inaccuracy and inconsistency. They present the manifestations of these problems and discuss further strategies using new computational technologies to avoid or solve these issues. The authors state that better or automatic data validation tools and more flexible data presentation methods should be developed. Effective strategies should be collected and case studies pointing out the best practices should be provided (Botsis et al., 2010).

Gupta et al. (2014) demonstrate in their retrospective single-centre study, that machine learning (ML) applied to information from a disease-specific database and the electronic administrative record (EAR) is capable of producing a satisfactorily performing predictive model for clinical outcomes. They claim that their study is the first using ML techniques on this data for cancer survival prediction. They built

one prognosis model for all cancers and improved the accuracy on rare cancers. The data set contained 869 patients and was used by Gupta et al. to predict survival at 6,12, and 24 months. They achieved to obtain AUCs ranging from 0.757 to 0.997 for 6 months, AUCs from 0.689 to 0.988 for 12 months and AUCs from 0.713 to 0.973 for 24 months.

Kop et al. (2015) compared 3 different machine learning models against the traditional logistic regression model in the prediction of colorectal cancer (CRC). They try to point out the benefit of using advanced data mining techniques in this domain and to generate a better performing predictive model than the ones suggested by the literature at the moment. The used data set contained more than 200000 observations and a vast amount of features regarding doctor consults, drug prescriptions, specialist referrals, comorbidity and lab test outcomes. The data set was divided into temporal, non-temporal, knowledge-driven (known predictive features) subsets and a subset solely consisting of the features age and gender as benchmark. Kop et al. then built models with SVM, CART and RF. The RF algorithm could outperform the existing solution using the LR model. For example, the obtained AUC for non-temporal data was 0.883 for the RF model compared to 0.792 in the traditional LR model. Further, different best-performing predictors were discovered. Due to the low number of patients with CRC, there still is an uncertainty regarding this results. However, the authors state to may have found new predictors for CRC and that their results should be validated in future research and other data sets. Kop et al. further concluded, that state-of-the-art data mining techniques lead to better performing predictive models than currently available solutions for this problem, described in the literature.

3.3.2 Heart Disease

Palaniappan and Awang (2008) presented in their paper their developed prototype called Intelligent Heart Disease Prediction System (IHDPS). This web-based prototype makes use of the classifiers DT, NB and ANN, where each has its unique strengths depending on the goal. The authors claim that IHDPS can answer complex "what if" queries. With the use of medical profiles like blood pressure, gender, age and blood glucose it is able to predict the probability of patients getting a heart

disease. Palaniappan and Awang claim that it enables the discovery of significant knowledge, e.g. relationship between attributes related to cardiac disease. Overall, Naïve Bayes was found to be the most effective model in terms of predicting heart disease, followed by ANN and DT.

Kurt et al. (2008) compared the performances of different classification techniques (LR, CART, multi-layer perceptron MLP, radial basis functions RBF and self-organizing feature maps SOFM) regarding their capability of predicting the presence of coronary artery disease (CAD). The performance was assessed using the ROC curve and the area under it (AUC), Hierarchical Cluster Analysis (HCA), and Multidimensional Scaling (MDS). The obtained AUC results are 0.783, 0.753, 0.745, 0.721, and 0.675 for MLP, LR, CART, RBF, and SOFM. Kurt et al. concluded that MLP appears to be the best technique to predict CAD in the given data set.

Oztekin et al. (2009) tried in their study to improve the prediction of outcomes following combined heart–lung transplantation. They had a dataset with more than 16000 cases and 283 features. Oztekin et al. developed ML-based predictive models and extracted the best predictors. They further applied three different feature selection methods, the first one is based on ML techniques, the second one is based on literature-review-defined features and the third one is based on common sense interaction variables. A consolidated subset of features was generated and used to develop Cox regression models. Two multi-imputed data sets were used and the resulting accuracy in them (10-fold-cross-validated) was in the range of 79-86% for ANN, 78-86% for LR and 71-79% for DT. The authors concluded that their integrated data mining methodology using Cox hazard models performs better in terms of prediction of graft survival, using different variables than the conventional approaches.

In Srinivas et al. (2010) authors claim that the field of health care is perceived as being rich in information yet poor in knowledge. They further claim that there is a lack of effective analysis tools for the purpose of exploring and discovering underlying relationships and trends in the data. Srinivas et al. examined the capability of classification algorithms like rule based algorithms, DT, NB and ANN for working with a vast amount of health care data in order to predict heart attacks. They used the One Dependency Augmented Naive Bayes Classifier (ODANB) and the Naive Credal Classifier 2 (NCC2) for the pre-processing of the data and effective decision

making. They predicted combinations of several target attributes. The common NB classifier performed with an accuracy in the range of 83.7%-84.14% in all data sets and overall better than the other classifiers.

Wu et al. (2010) built a model for the detection of heart failure more than 6 months before the diagnosis using machine learning techniques applied to EHRs. The most parsimonious model was obtained by using logistic regression with model selection based on the Bayesian information criterion. 10 variables were selected at average, while a high AUC was maintained. The heart failure could be predicted 6 months before the diagnosis with an AUC of 0.76, using LR and boosting. SVMs performed very poorly, probably because of the imbalance of the data.

The objective of the work done by Anbarasi et al. (2010) was to create a more accurate prediction model for the presence of heart disease using a reduced number of features. They used a genetic algorithm to find out the most predictive features, with the goal to indirectly reduce the number of tests which are needed from the patients. The reduction was possible and instead of 13 features, just 6 remained. Then three classifiers (NB, DT, classification by clustering) were used to perform a prediction with the same accuracy as obtained before the feature reduction. Overall, the DT classifier outperformed the others.

In Kumari and Godara (2011) authors analyzed different data mining classification techniques for cardiovascular disease prediction. They compared the methods on the basis of different performance measures (sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate). Following classifiers were compared: RIPPER classifier, Decision Tree, ANN and SVM. The obtained accuracy for RIPPER, Decision Tree, ANN and SVM was 81.08%, 79.05%, 80.06% and 84.12% respectively. Kumari and Godara therefore concluded that the SVM algorithm is capable of predicting cardiovascular disease with the highest accuracy and further shows the least error rate.

In Soni et al. (2011) a survey of currently used techniques in knowledge discovery in databases using data mining techniques in the medical domain, with focus on heart disease prediction is presented. Their findings obtained through conducting different tests were that DT outperforms almost always all the other applied methods and that the accuracy of DT and the NB classifier can be further improved by applying genetic algorithm feature selection.

Weiss et al. (2012) applied in their study two statistical relational learning (SRL) algorithms in order to predict primary myocardial infarction. They used EHR data using a subset of known risk factors as features and selected a cohort of 1153 observations. Weiss et al. further showed that relational functional gradient boosting (RFGB) outperformed all the other considered methods and that their methods therefore are capable of augmenting current epidemiological practices.

Shouman et al. (2012) identified the gaps in the research on heart disease diagnosis and treatment. They further propose a model to close those gaps. This was done in order to be able to discover if the application of data mining methods to the heart disease treatment data is capable of providing as a reliable performance as the one achieved in the diagnosis of heart disease.

Sun et al. (2012) presented an approach for enhancing known knowledge-based risk factors with complementary risk factors derived from EHR data, in order to obtain a well performing prediction model for heart failure. They used a sparse regression model with regularization terms which corresponded to knowledge and data-driven risk-factors. The EHRs consisted of 4644 heart failure cases and 45981 controls. Sun et al. were able to identify risk factors which were not known as such and they were therefore able to better predict the onset of heart failure. The obtained model performed better with those new factors than without (the AUC improved by over 20%) and additionally, these factors were confirmed as clinically meaningful by a cardiologist.

Eapen et al. (2013) tried to derive and validate prediction models for assessing the risk of 30-day re-hospitalization and mortality in older heart failure patients using EHRs. A comparison of patients which were classified as low-risk or high-risk patients showed odds of death of higher value (odds ratio: 8.82) and also higher odds of re-hospitalization (odd ratio: 1.99) and death/re-hospitalization (odds ratio: 2.95). Their built mortality model, based on a logistic regression model, showed overall a good discrimination of the risk groups.

3.3.3 Intensive Care Unit (ICU)

Calvert et al. (2016) developed and evaluated an algorithm which makes a prediction of patient mortality in the ICU with a higher accuracy than current systems, using

the relationship between the clinical features from the EHR. The algorithm, called AutoTriage, uses 8 features to assess the patients' 12h mortality with a score. Their algorithm yielded an AUC of 0.88, a sensitivity of 80% and a specifictiy of 81%, with a diagnostic odds ratio of 16.26. Calvert et al. therefore conclude that their solution provides an improvement with regard to specificity and sensitivity in patient mortality prediction over current solutions.

3.3.4 Admissions And Re-admissions

Futoma et al. (2015) described and compared in their work many predictive models for prediction of early hospital re-admissions. Some of them have never been applied to this area and clearly outperform traditionally used regression methods. The data set contains 3.3 million observations and 12 thousand features. NN consistently had better AUC values (between 0.638 - 0.734) in all data sets compared to penalized logistic regression (PLR).

3.3.5 Diabetes

Mani et al. (2012) used machine learning techniques combined with EMR data for type 2 diabetes risk forecasting. They build a model to assess the risk of the development of this disease between 6 months and one year later. Mani et al. concluded that making this prediction is feasible. They achieved to obtain an AUC greater than 0.8 in the best model (RF). RF had the best overall performance but in terms of human-understandability a decision tree model such as CART seems to be far more comprehensible.

3.3.6 Adverse Drug Events

Karlsson et al. (2013) investigated the use of machine learning classifiers in order to predict adverse drug events using electronic patient records (EPRs). As features they used age, gender, diagnoses and drugs. Some predictive models were built and an evaluation was done using different algorithms and subsets of features. The highest achieved AUC was 0.87 (RF). The RF algorithm outperformed the rule learner algorithm in all data sets.

3.4 Common Analysis Techniques In The Health Domain

Saeys et al. (2007) reviewed different feature selection techniques used in bioinformatics, as those techniques have become an important necessity. They present the different possibilities of feature selection and provide a basic taxonomy. Further, they discuss their use, variety and the potential in different fields of bioinformatics.

Saeys et al. (2012) evaluated several feature extraction/ranking methods derived from ML approaches. They performed experiments on synthetic and real world data. They concluded that methods using conditional error rates (CER) and mProbes are highly selective and do not select irrelevant features in most cases. A further conclusion they made is that using the performance of an model as a criterion for feature selection seems to be counter-productive.

Herland et al. (2014) reviewed the recent research, using tools and approaches from the field of "big data" for the analysis of different levels of health data (molecular, tissue, patient and population data). They also addressed questions regarding human-scale biology, clinical-scale and epidemic scale. Further they analyzed possible future work. As medicine is such a complex field they propose that research has to be done on all the levels in order to retrieve the most knowledge.

Jacobson and Dalianis (2016) applied deep learning techniques to EHRs in order to predict infections which are associated with the health care. They implemented a network of stacked sparse auto encoders and a network of stacked restricted Boltzmann machines. The best performance showed the Boltzmann machines which achieved a precision of 0.79 and a recall of 0.88.

Cheng et al. (2016) proposed in their paper a deep learning approach in order to phenotype using EHRs. They transformed the EHR for every patient in a time/event matrix. Then a convolutional neural network with 4 layers was built for the purpose of predicting and extracting phenotypes. They also investigated different temporal fusion mechanisms in the model. Then the model was validated on a real world EHR data set with the goal of predicting chronic diseases (chronic obstructive pulmonary disease (COPD): highest AUC 0.74, congestive heart failure (CHF): highest AUC 0.77).

Perer et al. (2015) utilized EMR data in order to extract common patterns of medical events such as diagnoses and treatments and they further explored how these patterns are related to the patient outcome. Their so called Care Pathway Explorer consist of a mining algorithm adapted to real-world patient data and a visualization tool with an interactive interface consisting of an overview and flow visualizations. Perer et al. used the system to perform an analysis on cohorts of hyperlipidemic patients with hypertension and diabetes pre-conditions. They further demonstrated the clinical relevance of the found patterns. Some of these findings correspond to already published knowledge and another part was prior to this unknown to the scientific medical community. Therefore, they concluded that their solution enables data-driven insights into the patient data.

3.5 Frailty

Different frailty models are described in the book "The Frailty Model" by Duchateau and Janssen (2007). The authors note that survival analysis techniques have been used in a variety of different disciplines, including biology, medicine and engineering. Recently there were more attempts made to work with more complex survival data, and models in this direction were developed and deployed. Duchateau and Janssen focus in their work on frailty models (parametric, semi-parametric) and further on similarities and differences between frailty and copula models. Frailty models represent hazard models with a multiplicative frailty factor: this factor determines how frail observations in a specific cluster are. These models are conditional models. The frailty factor itself is random, which induces the need to specify a frailty distribution in the model. A variety of distributions were studied in this work. Duchateau and Janssen discussed the current methods and demonstrated on examples how obtained results from statistical analysis are to be interpreted. All this with the aim to make the techniques more available to practitioners.

Swindell et al. (2010) tried to identify the predictors of long-term survival in older feminine patients (65-69 years old) and to develop a model using data from the Study of Osteoporotic Fractures (SOF). The data set contained 4097 observations (the youngest of the SOF cohort) and 377 phenotypic features. These features were analysed regarding their predictability regarding long-term (19-year) survival. The

feature representing the visual contrast sensitivity score appeared in the top 5 of the best predictors. Swindell et al. derived a 13-feature model, which shows a good performance (mean AUC: 0.673). The used features consisted of a measure of physical function, smoking behaviour, presence of diabetes, self-reported health, contrast sensitivity and functional status indices which reflect the sum of daily living impairments. A follow-up was done on average 20 years later. The output of the model (a multivariate index) was compared to multiple outcomes (test of cognitive function, geriatric depression, number of daily living impairments and grip strength). They state that their index needs further validation on other cohorts but the results suggest that components of their index are able to characterize the clinical presentation of "healthy aging". The 13-variable index for predicting long-term survival is given by a Cox PH model (mean C = 0.673 ± 0.001). The through forward search identified 13 features are listed here:

- Number of step-ups completed in 10 seconds
- Smoking: indicator with value 1 if subject is a current smoker
- Diabetes: indicator with value 1 if a subject is not diabetic
- Age at baseline examination (65 - 69 for all subjects)
- Response to Question: How is your health compared to others your age? (categories: excellent, good, fair, poor, very poor)
- Smoking: indicator with value 1 if subject is a past smoker
- Contrast sensitivity score, average of high and low spatial frequencies
- Pulse Lying Down (beats/60 seconds)
- Hypertension: indicator with value 1 if systolic blood pressure exceeds 160, diastolic blood pressure exceeds 90, or if subject used thiazide
- Past thiazide use: indicator variable with value 1 if the subject has previously used thiazide
- Height change since the age of 25 (self-reported at baseline exam)

- Participant's clinic throughout the study: indicator with value 1 if subject has attended clinic
- Marriage: indicator with value 1 if subject was married at the time of the baseline examination

A study done by Baylis et al. (2013) investigated the relationship between immune-endocrine axis and frailty and also mortality after 10 years in females and males with an age between 65 and 70 years. They worked on 254 observations of the Hertfordshire Ageing Study at baseline and also with the 10-year follow-up data. The baseline data consists of a health questionnaire data and immune-endocrine blood parameters. In the follow-up the Fried score for frailty (Fried et al., 2001) was calculated and mortality was assessed. Their findings were that higher baseline levels of white blood cell counts, lower levels of dehydroepiandosterone sulphate (DHEAS) and higher cortisol to DHEAS ratio could be related to a higher probability of frailty at the follow-up. The baseline white blood cell counts and the cortisol to DHEAS ratio appeared to be significantly different in observations which went on to be frail at the 10 year follow-up. Baylis et al. note that they have presented the first evidence that certain immune-endocrine biomarkers are related to the probability of frailty and mortality over a time of 10 years. They suggest a screening programme at the ages between 60 and 70 years in order to identify individuals with an increased likelihood of becoming frail, who clearly would benefit from an early on treatment in order to prevent the onset of the syndrome.

4. Materials and Methods

In this section the main methods, techniques and technologies that have been used in the scope of this thesis to fulfill the final goal, will be reviewed.

4.1 CRISP-DM

According to the in 2014 conducted web survey of Gregory the most widely used process model in knowledge discovery nowadays is the CRISP-DM model.

CRISP-DM was devised in 1996 and a consortium was formed, which obtained funding by the EU. The acronym was extracted out of "CRoss-Industry Standard Process for Data Mining". This standard process was supposed to be an industry-tool and also neutral in respect to application. In 1999 the first draft was completed and one year later a step-by-step data mining guide called "CRISP-DM 1.0" was published.

Since then it is the *de facto* standard methodology in the data mining community. It represents a hierarchical process model which contains a set of tasks. In total these are depicted at four levels of abstraction. Going from general to specific.
The methodology is divided in the reference model and the user guide. The first one represents an overview of the whole data mining process, consisting of every intermediate phase, outputs and the tasks (see figure 4.1). The user guide contains information about each phase and its tasks and leads through the data mining project.

The reference model shows the life cycle of the DM-project. Containing the tasks, phases and the underlying relationships, which strongly depend on the main goal, the user and for the most part on the data. All in all there are 6 phases, where the sequence is not fixed. Changing direction and eventually jumping to another

former phase is a requirement. The process is outcome oriented and depends on it. The result of the process dictates the next step. The arrows and their direction are showing the most relevant dependencies. The cyclical nature of the whole data mining process is represented by the outer circle. When a solution is deployed, this does not necessarily mean that the process is over. The solution can point out new questions and start a new process, which will benefit from the gained experience of the previous one.
Here, using the description given by Chapman et al. (2000), a short summary of the phases is presented:

Business understanding In the beginning one must comprehend the goals from a business perspective. After that, the arisen question or the obtained problem has to be translated into a well described data mining task. Then a preliminary plan has to be created in order to accomplish the goals.

Data understanding This phase begins with the initial collection of data. Further, getting to know the data is the main task. This is achieved by identifying quality problems and discovering or detecting subsets of interest. Then, with the obtained insights, a hypothesis can be formed.

Data preparation The objective of this phase is to create the final data set. The main activities here are the parameter selection, the cleaning of the data and the transformation of the data. This step depends on the next one and it is therefore likely that it has to be repeated or adapted.
The majority of the time the data scientist spends with understanding and preparing the data. Both tasks are very important and highly underestimated. (Perlich, 2016)

Modeling Here different models are chosen and calibrated in order to work in the best way with the available data. Besides, the data needs to be in a certain shape depending on the modeling technique. Therefore, the necessity to visit the previous step could emerge.

Evaluation In this step reviewing the built model from a business perspective is the main objective. Further, a review of the course of action that led to this result is of importance. If all the business goals were realized in a satisfying

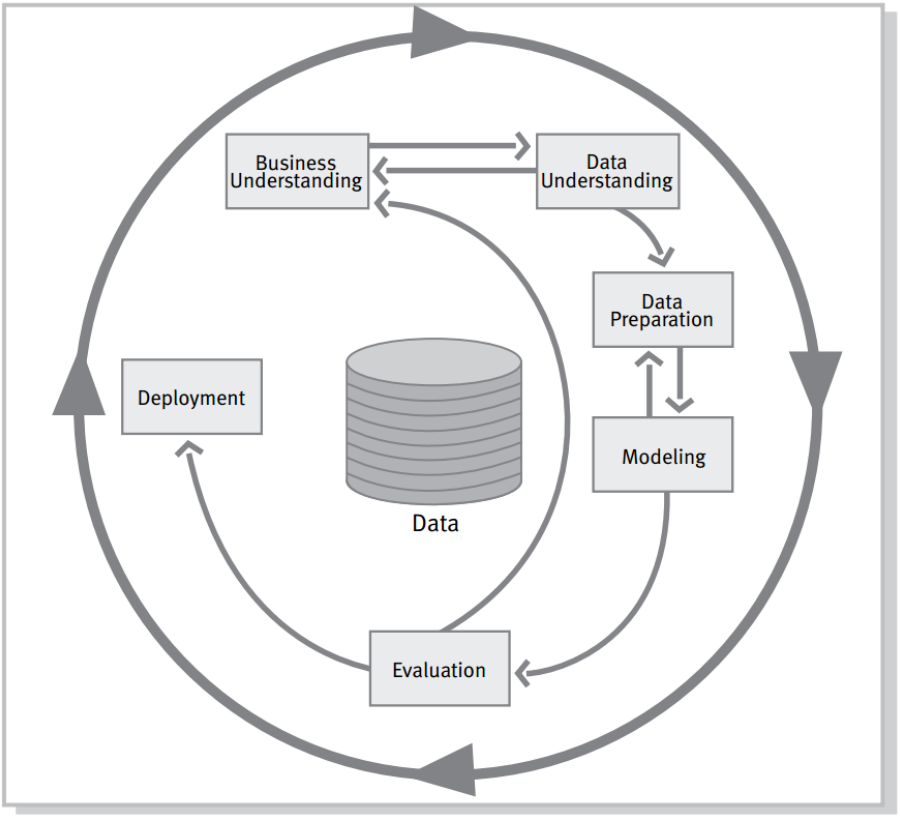


Figure 4.1: The CRISP-DM reference model. The 6 important phases in a data mining project and their relationships are depicted. (Chapman et al., 2000)

way, a decision on the application of the data mining results can be carried out.

Deployment Depending on the aim of the project, the resulting model or the obtained results have to be further processed for the customer in order to satisfy his needs. The range of work in this phase spreads from simply putting the results in a final report to creating a real time model which is deployed in the company.

4.2 R (programming language)

The programming language R is generally used for statistical computing and graphics. It represents a GNU Project and shows similarities to the S language, an environment created at the Bell Laboratories. Overall, R can be perceived as a different

implementation of S.
R is a tool which plays an important part in the area of statistics. Being a open source system, many different packages containing a variety of tools, methods and techniques are freely available. (The R Foundation, 2016)
Packages are also available for the sector of machine learning, data mining and multivariate statistics. Especially the package *e1071* created by the Department of Statistics - Probability Theory Group (Formerly: E1071) - placed at the Technical University of Vienna is one of the most used ones. This, according to Geethika Bhavya, who analyzed the most downloaded R packages from January to May 2015.

Also in the area of data mining and knowledge discovery R offers a vast amount of packages and implementations of different algorithms. More than 50 R packages were used within the scope of this thesis, the important ones are presented below.

4.2.1 Vizualisation

ggplot2

This today very popular package offers a sophisticated graphics language in order to create complex and elegant plots.

lattice

This package represents an improvement of the R standard graphics and enables the visualization of multivariate relationships.

4.2.2 Clustering

NbClust

The NbClust package provides 30 indices for determining the optimal number of clusters and proposes to the user the best clustering scheme. This is made possible by valuating the different results obtained by varying all combinations of the number of clusters, the distance measures and the clustering methods.

4.2.3 Imputation

mice

This package contains the implementation of Multiple Imputation (MI) using Fully Conditional Specification (FCS) also known as Multivariate Imputation by Chained Equations (MICE). This is a common technique for generating estimates to impute missing values by drawing from estimated conditional distributions of each variable given all the others (Shah et al., 2014). The package contains built-in imputation models for continuous data (predictive mean matching, normal), binary data (logistic regression), unordered categorical data (polytomous logistic regression) and ordered categorical data (proportional odds). The in the brackets mentioned methods are just examples, as there are many techniques available. Further, it is possible to impute continuous two-level data (normal model, pan, second-level variables). For each feature it is possible to build a customized imputation model. There is also the possibility to execute passive imputation, which can be used to keep consistency between the features. Additionally, many diagnostic plots are included, which allow an analysis of the quality of the imputations.(van Buuren et al., 2015)
Below a non-exhaustive list of specific features of the *mice* package can be found. It was taken from the paper about MICE by Buuren and Groothuis-Oudshoorn (2011):

- Column-wise specification of the imputation model.
- Arbitrary patterns of missing data.
- Passive imputation.
- Subset selection of predictors.
- Support of arbitrary complete-data methods.
- Support pooling various types of statistics.
- Diagnostics of imputations.
- Callable user-written imputation functions.

In a study called "comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome", performed by Ambler et

al. (2007), the mice imputations introduced the smallest amount of bias, the best coverage values and the best overall performance. Further, it outperformed the best hotdeck methods.

CALIBERrfimpute

This package contains the publicly available implementation of a random forest-based MICE algorithm from Shah et al. (2014). It was compared in two studies to parametric MICE settings. They used real world data (electronic health records) and came to the conclusion, that their implementation of random forest for imputing missing data, performs especially better in terms of conserving non-linear relationships. Both methods lead to unbiased estimates of (log) hazard-ratios, where the RF-implementation showed higher efficiency and the obtained confidence intervals appeared to be narrower. All in all, this method appeared to be quite suitable for the application on the data used in this thesis, as it outperformed the already well-working parametric MICE implementation.

Though, a mild weak-spot of their described method is, that it only has been validated in a few studies so far (for example theirs and one described in McEvoy et al. (2015)). Therefore, a generalization of the results should not be made too rashly. Anyhow, the data they used also consists of electronic health records and because of the described advantages in terms of conserving the inter-feature-relationships, this method was also used in the scope of this thesis.

4.2.4 Feature Selection

Boruta

This package contains a sophisticated feature selection algorithm, which uses a wrapper approach built around a random forest (Breiman, 2001) classifier. The random forest algorithm is already more explicitly described in 2.4.2. The term "Boruta" comes from the Slavic mythology and represents the name of the god of the forest. The algorithm is an enhancement of the already introduced idea to determine feature-relevance by doing a comparison of the relevance of real features and random probes (back then proposed as filter method). (Stoppiglia et al., 2003)

Here, a so called importance measure, which represents the loss of accuracy of the classifier caused by randomly performed feature permutations between objects, is used. Also the accuracy loss's average and standard deviation are calculated. Another importance measure is the Z score, which is calculated by the division of the average loss by its standard deviation. This measure isn't directly related to the statistical significance of the importance of the feature, because it is not normally distributed.(Rudnicki et al., 2006) Yet Boruta uses the Z score as importance measure anyway due to its ability to take into account fluctuations of the mean accuracy loss among the forest trees.

The R package *Boruta* was created by Kursa and Rudnicki (2010) and has already been successfully used in the scientific community. For example, it has been used to find powerful features for classifying different subtypes of pediatric patients with irritable bowel syndrome Saulnier et al. (2011). They achieved a classifying success rate of 98.5%. Further, Boruta was also successfully applied in a study to extract the most powerful features in terms of prediction, to discriminate between pregnant and non-pregnant participants (Aagaard et al., 2012).

Nevertheless, the authors of Kursa et al. (2010) have shown that random correlations of the data could potentially lead to the creation of dependencies between features, that are sufficiently strong to pass statistical tests of validity. Overall, they state that the importance of the feature in the machine learning method may rather be used as hint for the existences of a real relationship between features and not as proof. Thus, the obtained results should be examined with care and further analysis should be done.

4.2.5 Modeling

e1071 From this very widely used R package, the included implementations of support vector machines (SVM) and the naive Bayes (NB) classifier were used.

tree This package contains an implementation of Classification and Regression Trees (CART).

ipred This package contains Bagging for classification, regression and survival trees.

C50 This package was used for fitting classification tree models and rule-based

models using Quinlan's C5.0 algorithm.

randomForest This package contains the implementation of Breiman and Cutler's Random Forests (RF) for Classification and Regression (see section 2.4.2).

MASS Contains an implementation of linear discriminant analysis (LDA).

4.2.6 Evaluation

caret This package was used to calculate all the performance measures of the models, including accuracy, precision, sensitivity, specificity and the F_1-Score.

pROC This package contains different tools to calculate and visualize ROC curves and also to determine the AUC.

5. Results

The CRISP-DM model was used in this thesis and adapted accordingly, influenced by the suggestions of Niaksu (2015) and Bellazzi and Zupan (2008) described in section 3.1. CRISP-DM establishes the main tasks but does not establish a life cycle. Along the project development the different tasks are executed several times. In what follows the complete development is detailed. The document is structured according to the different main phases and the different tasks that are required to obtain the final goal. They will be exhaustively explained for each stage of the project.

5.1 Business Understanding

In this part the overall objectives and the data mining goals were determined. Further, the current situation was assessed and necessary activities were planned.

Business understanding is the stage of the project in which the main goal is defined and translated into data mining goals. As this research is framed as part of the FACET project, the main goal had to be aligned accordingly. In what follows the main goal is defined in detail. This definition has already been brought up in section 1.1.

5.1.1 Understanding of the Frailty Problem and Translation to Data Analytics

The goal of FACET is established as follows: *Now that people live longer, older adults need to live better and independently (i.e. without disability). Avoiding disability in older adults has a potential impact on over 13 million of EU citizens and an eco-*

nomic impact of 1,500 million euros per year, thus contributing to the achievement of both individual and social benefits. Consequently, the prevention of disability has become the most challenging concern for current Health Care providers. Disability cannot be reversed, but it is preceded, sometimes by several years, by a known frailty syndrome, which can be reversed, and thus prevented from worsening and its progression monitored. Frailty is characterized by a decreasing capacity to respond to demands, caused by diminishing functional reserve. The prevalence of frailty in people 65+ ranges from 7% to 16.3%, increasing with age, and it is the main risk factor for disability. Therefore, frailty assessment is a key tool for the prevention of disability by identification of people at risk.

The aim of the FACET platform is to provide an innovative solution for the assessment and follow-up of the functional status of elderly people in order to early detect frailty, to control its evolution and to prevent disability, by the integration of different proven technologies.

Therefore, an objective of this thesis is to perform analysis of the impact of different variables on the frailty of patients through data science tools, preparing the path for the alerts and the visualization of patterns that will be deployed in the service provided within the FACET project.

From the previous statement following data mining goals can be extracted:

1. Identification of risk/preventive factors regarding frailty, which can be used as predictors ("biomarkers").

2. Learning of accurate models for frailty prediction.

3. The validation of the models prior to deployment and the analysis of their suitability for predictive risk models.

5.2 Data Understanding

Data understanding is a paramount task of each data mining project development, which main goal is to understand the target data to be analysed very well. In the present research, the data was obtained within the scope of studying healthy aging

and the frailty syndrome. The study, called *Toledo Study for Healthy Aging*, is described by Garcia-Garcia et al. (2011) as follows.

The Toledo study is a population-based study conducted on 2,488 individuals aged 65 years and older. The study subjects were selected by a two-stage random sampling from the Toledo region. Institutionalized as well as community dwelling persons were selected. Data was gathered in 3 waves: first (2006 to 2009) information on social support, activities of daily living, comorbidity, physical activity, quality of life, depressive symptoms, and cognitive function was collected. Furthermore, anthropometric data and results of physical performance tests (walking speed, upper and lower extremities strength, and the stand-and-sit from a chair test) were collected and a blood sample was obtained. The diagnosis of the frailty syndrome was based on the Fried criteria (weakness, low speed, low physical activity, exhaustion, and weight loss)(Fried et al., 2001). In the second wave (2011-2013) and in the third wave (2015-2017), which is ongoing, additional parameters were added (urine parameters).
The patient data collection process in terms of time, number of patients and used parameters (set A and set B) can bee seen in figure 5.1. Here, UPM stands for "Universidad Politécnica de Madrid" and marks the data which was available in the scope of this thesis. Aber marks the data which was available for the Aberystwyth University. Their objective was retrieving biomarkers for the frailty syndrome using urinary data.

5.2.1 Definition of the Data Sets

From the aforementioned Toledo study a subset of data (in figure 5.1 marked with UPM) has been made available for this thesis. In particular a total of 474 anonymized electronic health records (EHRs) have been provided. Thereby, for each patient an EHR consisting of 284 parameters was provided. Further, a so called *codebook* was made available. It explains for each variable the meaning, the range and possible values. The *codebook* can be found in the annex A.2. The majority of attributes is from the first wave of the *Toledo Study for Healthy Aging* (2006-2009) and and only 21 come from the second study wave conducted in 2011-2013.

From the Toledo study a randomized sample was produced. It consists of 474

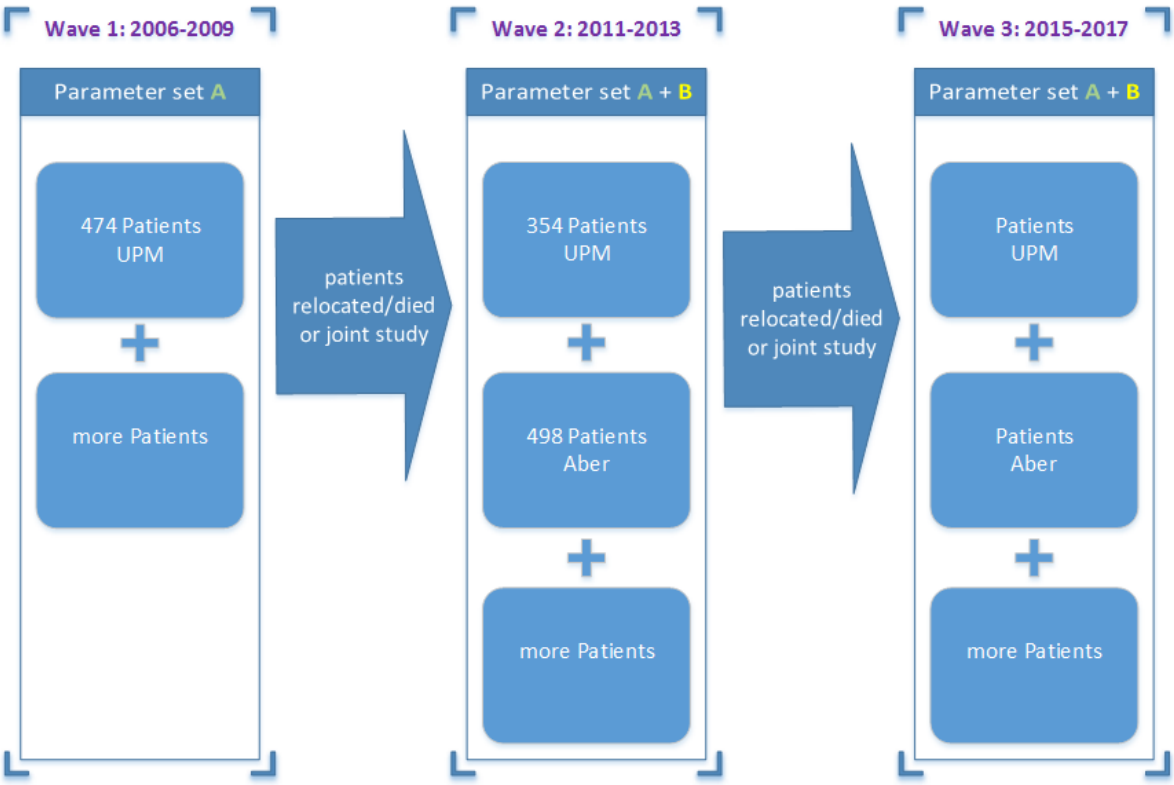


Figure 5.1: The diagram shows the evolution of the clinical data, which was collected at 3 different points in time. The number of patients changes as well as the available parameters (set A and B). UPM stands for "Universidad Politécnica de Madrid" and marks the data which was available in the scope of this thesis. Aber stands for "Aberystwyth University" and marks the data which was made available for them. They were mainly focused on finding biomarkers in the urinary data of the second wave.

patients, which are described by 284 attributes. This private and protected sample has been used in this thesis.

5.2.2 Definition of the Variables

As it has just been explained, patients are described by a set of 284 variables. One variable, the one representing the frailty stage (see description below), is the target variable for the predictive models. In the first stage the 283 remaining predictor variables were grouped according to their semantics into: i) demographic, ii) phenotype, iii) medication and iv) code features. The phenotype features then were

further split into physique, blood, cardio, disease, self reported disease, consumption and medical test attributes.
The medical test attributes were further divided into features corresponding to the Geriatric Depression Scale (GDS), Activities of Daily Living (ADL), Instrumental Activities of Daily Living (IADL), Mini-Mental-State-Examination (MMSE) and Mobility Scale (MS) attributes. In appendix A.1.1 the complete description of the variables can be found. Below you can find a short explanation for each medical test, which was carried out in the study:

Geriatric Depression Scale (GDS) This scale was created with the objective to obtain a reliable rating for depression in elderly. The applicant himself answers in the so called *short form* 15 different questions. Of those, 10 questions indicate the presence of depression when positively answered and the remaining 5 questions indicate the presence of depression when negatively answered. (Yesavage and Sheikh, 1986) (Yesavage et al., 1983a)

Activities of Daily Living (ADL) In this assessment also a questionnaire is used, which is answered by the patient. Here the goal is to estimate the patients' satisfaction in his daily activities, which contain hygiene, alimentation and independent access to necessities. There exist different variations of the ADL test, which differ regarding their contained number of questions. (Pincus et al., 1983)

Instrumental Activities of Daily Living (IADL) Like the ADL-test but mainly focused on instrumental activities. These include following daily tasks and responsibilities: food preparation, shopping, using the telephone, housekeeping, transportation, responsibility for own medications and the ability to handle finances. For each activity exist 3 to 5 questions, each yielding 0 or 1 point. The maximum for each category is 1 point. At the the end these points are summed up. This sum represents the IADL-Score with a range between 0 and 8. (Lawton and BRODY, 1970)

Mini-Mental-State-Examination (MMSE) The Mini-Mental-State-Examination represents standardized test for cognitive function or meausure of impaired thinking. The tested areas of cognitive function consist of orientation, regis-

tration, naming recall, calculation, writing, attention, repetition, comprehension, reading and drawing. The range of the result lies between total cognitive absence (0 points) and full cognitive function (30 points). (Folstein et al., 1975) (Cockrell and Folstein, 2002)

Mobility Score (MS) The MS questions belong to the Physical Activity Scale for the Elderly (PASE) questionnaire. They provide validated knowledge about the physical activity of the patients. Washburn et al. (1993)

Geriatric Depression Scale (GDS) The Geriatric Depression Scale (GDS) is a 30-item self-report assessment used to identify depression in the elderly people. It has been found to be a reliable and valid measure, which can be extracted from the GDS-questionnaire the patient himself has filled out. (Yesavage et al., 1983b)

Fried's Frailty Score This score corresponds to the score of frailty using Fried et al.'s Frailty Scale. The 5 used criteria are weight loss, exhaustion, physical activity, walk time and grip strength. Patients with no deficits in all criteria score 0, which means they are not frail. Those who have deficits in 1 criterion or 2 criteria are called intermediate frail or pre-frail (this term was used in this thesis). All higher scores lead to the classification frail. (Fried et al., 2001)

5.2.3 Data exploration and quality assessment

In what follows, performed tasks will be described in order to gain understanding of the data prior to modeling: i) data visualisation and analysis of values, ii) outlier detection, iii) ontology-guided PCA and iv) cluster analysis.

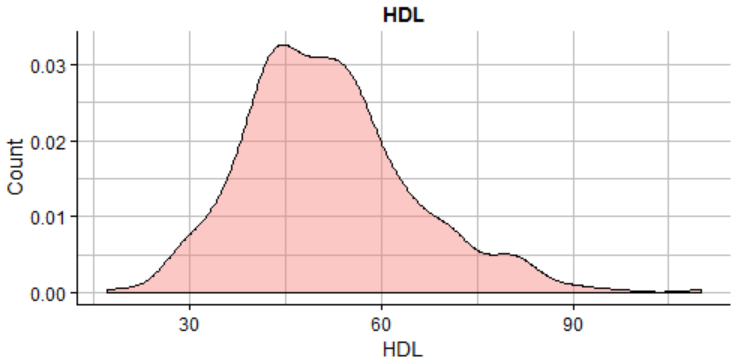
i) Data Visualisation and Analysis of Values

The retrieved data set was analysed using different statistical visualisation techniques like plotting the histogram, the kernel density function estimate and boxplots. Further, the values of each feature were inspected and compared to the values they should have according to the provided *codebook* (see annex A.2). Moreover, statistical measures were calculated and analyzed. The provided variables were divided according to their corresponding data type into continuous, categorical and

binary variables. Depending on this data type, different visualisations were realized and statistical measures calculated. For simplicity and clarity of the document only examples for each type of variable are presented. The description and analysis of each variable can be found in annex A.1.1 and annex A.1.2

Continuous Variables The variable HDL, which represents the measured content of high-density lipoprotein in mg/dL in the blood, serves as an example for continuous variables. The built description table is shown in table 5.1.

Table 5.1: Description of HDL

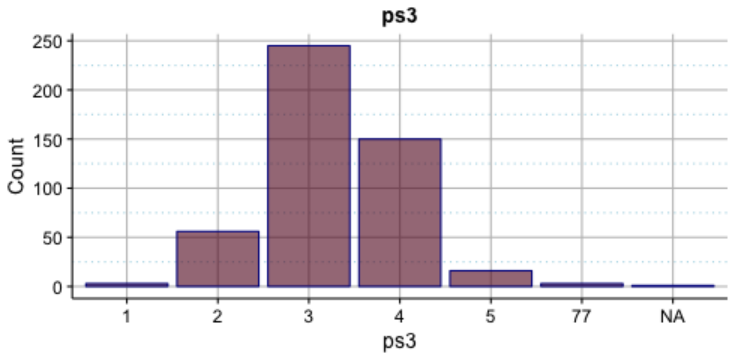
HDL								
Meaning	This feature gives numeric information about high-density lipoprotein (HDL) [mg/dL].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	17.00	110.00	51.67	50.00	176.16	13.27	43.00	9.07
Distribution								

In the first row the variable name HDL is presented, followed by a short description of the meaning in the second row. Relevant statistical measures like the sample minimum, maximum, average, median, variance σ^2, standard deviation σ and the number respectively the percentage of missing values are shown in the third row. In the last row a figure depicting the distribution of the values (kernel density estimate) is shown.

Categorical and Binary Variables The variable $ps3$, which gives categorical information about the current health status of the patient compared to other people with the same age in the view of the patient (question: "How would you judge your

health compared to other people of your age?"), serves as an example for categorical variables. The built description table is shown in table 5.2.

Table 5.2: Description of *ps*3

*ps*3	
Meaning	This feature gives categorical information about the current health status of the patient compared to other people with the same age in the view of the patient. Asked question: "How would you judge your health compared to other people of your age?"
Statistics	mode: 3.00; levels: 7.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Much worse 2: Sligthly worse 3: The same 4: Better 5: Much better 6: Undetermined 77: Not available

Once again, the first row contains the variable name, followed by a short description of the variable meaning in the second row. Relevant statistical measures like the mode (most frequent observation), the different levels and the number respectively the percentage of missing values are shown in the third row. The last row contains an explanation for each appearing level in the feature.

Binary variables were analysed in the same way and are also presented in this form.

ii) Outlier Detection and Missing Value Detection

Many binary and categorical features contained values like 77,88 and 99. For example, the feature $tab1$, which refers to the answer to the question "Have you smoked at least 100 cigarettes in your entire life?", contains aside from the valid values 1: "yes", 2: "no" and 3: "unknown" also 88 and 99. At the first glance they may appear as outliers but after further investigation, the statistician of the study stated that they have indeed a meaning. The significance of 88 is that the patient did not answer, 77 that he did not know how to answer and the meaning of 99 is that he did not want to answer. What they do have in common is the core significance that the patient did not answer and that therefore no information regarding the feature itself is available. One could claim that the reason why they did not answer (was not able, did not want to) also contains information which could be used, but investigations in this direction were not aim of this thesis. Anyhow, these values have to be treated differently as can be seen in section 5.3.4. The method of assigning special values was also used for the answers of questions like "For how many years did you smoke?" (the reply was a numeric value representing the number of years), where for a population having ages between 65 and 95, values like 77 are quite likely to appear. Now, when finding such a value, it is not clear if it stands actually for the value of 77 years, or if it has some other special meaning like 77: "could not remember".

Another issue stated by the doctors is that sometimes patients don't want to answer questions because they are simply not able to, this because of analphabetism. Therefore, also many values corresponding to these questions are missing. However, this issue was ignored for this investigation.

There are two features related to income, namely $Individualincome$ and $Householdincome$. The first one has 8.44% missing values and the second one 13.29%. The missingness could base upon the fact, that people with a high income as well as people with a relative low income, are more likely to not state their financial situation (this possibly out of shame or discretion).

There are 37 features where more than 60% is missing. 12 of those features are

follow up questions to a previous asked principal question. For example the feature $tab1$ contains the answers to the question "Have you smoked at least 100 cigarettes in your entire life?", when answered with 2 (which stands for no) the follow up question, represented by $tab1a$ ("If yes, Did you smoke cigarettes daily, occasionally, or not at all?"), has not been asked. So as a matter of fact, these values are not missing at random, but rather the question was not applicable for these observations.

Features representing codes and IDs of the hospital (see table A.13) do not contain relevant information with regard to frailty prediction, as they were created for organizational reasons and do not contain information regarding medical/phenotypic/demographic aspects.

iii) Ontology-Guided Principal Component Analysis

Once the variables had been explored, the following step was to try to find similarities and relationships of the predictors and the target variable in order to get insights that could help prior to the predictive analysis.

Approach The here used approach is based on the work from Wartner et al. (2016), which describes how to execute principal component analysis (PCA) within an ontology-guided data infrastructure for scientific exploratory purposes. The goal is to obtain indications of unsuspected relationships and similarities between the features by further including doctors in these analyses.

Description The PCA was used to reduce the high dimensional data set and to analyse the data in 2-dimensional plots. According to the doctor's recommendations, following variables were used: education status, income, BMI (self-derived, see section 5.3.5), Geriatric Depression Scale score, total comorbidities (self-derived, see section 5.3.5), Mobility Score (self-derived, see section 5.3.5), gender and polypharmacy. The here used term "recommended variables" refers to attributes which are scientifically proven to be related with frailty, or in suspicion to be related with it. These factors can be found in section 5.3.7
The first principal component (PC1) is the linear combination of the used subset of features that has maximum variance among all possible linear combinations. It therefore accounts for as much variation in the data as possible. In this case PC1

has a variance of 23.59%. The second principal component (PC2) is the linear combination of the used subset of features that accounts for as much of the remaining variation as possible. Given the constraint that the correlation between the first and second component is 0. In this case PC2 represents a variance of 11.98%. The third principal component, which is for obvious reasons not shown in the 2-dimensional plots, represents a variance of 9.12%. All following principal components of higher order have the same properties. They account for the remaining variation and are also not correlated with the other principal components. For this two-dimensional presentation the first two principal components were used. They make up 35.57% of the total variation of the data, which is sufficient in this case because the PCA is here only used as a visualization tool for exploration. The total variance shown in the 2-dimensional plot, made up by PC1 and PC2, is also a measure for the report quality (Wartner et al., 2016).

Results It can be seen in the resulting PCA plot 5.2 that non-frail patients (green), pre-frail patients (yellow) and frail patients (red) appear in overlapping areas.

In figure 5.3 the loadings (the eigenvectors multiplied by the square root of the corresponding eigenvalues; they do also contain the variance along the principal components), themselves were plotted in this 2-dimensional principal component plane. (Wartner et al., 2016)

Interpretation Closeness between the features in the PCA plot can be seen as indicator that their might be a relationship. As can be seen in figure 5.3 a high mobility score (MS_score) seems to have a strong relationship to the ability to sit down and up in a chair ($silla$), and they together seem to correlate with the physical activity score ($pasetotal$). This observation is not uncovering an unknown fact, as these three are representing measures of physical activity. More interesting seems to be the relationship between needed time to walk ($marcha$) and geriatric depression score ($gdstotal$), which may reveal that needing more time to walk a certain distance and depression are correlated. The relative closeness of age ($hi8$), presence of polypharmacy ($polypharmacy$) and the number of comorbidities (COM_total) is more clear, according to the doctor's statements. Also interesting seems to be the apparent relative relationship between the income ($INCOME$) and the grip

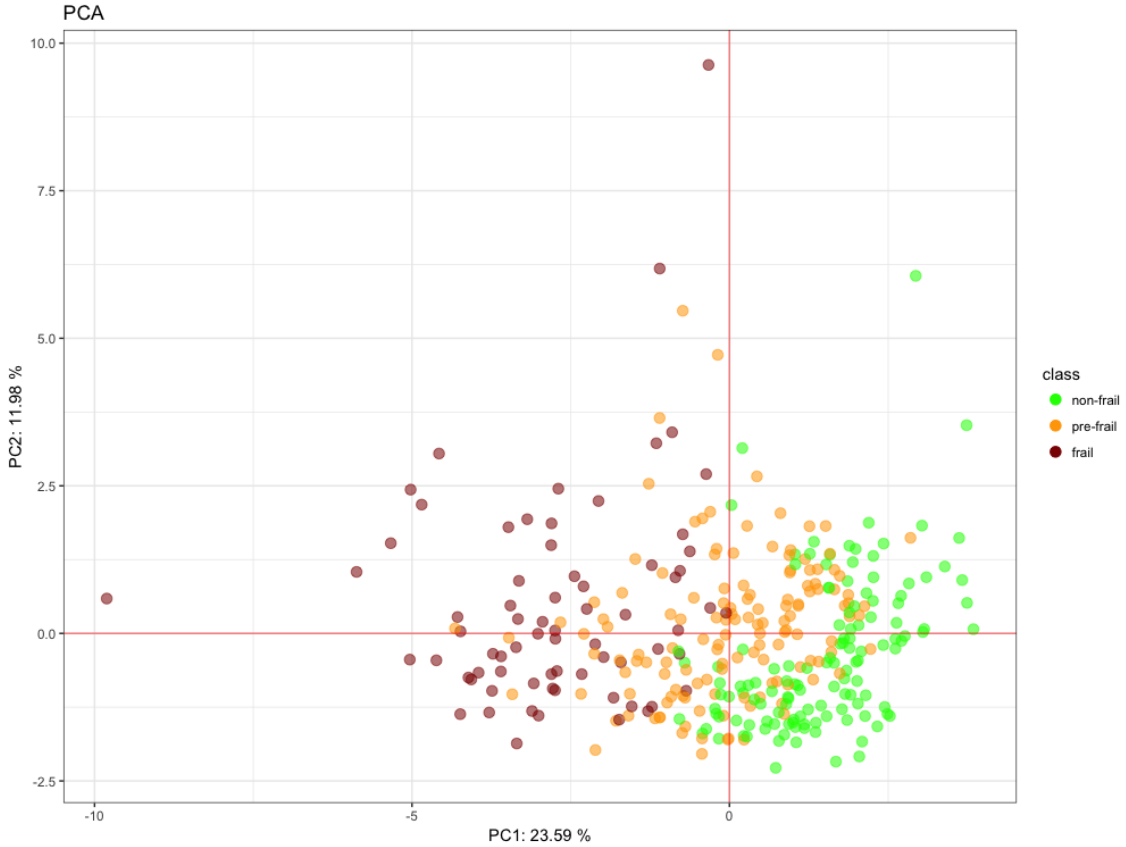


Figure 5.2: The 2-dimensional principal component plot shows the observations coloured corresponding to their frailty status. The PC plot was created using the frailty related features, which where described as risk factors or as preventive factors from the doctors.

strength ($fuerza1a$) of the patients.

The mobility score (MS_score), the feature representing the ability to sit down and up in a chair ($silla$) and the physical activity score ($pasetotal$) seem to appear in the direction of the healthier observations, which can be observed in figure 5.2 (non-frail: green coloured). A high geriatric depression scale score ($gdstotal$), having many comorbidities (COM_total), needing a long time to walk ($marcha$) and *polypharmacy* could be associated with frailty because they seem to be in the direction of the frail (red) observations (figure 5.2) and also in the direction of the feature representing frailty ($FRAGIL$) in the loading plot. These observations have to be made very carefully as the target variable $FRAGIL$ itself was also used in both PCA-plots. This should be kept in mind when observing the visualisations. For example, the circumstance that the in figure 5.2 shown observations are kind of clustered according to their frailty status. The separation is therefore not mainly

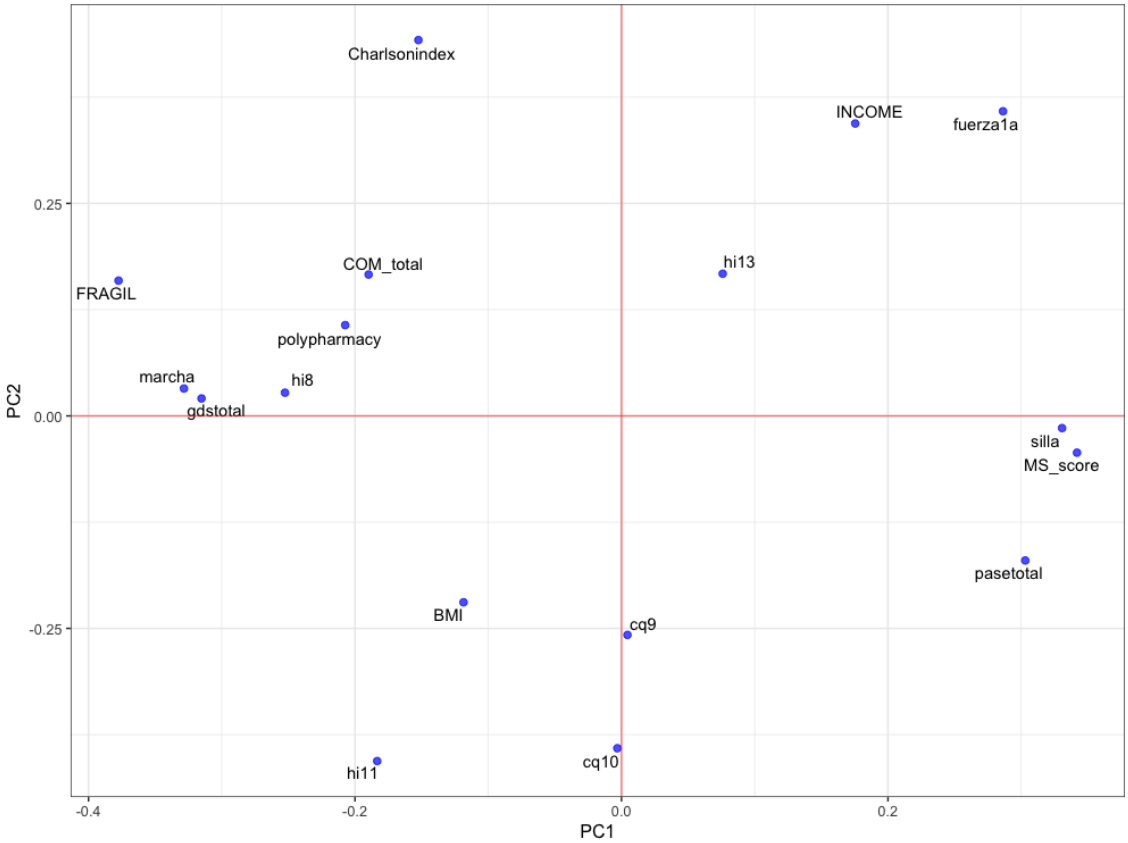


Figure 5.3: The 2-dimensional loadings plot of the frailty related features, which where described as risk factors or as preventive factors from the doctors.

based on the other variables but on the target variable which describes the frailty status ($FRAGIL$) itself.

Some of these relationships seem to be quite logical, for example that high educated people (education feature $hi13$) are more likely to have a higher income ($INCOME$), however others require further investigation and built assumptions have to be validated in the following steps. Wartner et al. (2016) states that it is very dangerous to use the PCA without further exploration as even promising looking visualizations might have no worth. Therefore, there is the need to further check the corresponding key-features. However, it can be seen that all these features do contribute to the variability of the observations as they are relatively far from the centre of the visualization, which apparently makes them quite usable.

iv) Cluster Analysis

For further exploration, the following step was to try to find groups of patients that behave similarly in order to get insights that could help prior to the predictive analysis. The objective is to build clusters using certain frailty related factors and to analyse the distribution of frailty and other features of interest in these sub-population-groups.

Variables for the clustering It was decided to cluster the patients according to the variables representing education status, financial situation, BMI, Geriatric Depression Scale (GDS) Score, comorbidities and mobility score.

Used Technique For the clustering the k-Means algorithm, which is described in section 2.5, was used.

Parameter Tuning In order to determine the optimal cluster number the Akaike information criterion (AIC), Bayesian information criterion (BIC) and the within-sum-of-squares (WSS) were used. In figure 5.4 the corresponding scores are shown. AIC and BIC serve as penalty score, that is why one looks for a cluster number with low values in those two. In practice also the "elbow" in the WSS-curve is searched, as described in section 2.5. Further, the package *NbClust* was used in order to have an additional opinion on which number of clusters should be chosen. This implementation provides 30 indices as basis for a decision. According to the majority rule of *NbClust*, 4 was proposed as the best number of clusters. Considering figure 5.4 and the result from *NbClust*, 4 was finally chosen as the number of clusters.

Results In figure 5.5 the results are shown. Presented is the composition of mean feature values for each cluster. In 5.6 the clusters are coloured according to the frailty status of their contained observations. Additionally, a normalized view is given in order to better examine the distribution. The same was done for gender in figure 5.7 and for the polypharmacy status in figure 5.8. Moreover, it can be seen in figure 5.5 that cluster number 3 seems to be quite interesting, as it differs a lot from the others in terms of composition of the mean feature values (the cluster centers). The number

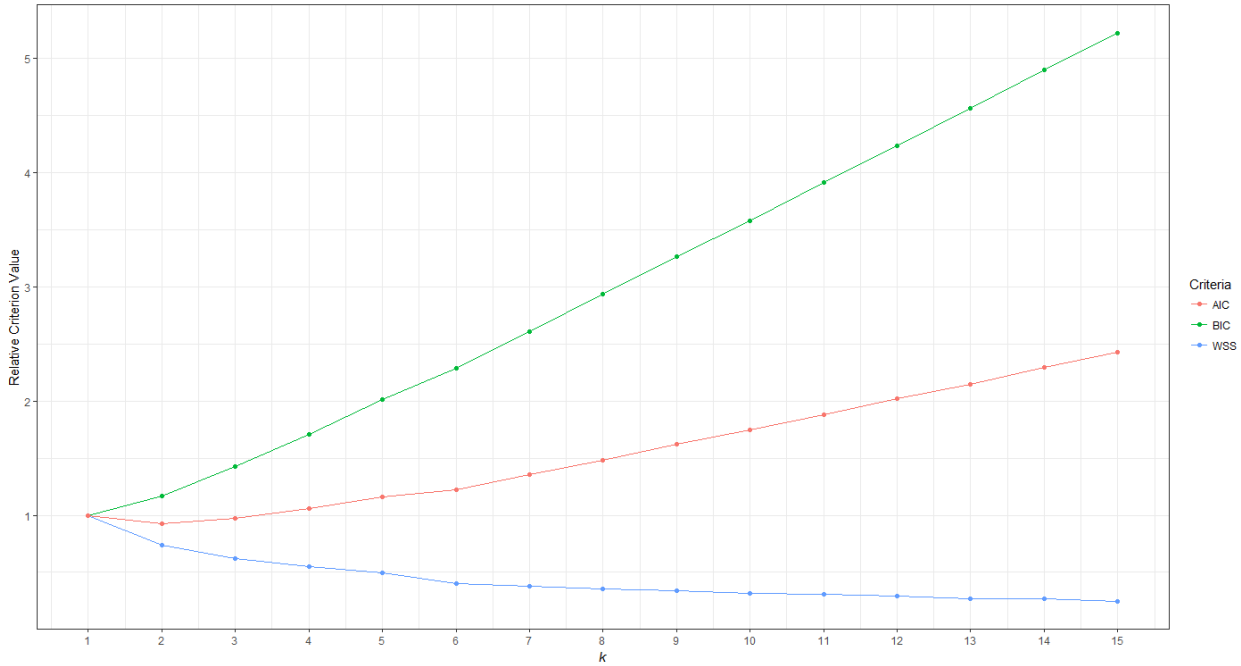


Figure 5.4: This plot shows the AIC, the BIC and the WSS scores. It serves to determine the optimal number of clusters.

of comorbidities (COM_total) seems to be quite high (highest of all clusters), the mobility score (MS_score) extremely low (lowest of all clusters) and the geriatric depression score (GDS) is elevated (2nd highest). One could assume, that this cluster captures a lot of the frail population, which was also stated by the doctors with whom this results were discussed. Interestingly, education ($EDUCATION$) and income ($INCOME$) is also low in these observations. In complete contrast stands cluster number 1. It contains more educated (elevated $EDUCATION$, highest of all clusters) patients with a low number of comorbidities (COM_total) and a high mobility score (MS_score) and also a higher income on average. The GDS in this cluster is also the lowest in comparison to the others. Therefore, for this cluster was assumed that the healthier part of the population is represented here. The body mass index (BMI) seems to be in all cluster more or less the same and does therefore not contribute a lot in separating the observations. Cluster number 2 contains mobile but depressed patients (highest GDS of all clusters) with the lowest education and also a low income. Cluster 4 seems to have very good parameters in terms of depression, mobility and comorbidities (lowest), and is therefore considered to represent the healthier observations.

In order to validate the assumptions the cluster observations were coloured ac-

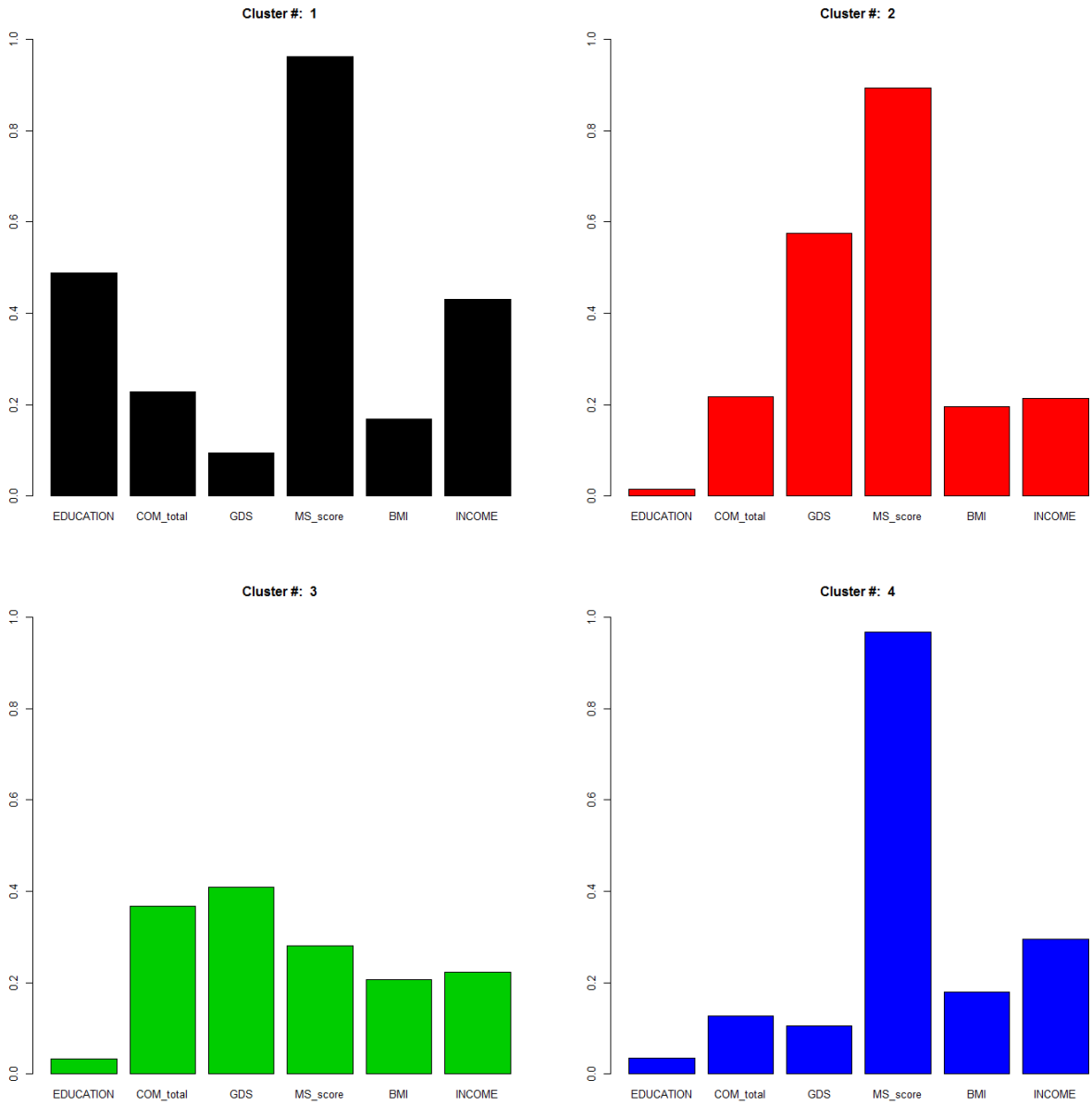


Figure 5.5: The individual composition of the 4 different clusters. The value represented by the bars is the mean value of the feature in the cluster.

cording to their frailty status as can bee seen in figure 5.6. Cluster 3 contains primarily frail patients, as has already been assumed. Also the assumption that cluster 1 and cluster 4 contain healthier subjects has been confirmed. Interestingly, cluster 2 contains mainly pre-frail and frail observations.

Now the distribution between the genders is examined. Therefore, the observations for each cluster were coloured according to their gender as can be seen in figure

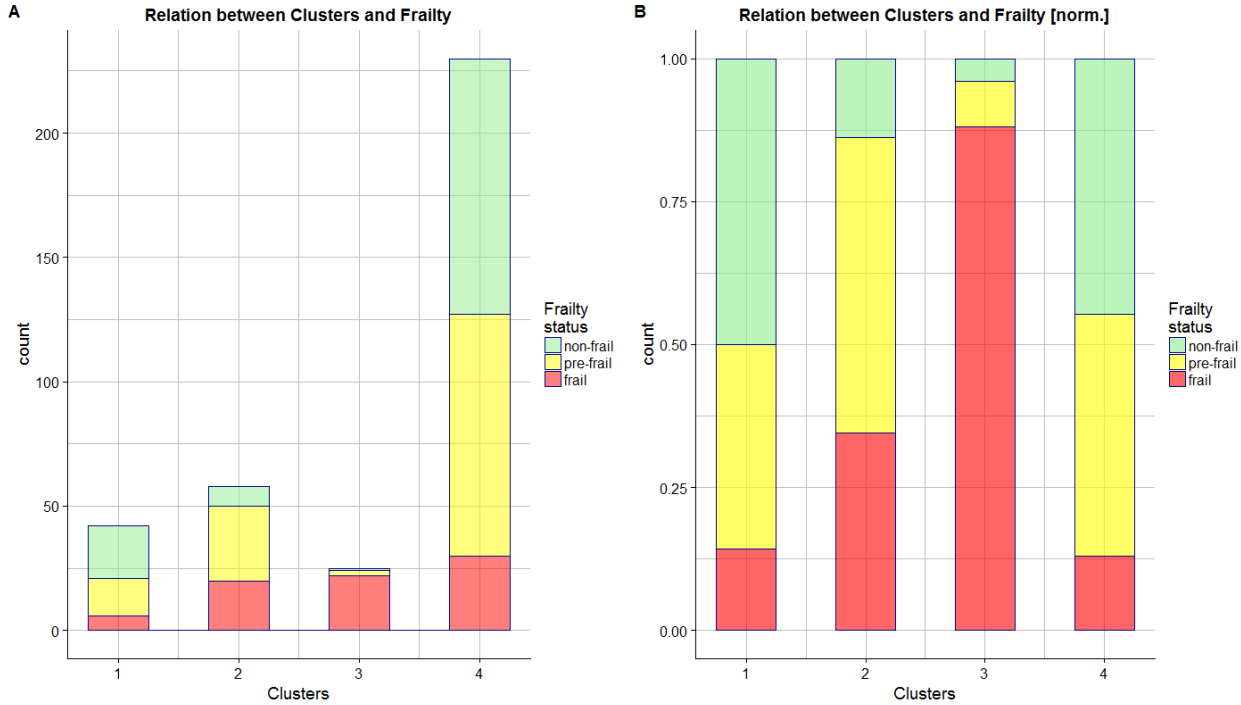


Figure 5.6: (A) The observations for each cluster are coloured according to their frailty status. (B) Here the observations are normalized for each cluster.

5.7. Cluster 1 and 4 seem to be quite equally distributed. Interesting is that the "frail" cluster 3 mainly contains women as well as the mixed pre-frail/frail cluster 2, which confirms the observations of the doctors. They stated that being female elevates the risk of being/becoming frail.

Now the clusters were coloured according to the amount of patients who take more than 4 medications (polypharmacy), as can bee seen in figure 5.8. Cluster 1 and 4, the apparently healthier clusters, contain less than 50% polypharmacy patients. The clusters which are more associated with frailty, number 2 and 3, contain more than 75% polypharmacy patients.

5.2.4 Final Data Quality Report

The data set contains 474 observations and 284 features including the target variable representing the frailty status. 176 features are more than 90% complete and in 41 features more than 50% of the values are missing. In order to make use of all the observations and therefore of the contained information, a special strategy to

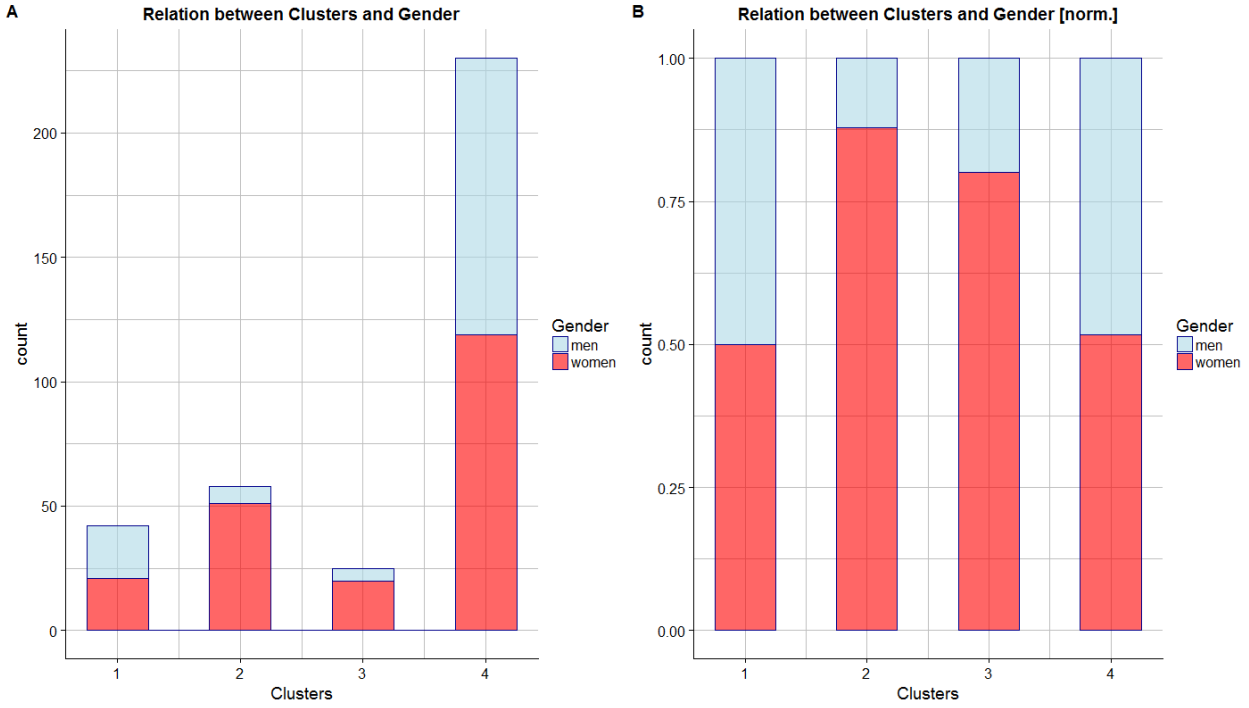


Figure 5.7: (A) The observations for each cluster are coloured according to their gender. (B) Here the observations are normalized for each cluster.

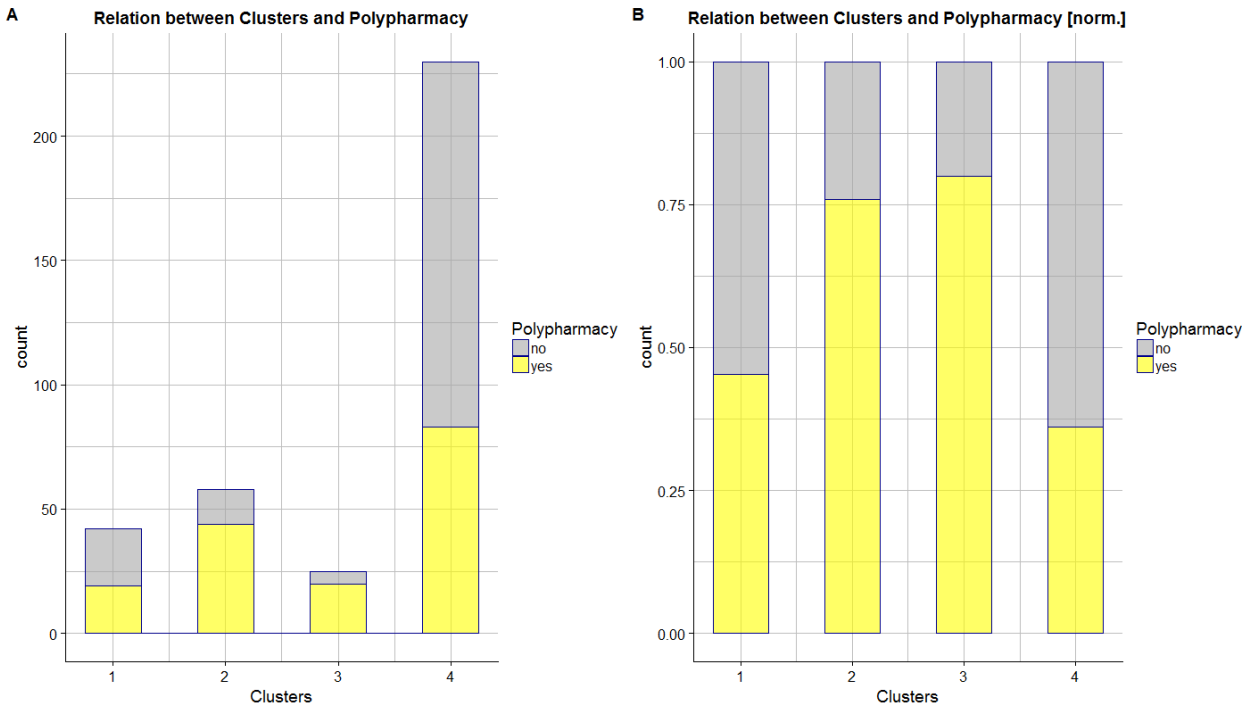


Figure 5.8: (A) The observations for each cluster are coloured according to their polypharmacy status. (B) Here the observations are normalized for each cluster.

deal with missing data is clearly necessary. Through analysis of known frailty related factors via PCA and clustering methods, it can be assumed that the from the doctors described relationships are also present in the data. The presence of different values, which are actually representing missing information requires further processing. For many features a special treatment is necessary in order to better capture their actual meaning as the current values do not sufficiently reflect it. In the next chapter the aforementioned issues will be treated.

5.3 Data Preparation

In this phase the data was cleaned, prepared and when necessary transformed. Further, new features were derived and the quality of the features in terms of predictiveness was assessed.

5.3.1 Cleaning and Transformation

Data File Preparation

First special values in the given data set were investigated. The data set was provided in form of an Microsoft excel-file. The missing values were fields containing the value "$\#NULL!$". This value was replaced by "NA", so that it is readable when loaded into the programming environment of R.
Variables, which used as decimal separator the comma, were treated and the comma was replaced by a point.

Removal of Unnecessary Features

For this thesis it was decided to exclude information regarding drugs. On the one hand because the information presented is not sufficiently structured and the preprocessing required exceeds the time for the thesis and on the other hand, because doctors preferred to have the first predictive model only with phenotypical parameters and results of the different tests. The drug related features only contain the ATC codes for drugs, the compound name and the commercial name with no information about intake frequency nor dosage and on top of that, the information is

weakly structured. Hence, drug related features, starting with the prefix "drug_", were excluded from the analysis. Only the feature *num_drug*, containing a numeric value representing the total number of drugs a patient is taking, is left for further processing.
Moreover, features which contain certain codes, assigned from the hospital or the blood laboratory, ending with the suffix "_code", were also removed as they do not contain relevant information.
Features which belong to the follow-up study conducted in the years 2011-2013, were discarded, as there were only 21 of them (and the remaining 264 are from the earlier wave) and therefore a temporal analysis was not possible. Also features, which in a statistical sense contain no information, were excluded. An example therefore is the feature *cq*8, which describes binary the presence of leukemia or polycytemia. As all the observations have the same value "2" (meaning "not present"), this feature was excluded.
Summing up, a total of 196 variables were left for further analysis.

5.3.2 Labelling of Unlabelled Observations

As has been previously noted, 3 out of the 474 records do not contain information regarding the frailty status of the patient. The majority of the related variables, which are used to determine the frailty status according to Fried et al. (2001), was available. Thus, the missing frailty status could be imputed using the *mice* package of R. For the imputation all available features were used. Further, the doctors were consulted and obtained multiple imputed frailty estimates were used as suggestions. Further, the values of the frailty related features (see table table A.1) were considered for the diagnosis using Fried's criteria (Fried et al., 2001).

After analysing the imputations and reviewing the health records of the patients with the physicians, the 3 missing frailty status could be determined.

5.3.3 Outlier Treatment

Statistical techniques and the from the hospital provided *codebook*, which contains a short description of each feature including the range and the meaning of appearing categories, were used to inspect the data set regarding potential outliers. Not

described appearing values were examined from a statistical point of view using the informal box plot method, described in section 2.3.2. Additionally, the kernel density estimate was analyzed. As a demonstrative example therefor serves the feature $p38gpt$, which represents the glutamic-pyruvic transaminase (GPT) level in U/L. It's Gaussian kernel density estimate is shown in figure 5.9. The majority

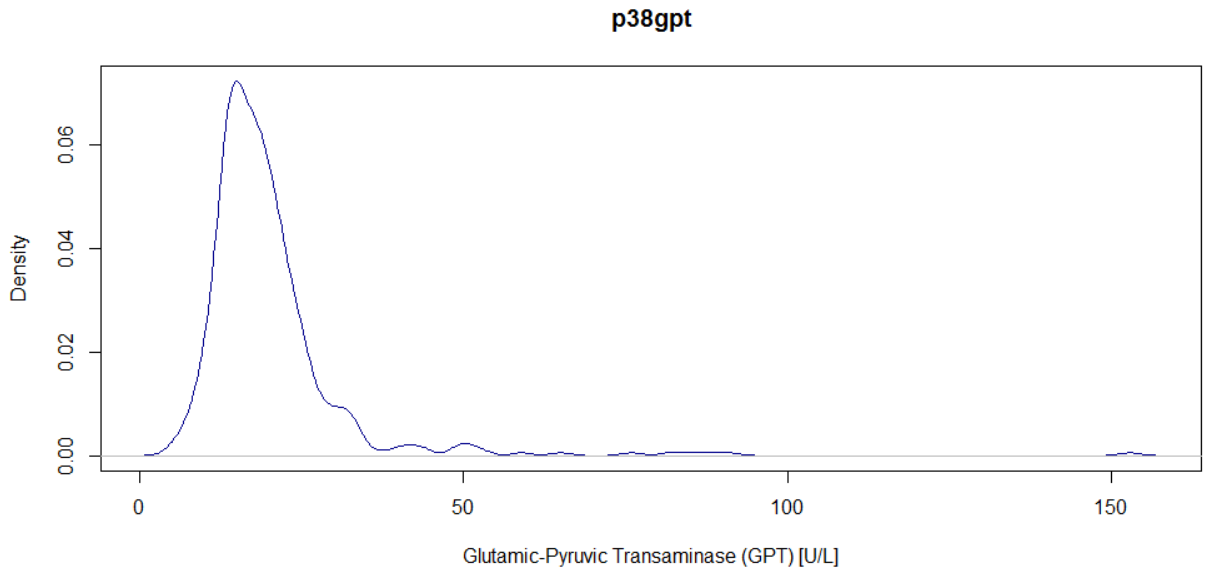


Figure 5.9: In this figure the Gaussian kernel density estimate for the feature $g38gpt$ is shown. It represents an estimate of the probability density function of the appearing glutamic-pyruvic transaminase (GPT) levels in U/L.

of the observations show levels between 0 and 50 U/L, as can be seen quite nicely in figure 5.9. The value(/values), which are appearing at approximately 150 U/L, requires(/require) further investigation as it(/they) could be an outlier(/outliers). For further investigation the variable is explored in a box-and-whisker plot, which is shown in figure 5.10. Here statistical outliers are presented as little circles.

Also in this plot a single outlier with the value of 153 U/L appears. After that exploration, domain-knowledge was used to analyse the significance of that certain value. Further, the doctors of the hospital were involved in the decision if the value is plausible and should be kept, or if it should be discarded. Moreover, possible values were discussed with the doctors and a threshold was established, exceeding values then simply were set to not available (NA). In the example of $p38gpt$ it was decided, after consulting literature and the medical doctors, to exclude values higher

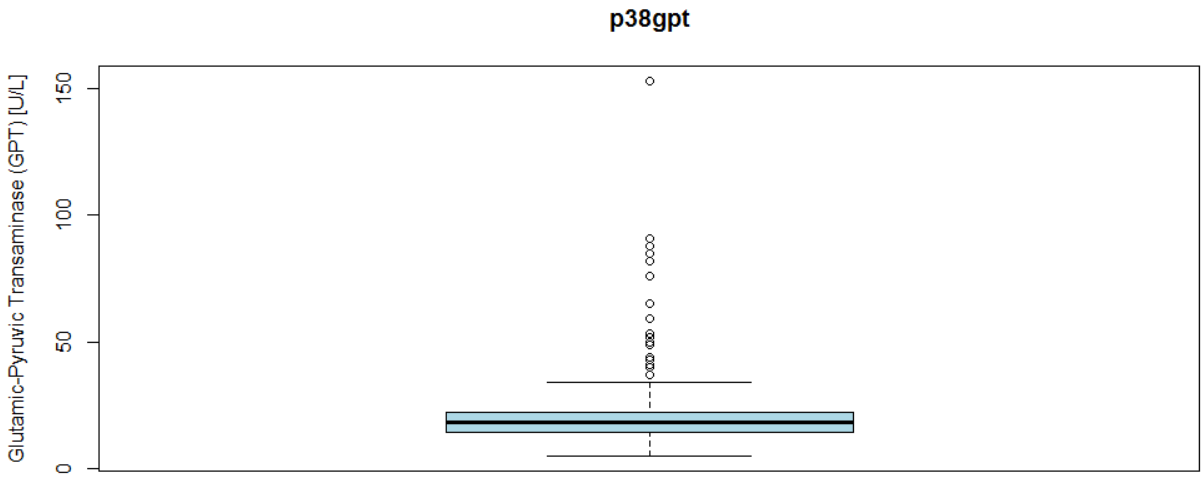


Figure 5.10: In this figure the box-and-whisker plot of the feature $g38gpt$ is shown. The y-axis represents the glutamic-pyruvic transaminase (GPT) level in U/L.

than 150 U/L by setting them to not available (NA). Like in the given example of $p38gpt$, this procedure was executed for each variable of the data set.

Many categorical features contain the values 77, 88 and 99 which are outside the expected range and have the core significance that no answer was given, these values were also set to not available (NA).

5.3.4 Feature Transformation

Now that the data is cleaned, the following step was to make use of certain features by transforming them. They could otherwise not have been used in the CRISP-DM phases which will follow.

There are 37 features where more than 60% of the values are missing (explained in section 5.2.3). Overall, follow-up questions contain a relative high percentage of missing data, as negative answered primary questions are not followed by the secondary ones. Some of them were transformed in order to make use of the contained information in all the follow-up questions' related features. This allows them to remain in the modeling pipeline. Affected are follow-up questions where the previous principal question was answered negatively, for these observations the value was set

to not available (NA) by the investigators.

Approach For some features it is possible to replace their not available values by a numeric value while still conserving the meaning. In the example of $tab1a$ (a feature referring to smoking frequency), for all observations which contain the value 2 in the feature $tab1$ (which means they have never smoked in their life), the $tab1a$-values were assigned to the value 5 instead of not available. The original categorical levels are 1: daily, 2: occasionally and 3: undecided. The value 5 was used to clearly separate it from a smoker.
This concept of assigning values for features representing follow up questions, which mainly contain missing values, makes them usable in the steps to come.

However, for other features it did not seem that easy to find a value which represents the meaning in the same way. Yet, for these features it was also considered to assign a new level, while trying to partly maintain the meaning. For example the feature $em1$ representing the answer to the question "Are you able to walk at home?" is followed by $em1a$ representing the answer to the question "If answered YES; Do you get tired when doing it?". If feature $em1$ contains the value standing for "no", $em1a$ doesn't contain a value for this observation. A possible solution for this problem could be assigning, additionally to the available levels "yes" (numeric: 1) and "no" (numeric: 2), the invented value "more than tired, I can't do it" (numeric: 0).

As the conclusion was made that this is not really perfect, respectively not good practice in respect of conserving the meaning of the values, these kind of features were removed, because most of the values were missing anyway and the introduction of a bias can't be ruled out.

5.3.5 Feature Creation

After the data had been cleaned and different features had been transformed, the following step was to extend the available data set by creating new features, using the available ones. The doctors presented different frailty associated variables. Some of them were not present in the data, but others could be calculated or extracted. Overall, following features could be derived:

- Total number of comorbidities (categorical, range: 0-5), used features $ccv1$, $ccv2$, $ccv4$, $ccv6$ and $ccv8$. Calculation:

$$COM_{total} = (-1)((ccv1-2)+(ccv2-2)+(ccv4-2)+(ccv6-2)+(ccv8-2)) \tag{5.1}$$

- Score of the mobility score related principal questions (categorical, range: 0-5), used features: $em1$, $em2$, $em3$, $em4$ and $em5$. Calculation:

$$MS_{score} = (-1)((em1-2)+(em2-2)+(em3-2)+(em4-2)+(em5-2)) \tag{5.2}$$

- The body mass index (BMI) (Quetelet, 1842) (continuous, expected range: 15-40), used features: height in cm ($altura1$) and weight in kg ($peso1$). Calculation:

$$BMI = \frac{peso1}{(\frac{altura1}{100})^2} \tag{5.3}$$

The general income of a person is also of interest, because it seems related to frailty. Patients with higher income seem to have more possibilities in terms of treatment and health support. In the data set is a categorical variable describing the income of the patient himself and further a variable which describes the income in the household the patient lives in. Combining these two could be a possible approach for creating a new, maybe suitable, feature. However, it does not seem as easy as with the already created features. Simply combining these features is one possibility, but could play out as a too crude way of doing it. An alternative would be to give those two different incomes weights and to build the sum afterwards. One could argue that the income of the patient himself is of higher importance, as the relationship to the third parties living with him is not quite clear and therefore, also not the given financial support. The calculation of the weighted sum of incomes for creating a feature called $INCOME$ using $Householdincome$ and $Individualincome$, depending on the chosen weights $w1$ and $w2$, can be seen in formula 5.4.

$$INCOME_{weightedsum} = w1 \cdot Householdincome + w2 \cdot Individualincome \tag{5.4}$$

5.3.6 Imputation of Missing Data

Now that the data is prepared and new features have been derived, the following step was to make sure all the observations can be used in the modelling phase. Therefore, it was decided to calculate different estimates for each missing value. Thus, missing values are imputed (filled) with estimates.

In table 5.3 the features where more than 5% of the values are missing can be seen. These measures are referring to the already in the previous steps pre-processed data set. An important step before applying imputation techniques, is to assess the reason for missingness. As already mentioned in section 2.3.3, three types of missing data exist and they are called Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). The assumed reason for the missingness and the according applicability of imputation techniques is also presented in 5.3. Features where more than one third of the values are missing were excluded from further investigations. They are marked in bold. Overall, all MNAR cases can be found in features which represent follow-up questions, they therefore were only be answered if the underlying basis question was answered positively. For them no imputation is possible because they can't be derived from other features.

The other ones were described as missing at random, which may seem in some cases debatable, for reasons already discussed in 5.2.3.

Implementation In order to use all the available information contained in the data set, different imputation settings using the MICE implementation, more specifically the CALIBERrfimpute expansion of it, were considered. They are described in section 4.2.3 and 4.2.3

Configuration Following configuration, regarding the imputation method, was chosen:

- For continuous features: **rfcont** for numeric random forest imputations
- For binary, ordered and unordered categorical features: **rfcat** for categorical random forest imputations (factor, >= 2 levels)

Feature Name	Percentage of Missing Data	Reason for Missingness	Imputation Possible
tab1a1a	**75.11**	MNAR (follow-up question)	no
alch1a1	**91.14**	MNAR (follow-up question)	no
alch1a2	**98.73**	MNAR (follow-up question)	no
alch1a3	**98.95**	MNAR (follow-up question)	no
alch1b	**82.91**	MNAR (follow-up question)	no
alch2	19.20	MAR	yes
alch2a	**86.29**	MNAR (follow-up question)	no
alch2b	**86.50**	MNAR (follow-up question)	no
alch2c	**86.92**	MNAR (follow-up question)	no
p15dd	17.72	MAR	yes
p44pcrh	14.98	MAR	yes
lawton2008	6.33	MAR	yes
mmse2008	15.82	MAR	yes
gdstotal	9.49	MAR	yes
Depression	9.49	related to {gdstotal}	no
INSULINA	11.60	MAR	yes
HDL	9.07	MAR	yes
LDL	9.07	MAR	yes
TESTOTOTAL	**37.97**	MAR	yes
TESTOLIBRE	**37.97**	MAR	yes
em2a	8.44	MNAR (follow-up question)	no
em2b	8.44	MNAR (follow-up question)	no
em3a	14.35	MNAR (follow-up question)	no
em3b	13.92	MNAR (follow-up question)	no
em4a	7.81	MNAR (follow-up question)	no
em4b	7.59	MNAR (follow-up question)	no
em5a	25.95	MNAR (follow-up question)	no
em5b	26.16	MNAR (follow-up question)	no
enpot1	17.93	MAR	yes
enpot2	18.78	MAR	yes
enpot3	18.14	MAR	yes
enpot4	22.57	MAR	yes
enpot6	12.87	MAR	yes
enpol1	13.08	MAR	yes
enpol2	13.29	MAR	yes
enpol3	13.29	MAR	yes
enpol4	13.29	MAR	yes
enpol5	13.29	MAR	yes
enmem1a	18.99	MAR	yes
enpmem2	19.41	MAR	yes
enpat1	**51.05**	MAR	yes
enpat2	**61.60**	MAR	yes
enleng1	13.92	MAR	yes
enleng2	13.08	MAR	yes
enleng3	13.50	MAR	yes
enleng4	13.29	MAR	yes
enpprx1	13.92	MAR	yes
enpprx2	13.50	MAR	yes
cognitive_impairment_MMSE_educative_level	17.09	MAR	yes
Individualincome	8.44	MAR	yes
Householdincome	13.29	MAR	yes
numpersonsfamilyunit	18.78	MAR	yes
IGF1	27.00	MAR	yes
cq6a	**98.73**	MNAR (follow-up question)	no
INCOME	13.71	MAR	yes

Table 5.3: Overview of features with more than 5% missing values. Additional information for the reason of missingness and the applicability of imputation methods is given. Features where more than one third of the values is missing are presented in bold.

Due to the size of the data set and the high number of features, the imputation could at first not be done at once, regarding the computational cost. Using all features as predictors for each feature when building the imputation model was tried with dif-

ferent settings, but primarily aborted because it would have taken 2 to 4 days, also because of the high number of iterations the monte carlo markov chain (MCMC) algorithm would have needed to produce converging estimates. Hence, at first the decision was made to make different splits of the data set. One option was to separate the features according to their semantics. Here the extent of dissection was also varied in order to find the best imputation, not only with regard to maintain the inter-feature relationships but also with regard to computational complexity. Another option is splitting the data set by choosing randomly subsets of a certain size and perform imputation inside these sets. It seemed to be computationally bare-able using thirds of the data and therefore working with three different feature sets. However, all these considerations regarding splitting the data were abandoned, on one hand because it would have definitely lead to the obscuration of inter-feature relationships between the subsets and on the other hand, it would not have been conform to the MICE instructions shown in Buuren and Groothuis-Oudshoorn (2011)'s paper. Further, it is a rule to use as much information as possible as this leads to multiple imputations which have a minimal bias and a maximal certainty (Buuren and Groothuis-Oudshoorn, 2011). So there had to be found another way to lower the immense computational cost. Fortunately, in the function *mice*() the used predictors for each imputation model for each feature can be customized. One way is selecting manually every predictor for every imputation model and another way is to use statistical measures for the selection. Consequently, is it for example possible to just consider variables which show a correlation higher than a certain percentage. Additionally, only such variables which are more than a certain desired percentage complete will be used. This still is computational cost-full but a supercomputer was available and therefore, the imputation could be executed using also low correlated features as predictors. For the first imputation only predictors, which correlate more than 7% and are more than 80% complete were selected by configuring the parameter *pred*. The overall configuration of the *mice()* function can be seen in following code-fragment.

```
mice(data, seed = 219,
    pred = quickpred(imp, mincor = 0.07, minpuc = 0.8),
    defaultMethod = c("rfcont","rfcat","rfcat","rfcat"),
    m = 5, maxit = 70, MaxNWts = 9000)
```

Here, *MaxNWts* depicts the maximal number of weights used by the inner neural network. The argument *maxit* was used to set the maximal numbers of iterations to 70. As creating 5 different imputations was desired, the parameter m was set to 5. The argument *defaultMethod* contains the different methods for the different data types, which were already mentioned earlier. Using *pred*, different restrictions regarding minimum correlation and completeness of the predictors were added. The first argument represents the data set in matrix form for which the imputations should be computed. The parameter *seed* can be used to set the number for initializing the pseudo-random generator.

The mean and the standard deviation for each variable at each iteration can be observed in the received imputation object. These values were plotted for the features with the highest amount of missing values in order to see if median and variance of the different imputations do converge. It seemed that 70 iterations are quite sufficient in this regard.

Results As can be seen in image 5.11, the kernel density estimates of the imputed values are approximating the "true" kernel density estimate of the original values. Especially coherent distributions can be observed for the features *HDL* and *LDL*, where all 5 imputations show a similar appearance. For the attribute *tads* only one imputation seems to have captured the kernel density estimate of the original values.

Null imputation is a task that on its own requires a lot of work due to the vast amount of decisions that have to be made. In fact for each attribute a deep analysis is required. In this work 157 attributes are given for which data imputation is required. Due to the fact that the main goal of the thesis is showing that prediction of frailty is feasible rather than analysing the most efficient algorithm for a prediction, quite enough effort has been dedicated to null imputation. However, a deeper analysis would be needed in order to answer questions related to the statistical analysis of the multiple imputations and also to the obtained statistical results, which are pooled into a final point estimate plus standard error, applying Rubin's pooling rules (Van Buuren, 2012).

The obtained imputations are then examined using visualisation tools. One possibility to check if the obtained imputations are reasonable, is to compare the kernel density estimates of the observed and the imputed values for ideally all variables.

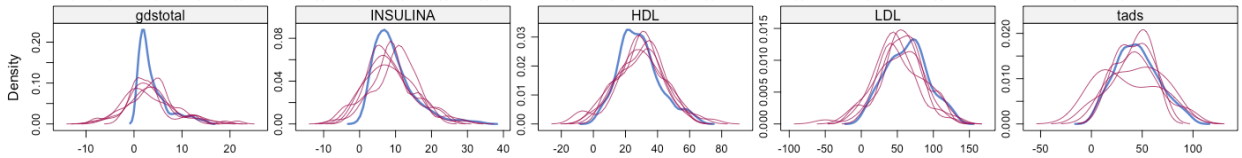


Figure 5.11: This plot shows the kernel density estimates for the original data (blue) and the 5 different imputations (red) for the features *gdstotal, INSULINA, HDL, LDL* and *tads*.

As this would not have been feasible within the scope of a master thesis, only features with more than 5% missing values were examined. Further, the kernel-density function was plotted and analysed for each feature and each imputation in order to evaluate the quality.

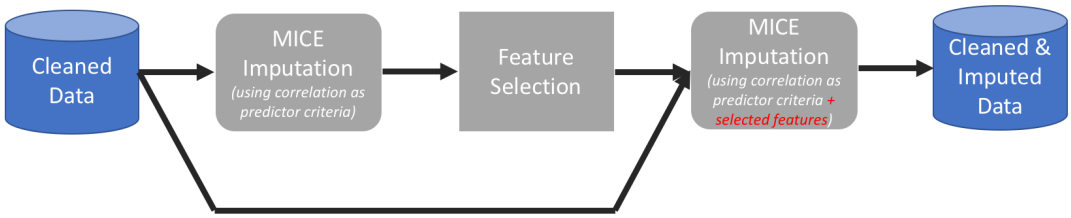


Figure 5.12: This figure illustrates how the imputation and the feature ranking process are connected. At first, the imputation models are built using features, which show a minimum correlation (here 7% was used) to the feature to be imputed. After that, the obtained 5 different data sets are used for the feature selection process. Knowing the selected features, the imputation is re-done. This by using as predictors additional to the correlated features also the selected ones.

The second imputation was done the same way, but this time also the selected features were included for every imputation model. This is recommended by Buuren and Groothuis-Oudshoorn (2011). The connection between the imputation and the feature selection process is demonstrated in figure 5.12.

The overall configuration of the *mice()* function for the second imputation can be seen in following code-fragment.

```
mice(data, seed = 219,
    pred = quickpred(imp, mincor = 0.07, minpuc = 0.8,
    include = selected_features),
    defaultMethod = c("rfcont","rfcat","rfcat","rfcat"),
    m = 5, maxit = 50, MaxNWts = 9000 )
```

The only difference is that by adding the parameter *include = selected_features* to the attribute *pred*, the selected features are used additionally for every imputation model.

Here, the obtained imputations were also analysed as it has been done before. With the help of density plots of the imputed and the original values, once again the quality of the imputations was assessed. The obtained 5 different imputed data sets then were used for the modelling process.

5.3.7 Dimensionality Reduction

As one objective is to predict the frailty syndrome with a subset of features, which are highly predictive, the most predictive features were determined using feature ranking methods. Further, the obtained results were compared with the suggested factors from the doctors of the Toledo study, which are listed below.

Factors associated with and increased prevalence or incidence of frailty:

- Older age
- Female
- Lower educational level
- Depression
- Sedentariness
- Some chronic diseases (Diabetes, Ischemic Heart Disease, COPD. Heart Failure, Cognitive Impairment/Dementia, osteoarthritis)
- Multiple comorbidities ($\geq$ 3 chronic diseases)

- Low income

There are also some protective variables:

- Physical exercise
- Vitamin D
- Protein calorie supplementation
- Mediterranean diet
- Reduction of multiple medications (polypharmacy 5 or more)
- Stopping to smoke
- Reduction of alcohol consumption

Risks of adverse outcomes are:

- Disability
- Falls
- Hospitalization
- Permanent institutionalization
- Death

Prognostic indicators in chronic diseases and surgery:

- Diabetes
- COPD
- Hypertension
- Chronic kidney disease
- Heart failure
- Oncology
- Major cardiac and abdominal surgery

Feature Selection

In order to make just use of the features which are indeed predictive and therefore beneficial for the final predictive model in terms of performance, different feature selection methods were considered. Finally, it was decided to use the *Boruta* algorithm. For further explanation and description see section (4.2.4).

Implementation The R package *Boruta* was used to perform feature selection on the data set using a random forest wrapper method. This selection was done with regard to the categorical target variable frailty.

Procedure At first, the features which are directly related to the target variable representing the frailty status ($FRAGIL$) were excluded from these process as it was the goal to use features, which could rather be used for a prediction than for a direct diagnosis. Among these features are those related to Fried's questions (Fried et al., 2001) for determining the frailty score (binarized weight loss *ppeso*, binarized exhaution $exhaustion$, binarized physical activity score $pasefrag$, binarized needed time to walk $marchafragil$, binarized grip strength $fuerzafragil$) and those, which were used to determine or calculate them. This includes: numeric grip strength in kg ($fuerza1a$), number of times the patient is able to stand up from the chair in a time of 30 seconds (used for determining $exhaustion$, called $silla$), needed time to walk: needed seconds for a distance of 3m ($marcha$) and numeric physical activity score for elderly ($pasetotal$). The feature used for determining weight loss was not contained in the data set.

For each imputed data set the feature selection process using the Boruta algorithm was executed. For the sake of obtaining reliable and stable results, the method was configured to use 1000 trees for the random forest algorithm and to perform 1000 runs in order to avoid so called tentative results. At the end, 5 different sets of selected features were present. The finally chosen selected features were those, which appeared at least 3 times in the 5 different Boruta sets. The complete feature selection process, which begins after the first executed imputation procedure and provides the selected features for the second imputation, is shown in 5.13.

Figure 5.13: This figure shows the overall feature selection process. At first, the Boruta algorithm is applied on each imputed data set. Then, the 5 different selected feature sets are compared and features which appear in 3 or more selected sets are chosen for the final feature set.

Results In figure 5.14 the result of the feature selection is presented. The variables are ordered by importance, the rejected ones are coloured red, the selected ones green and those, for which no decision could be made, are yellow. All the importance measures of the features were compared to randomly permuted copies of themselves, so called shadow attributes. The Z Score of the most important shadow attribute was used as separator between selected and rejected features. Features where no decision could be made were marked tentative and coloured yellow.

By using the function *TentativeRoughFix* those features, with a median importance higher than the maximal one of the shadow attributes, were selected and the others rejected. This is a simple test for judging these tentative attributes. Tentative attributes could also be resolved by increasing the number of importance runs of the Boruta algorithm. That is why instead of the default 100 runs, 1000 runs were used. The finally selected features can be seen in table 5.4.

After the feature selection, the obtained final variables were used for another imputation round. As suggested by Buuren and Groothuis-Oudshoorn (2011), the features which are powerful in terms of predictiveness should always be used in the imputation for each feature. That is why they all were included in each imputation

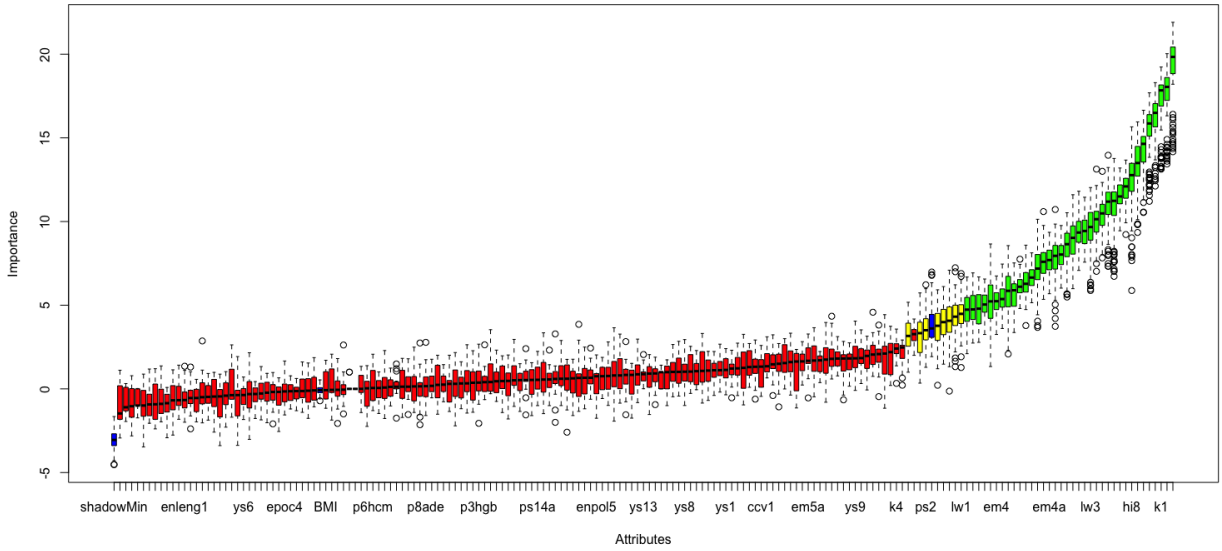


Figure 5.14: This image shows the attributes and their importance measure, by which they were selected (green) or rejected (red). This decision was made by comparing their importance measure to randomly permuted copies of themselves, the so called shadow attributes (Kursa et al., 2010). Features which could neither be selected nor rejected were marked tentative (yellow).

model.

Interpretation The selected feature set shown in 5.4 seems to be the most useful subset of features in the given data, regarding binary frailty classification. Interestingly $p40falc$ is also part of these well-suited predictors. This variable represents the blood alkaline phosphatase level in U/L. Less surprising is that age ($hi8$) is also among these features. Also the height of a person ($altura1$) seems to be predictive. The known frailty related variables representing depression ($Depression$, $gdstotal$, $ys2$, $ys4$) and polypharmacy ($polypharmacy$, num_drug) do also appear in the selected set. The variable MS_score, which combines 5 questions about mobility and was derived in this work, can also be found. Further, Mobility Scale related variables are also present ($em2a$, $em3$, $em4a$, $em5$). Question-features from the Mini-Mental-State-Examination ($mmse2008$, $cognitive\_impairment$ -$\_MMSE\_educative\_level$, $enpmem2$), the Instrumental Activities of Daily Living questionnaire ($lawton2008$, $lw1$, $lw2$, $lw3$, $lw4$, $lw5$, $lw6$) and the Activities of Daily Living questionnaire ($katz2008$, $k1$) seem also be very predictive. Also features reflecting the self-reported

Selected Features	Description	Type
altura1	Height (cm)	numeric
cognitive_impairment_MMSE_educative_level	Presence of cognitive impairment	binary
Depression	Presence of depression	binary
em2a	Mobility Scale follow-up question (tiredness when going out)	binary
em3	Mobility Scale question (stair-climbing ability)	binary
em4a	Mobility Scale follow-up question (tiredness when walking outside)	binary
em5	Mobility Scale question (walking outside ability)	binary
enpmem2	MMSE follow-up question (remembering objects ability)	categorical
gdstotal	Total GDS	binary
hi8	Age in years	numeric
k1	ADL question (difficulty washing)	categorical
ktaz2008	Number of ADL abilities	numeric
lawton2008	Number of IADL abilities	numeric
lw1	IADL question (difficulty using telephone)	categorical
lw2	IADL question (difficulty shopping)	categorical
lw3	IADL question (difficulty cooking)	categorical
lw4	IADL question (difficulty doing light housework)	categorical
lw5	IADL question (difficulty doing heavy housework)	categorical
lw6	IADL question (difficulty using public transportation)	categorical
mmse2008	Total MMSE score	numeric
MS_score	Sum of mobility score main features (em1,em2,em3,em4,em5)	numeric
num_drug	Number of drugs (drug intake)	numeric
p40falc	Alkaline phosphatase [U/L]	numeric
polypharmacy	Presence of polypharmacy	binary
ps1	Self-reported health status	categorical
ps3	Self-reported health status compared to people the same age	categorical
ps6	Capacity of dealing with problems	categorical
ps7	Capacity of dealing with tasks	categorical
ys2	GDS question (dropped activity of interests)	binary
ys4	GDS question (boredom)	binary
reum1	Presence of joint inflammation (>4 weeks in a row)	categorical

Table 5.4: Obtained final selection of features using the Boruta algorithm and a voting system. When a feature was selected by the Boruta algorithm in at least 3 different imputed data sets (out of 5), it was included in the final selection. In total 33 features were selected for the binary classification problem (non-frail/frail).

health-status were selected ($ps1$, $ps3$, $ps6$, $ps7$). Further, a feature reflecting a question regarding rheumatic disease ($reum1$) appears in the final selection.

5.4 Modelling and Evaluation

Once data had been prepared, the following step was to build predictive models. As can be seen in the sections to come, different techniques have been applied. Later the received results have been compared and validated. In what follows, one can find the model settings (section 5.4.1), the data set preparation (section 5.4.2), the modeling and validation schema (section 5.4.3), the model performance (section 5.4.4) and lastly the evaluation of the models (section 5.4.5).

5.4.1 Classification Model Settings

As learning algorithms for the predictive models the Naïve Bayes (NB) algorithm, classification and regression trees (CART), bagging CART, C5.0, random forest (RF), support vector machines (SVM) and linear discriminant analysis (LDA) were used. The different algorithms were implemented in the R environment using different third party packages, which are listed below. Further, changed configurations, which differ from the default settings are described in this listing.

Naïve Bayes

The Naïve Bayes classifier *naiveBayes* of the R package *e*1071 was used in its standard configuration.

CART

The classification and regression tree algorithm *tree* of the same titled R package was used in it's standard configuration.

Bagging CART

The bagging CART implementation *bagging* from the R package *ipred* lead to the best results, when using 55 bootstrap replications.

C5.0

The best accuracy for the C5.0 algorithm (from the R package *C50*) could be achieved using 50 iterations for the multiclass classification and 55 iterations for the binary classification.

Random Forest

The best accuracy in the random forest implementation "randomForest" from the R package with the same name was achieved, using 1000 trees, no replacements in the inner sampling of cases and 5 as number of variables randomly sampled as candidates at each split.

Support Vector Machines

The best setting for this algorithm was using as type the C-classification, as kernel the radial basis function and as tolerance of termination criterion the value 10^{-3}. The degree was set to 3, the 'C'-constant of the regularization term in the Lagrange formulation was set to 10 and the gamma of the radial basis function was set to 0.07.

Linear Discriminant Analysis

This method from the R package *MASS* was used in its standard configuration.

5.4.2 Data Set Preparation

In order to utilize the data in the best way, it has been shown that sometimes it is beneficial for the performance of the learning algorithms to transform the data to different ranges and also to change the distribution. This was also considered in this work and therefore, every algorithm was used on the z-score standardized, the Min-Max normalized and the raw data set. Where the raw form represents the data after completion of the preprocessing phases.

Standardized z-scores The standardized form represents the data after building the standardized z-scores, using the formula 5.5. Here x represents the raw value, μ the mean of all the values of the feature and σ the standard deviation of all the values of the feature. This formula is applied to each value x_i and as a result the standardized feature has a mean of 0 and a standard deviation of 1.

$$z(x_i) = \frac{x_i - \mu}{\sigma} \tag{5.5}$$

Min-Max Normalization Min-Max normalization is a method where the values of the data are transferred into a range of $[0, 1]$. Where the lowest appearing value x_{min} is set to zero and the maximal value x_{max} is set to 1. The used formula is shown in equation 5.6. Here each value x_i is Min-Max normalized using its current

value, x_{min} and x_{max}.

$$mm(x_i) = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{5.6}$$

For each learning algorithm the 3 aforementioned differently prepared data set variants were used and the resulting performances were compared. Then for each algorithm the variant which leads to the best performance was chosen. The results can be seen in table 5.5.

Learning algorithm	**raw form**	**z-score standardization**	**Min-Max normalization**
Naïve Bayes			X
CART	X		
Bagging CART	X		
C5			X
Random forest		X	
Support vector machines (RBF Kernel)		X	
Linear discriminant analysis		X	

Table 5.5: Selected data preparation for each algorithm: Here the data preparation form, which leads to the best performance, is marked with X.

5.4.3 Modeling and Validation Schema

After preparing the data for the modeling phase, the next step was building the models and validating them. In image 5.15 the procedure for modelling and evaluating is presented. At the beginning each obtained imputed data set is used to build the different models (e.g., RF, DT, SVM), which are tested in a cross-fold validation setup. The resulting performance measure values of each model for each imputation are then compared and the one with the overall best performance is chosen as final model. Therefore, 5 different final models are obtained at the end. Afterwards they can be used as a ensemble classifier, which provides one result for new unseen instances.

In order to evaluate the out of sample error of the built models, as mentioned before, 10-fold cross-validation was performed. Due to the fact that the classes are imbalanced, a stratification technique was implemented. The scheme can be seen in figure 5.16. At first, the observations were split according to their frailty status (2 classes). Afterwards, the 10 folds were created separately for each class and then fused according to the fold-number. The observations were chosen randomly.

By using multiple 10-fold cross-validations, a first estimate of the generalization

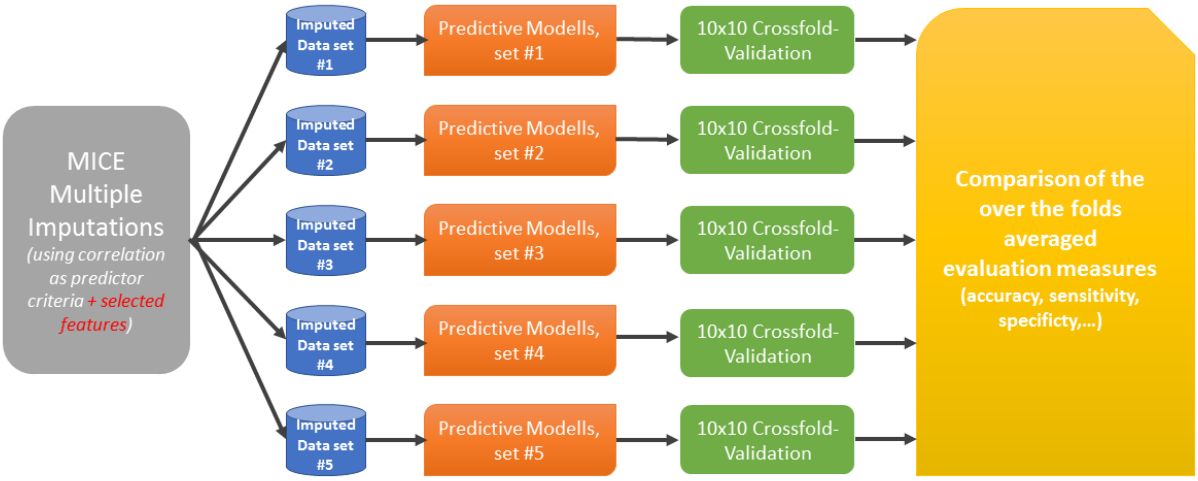


Figure 5.15: This image shows the general procedure. Firstly, models are built using the 5 different obtained imputed data sets. Secondly, the models are evaluated in a cross-fold validation setup. The resulting performance measure values (e.g. accuracy, sensitivity, specificity) are then combined in one final result, by averaging the 5 different results.

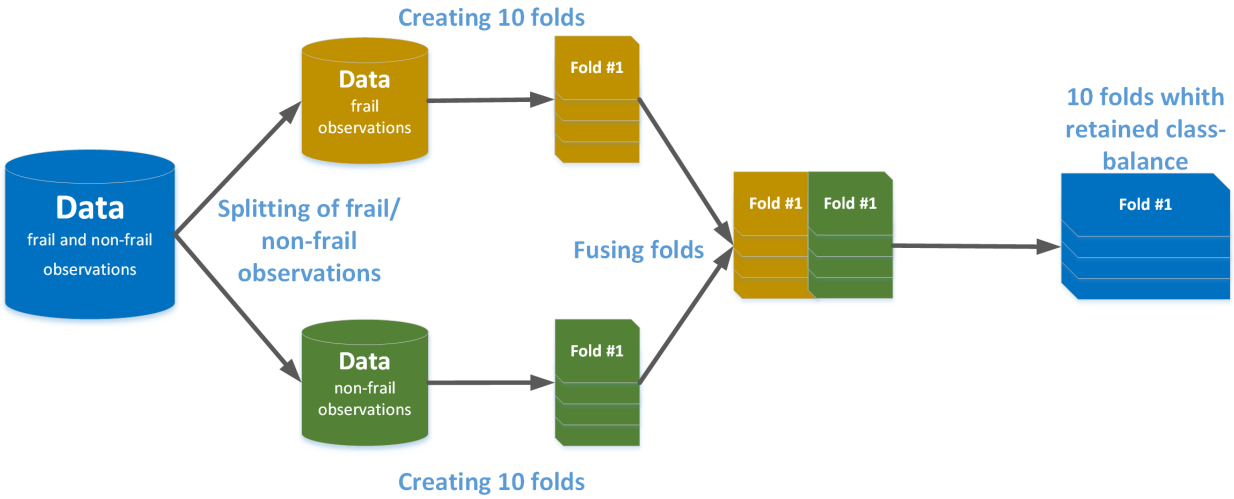


Figure 5.16: This image shows the stratification procedure. First the data was split according to the frailty classes frail and non-frail. Then for each class 10 folds were created. Afterwards the folds were fused together according to the fold number and as a result, 10 folds were available and the original class distribution was retained.

error is obtained. Though, according to Bellazzi et al. (2011) this is hardly sufficient,

so following their recommendation the prediction performance should also be tested on a independent data test set. Unfortunately, the provided additional data set by the Getafe Hospital does not contain the same here selected features.

As for the modelling phase the 5 different imputed data sets were considered, 5 different best performing classifiers were obtained. Thus, the final predictive model is, as mentioned before, an ensemble classifier, which can be used on new unseen instances. The final predicted class is the result of the 5 different votes, where each vote is the corresponding classification result of each model.

5.4.4 Model Performance

The model performances were obtained by averaging each performance measure for the 10 different 10-fold cross-validation setups. The obtained results can be seen in table 5.6. For each performance measure, the over the folds averaged value including the standard deviation is shown. The highest obtained value for each performance category is marked in bold.

5.4.5 Evaluation

For this research two different evaluations are required. First, the analyses of the performances of the models (see section 5.4.5) and later, the analysis of how the models actually fit the business goals (see section 5.4.5).

Analysis of the Model Performances

The overall best performances in nearly all measures have Random Forest and SVMs with a radial basis function as kernel. Followed by bagging CART, LDA, C5 and CART. Striking is the high obtained specificity and precision of the Naïve Bayes classifier, while it performs very poorly in the other measures. In this case specificity represents the ratio of predicted real non-frail patients to all non-frail patients. Thus, this classifier shows an extraordinary performance in the task of detecting non-frail patients. The highest values for accuracy and AUC are always achieved by Random Forest and SVMs, which do not differ significantly in their results. The highest scores in table 5.6 in each category for each imputation are marked in bold. The

Imputation 1						
Prediction method	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**F_1-Score**
Naive Bayes	73.20 ± 5.97%	0.756 ± 0.052	0.656 ± 0.102	**0.856 ± 0.079**	**0.885 ± 0.054**	0.749 ± 0.067
CART	72.77 ± 5.20%	0.710 ± 0.061	0.782 ± 0.108	0.639 ± 0.168	0.789 ± 0.065	0.778 ± 0.049
Bagging CART	75.51 ± 7.16%	0.731 ± 0.070	0.830 ± 0.086	0.633 ± 0.084	0.786 ± 0.048	0.806 ± 0.060
C5	**77.83 ± 7.13%**	0.752 ± 0.086	**0.860 ± 0.056**	0.644 ± 0.164	0.804 ± 0.075	**0.829 ± 0.051**
Random forest	77.64 ± 5.62%	0.755 ± 0.053	0.844 ± 0.089	0.667 ± 0.087	0.806 ± 0.041	0.823 ± 0.050
Support vector machines (RBF Kernel)	77.64 ± 6.55%	**0.762 ± 0.065**	0.824 ± 0.09	0.700 ± 0.099	0.819 ± 0.053	0.819 ± 0.057
Linear discriminant analysis	75.11 ± 5.34%	0.739 ± 0.042	0.789 ± 0.096	0.689 ± 0.047	0.805 ± 0.023	0.795 ± 0.055

Imputation 2						
Prediction method	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**F_1-Score**
Naive Bayes	72.78 ± 6.47%	0.750 ± 0.059	0.656 ± 0.109	**0.844 ± 0.094**	**0.878 ± 0.063**	0.745 ± 0.072
CART	70.89 ± 5.94%	0.699 ± 0.057	0.741 ± 0.098	0.656 ± 0.104	0.781 ± 0.047	0.757 ± 0.058
Bagging CART	75.11 ± 6.59%	0.729 ± 0.072	0.820 ± 0.089	0.639 ± 0.134	0.792 ± 0.066	0.802 ± 0.054
C5	77.39 ± 7.35%	0.745 ± 0.093	**0.867 ± 0.057**	0.622 ± 0.192	0.797 ± 0.082	**0.828 ± 0.050**
Random forest	77.01 ± 6.65%	0.752 ± 0.064	0.827 ± 0.101	0.678 ± 0.101	0.809 ± 0.052	0.815 ± 0.060
Support vector machines (RBF Kernel)	**77.63 ± 7.01%**	**0.761 ± 0.071**	0.827 ± 0.085	0.694 ± 0.102	0.816 ± 0.057	0.820 ± 0.060
Linear discriminant analysis	76.14 ± 5.15%	0.752 ± 0.046	0.792 ± 0.081	0.711 ± 0.057	0.817 ± 0.032	0.803 ± 0.050

Imputation 3						
Prediction method	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**F_1-Score**
Naive Bayes	73.41 ± 5.64%	0.757 ± 0.057	0.664 ± 0.083	**0.849 ± 0.102**	**0.885 ± 0.069**	0.755 ± 0.056
CART	73.21 ± 5.75%	0.728 ± 0.07	0.746 ± 0.064	0.709 ± 0.14	0.815 ± 0.067	0.776 ± 0.045
Bagging CART	78.28 ± 3.92%	0.764 ± 0.057	**0.841 ± 0.058**	0.688 ± 0.148	0.823 ± 0.062	0.828 ± 0.026
C5	74.06 ± 7.12%	0.709 ± 0.089	0.837 ± 0.057	0.581 ± 0.181	0.774 ± 0.073	0.802 ± 0.048
Random forest	77.62 ± 6.65%	0.762 ± 0.076	0.820 ± 0.068	0.704 ± 0.134	0.824 ± 0.068	0.820 ± 0.052
Support vector machines (RBF Kernel)	**79.32 ± 5.00%**	**0.779 ± 0.056**	0.838 ± 0.049	0.720 ± 0.09	0.833 ± 0.048	**0.834 ± 0.040**
Linear discriminant analysis	78.47 ± 4.77%	0.773 ± 0.051	0.821 ± 0.059	0.726 ± 0.085	0.833 ± 0.045	0.825 ± 0.040

Imputation 4						
Prediction method	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**F_1-Score**
Naive Bayes	72.78 ± 5.89%	0.750 ± 0.061	0.657 ± 0.083	**0.843 ± 0.111**	**0.881 ± 0.075**	0.749 ± 0.057
CART	71.26 ± 5.83%	0.697 ± 0.053	0.762 ± 0.095	0.631 ± 0.083	0.774 ± 0.043	0.765 ± 0.058
Bagging CART	76.38 ± 5.77%	0.747 ± 0.069	0.817 ± 0.076	0.676 ± 0.147	0.812 ± 0.065	0.811 ± 0.046
C5	74.25 ± 7.13%	0.712 ± 0.085	**0.837 ± 0.057**	0.587 ± 0.157	0.774 ± 0.07	0.803 ± 0.052
Random forest	76.99 ± 5.90%	0.755 ± 0.069	0.817 ± 0.069	0.693 ± 0.136	0.819 ± 0.067	0.815 ± 0.046
Support vector machines (RBF Kernel)	**78.47 ± 5.14%**	0.771 ± 0.057	0.827 ± 0.053	0.714 ± 0.092	0.829 ± 0.049	**0.827 ± 0.041**
Linear discriminant analysis	78.06 ± 5.39%	**0.772 ± 0.057**	0.807 ± 0.061	0.737 ± 0.091	0.837 ± 0.049	0.820 ± 0.045

Imputation 5						
Prediction method	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**F_1-Score**
Naive Bayes	73.41 ± 5.45%	0.756 ± 0.053	0.664 ± 0.088	**0.849 ± 0.098**	**0.885 ± 0.066**	0.754 ± 0.057
CART	71.67 ± 7.79%	0.702 ± 0.087	0.762 ± 0.100	0.642 ± 0.166	0.786 ± 0.089	0.769 ± 0.066
Bagging CART	76.79 ± 4.69%	0.749 ± 0.053	0.827 ± 0.071	0.671 ± 0.115	0.809 ± 0.049	0.815 ± 0.039
C5	75.31 ± 4.08%	0.726 ± 0.055	**0.837 ± 0.065**	0.615 ± 0.138	0.787 ± 0.055	0.808 ± 0.030
Random forest	78.03 ± 5.10%	0.764 ± 0.060	0.830 ± 0.073	0.698 ± 0.129	0.824 ± 0.061	0.824 ± 0.041
Support vector machines (RBF Kernel)	**78.47 ± 5.39%**	**0.771 ± 0.059**	0.827 ± 0.055	0.714 ± 0.092	0.828 ± 0.049	**0.827 ± 0.043**
Linear discriminant analysis	77.62 ± 5.35%	0.769 ± 0.058	0.800 ± 0.063	0.737 ± 0.102	0.836 ± 0.054	0.816 ± 0.045

Table 5.6: 10-fold cross-validation results for the binary classification models for each imputed data set, working with the two classes $non-frail$ and $frail$. The highest obtained value for each performance category is marked in bold.

variation of the results between the different imputed data sets is also very small, which indicates that also the variation of the imputed values is small. For example, the accuracy of SVM averaged over all imputed data sets is 78.31 ± 0.70%. The standard deviation is not even one percent. The RF algorithm performed slightly inferior with an averaged accuracy of 77.46 ± 0.45%. Here the standard deviation is below a half percent.

Analysis of the Business Goal Compliance

The data mining goals, which were derived from the business goals, described in section 5.1.1 are now checked for compliance with the results. Regarding the finding of suitable predictors/"biomarkers" it can be said that such have been found. They seem to be consistent with known frailty risk factors or preventive factors found by the medical community. Interesting seems to be the finding that the feature $p40falc$, representing blood alkaline phosphatase level in U/L, is highly predictive. This certainly requires some follow up investigations, as this could possibly be a new biomarker for frailty detection. The doctors said that this variable is probably a good predictor, because it gives information about inflammation processes in the body. They are already investigating it, in the scope of the FRAILOMIC initiative (Lippi et al., 2015), which is a research project aiming to identify the factors that turn frailty into disability. The doctors conformed that the found biomarkers are related to frailty. They commented also on the missingness of the gender feature. According to them, it's one of the important markers for determining frailty and they were surprised that it did not appear in the final predictor set. It is possible that the feature selection algorithm found this variable to be redundant and that the contained information is already provided by other features. The variable height is, for example, highly correlated to the gender variable (correlation coefficient = 0.725).

The built models achieved an accuracy of more than 78% for binary classification of the frailty syndrome, without using features, which are directly related to the target or used to build it (see Fried's frailty criteria and stages (Fried et al., 2001)). The results show, that it is feasible to build predictive models for the frailty syndrome using data from electronic health records.

6. Discussion and Lessons Learned

In what follows, the methodology, obtained results and insights are discussed.

Overall, CRISP-DM has been successfully applied in a real medical environment. It has been shown that the integration of doctors in the CRISP-DM loop (data understanding, data preparation, modeling and validation phase) seems to be highly beneficial for the obtained models, in terms of validity, robustness and accuracy.

In particular, concerning the business understanding, the problem of frailty as described by the physicians, has been translated into sophisticated data mining tasks.

However, prior to being able to apply data mining techniques to the raw data (that is to say, the data provided by the physicians) it had to be understood, cleaned and prepared in order to be the input for the different data mining algorithms. Consequently, in the data understanding phase statistics, clustering methods and different visualisation techniques have been applied in order to understand the data, to find null values, to detect outliers and to find underlying correlations and relationships.

Further, a deeper understanding was acquired through the help of doctors and by consulting literature. The analysis of all the features helped to determine their particular importance in the frailty prediction. The application of the ontology-based PCA approach described by Wartner et al. (2016) was able to deliver some insights which were further investigated.

Following the data understanding, the data had to be prepared. On the one hand to clean inconsistencies detected in the previous stage and on the other to include semantics given by doctors. Further, multiple estimates for missing values were computed using imputation methods. As the final step of this stage, tables

were produced which serve as an input for the algorithms.

Moreover, it has been shown that this phase is of highest importance and has proven to be the most time-consuming part. Performed manipulations in this phase had a high impact on the results regarding quality and accuracy. Especially the imputation of null values was a complex and difficult task, given that deriving valid and probable estimates while trying to establish a valid model became apparent as very hard to achieve.

Using a random forest wrapper based feature selection method, potential predictors were identified. Further, previously known predictors for frailty, from the medical community, could be used to validate the built model and vice versa, the feature selection process confirmed their predictability. The present work has identified potential biomarkers for frailty prediction, which were conformed by the doctors. Most of the found predictors are variables describing the mobility, the mental state and the capability of performing daily tasks. According to the doctors, the variable describing the gender of each patient should be a predictor. However, the final predictor set does not contain the gender variable which could be due to redundancy. Maybe the contained information is covered by other variables such as height, which correlates strongly with gender. The feature selection algorithm may have discarded gender because of this. This manifests that further analysis with a bigger population is required in order to understand the role of this variable in particular but also for all the found potential predictors.

A very interesting finding seems to be that the feature representing the alkaline phosphatase levels was also found to be a suitable predictor. Given that it is a marker for inflammation this was also considered as plausible by the doctors. This feature is currently also being investigated by the FRAILOMIC initiative, which has the goal to find factors that are responsible for turning frailty into disability.

Predictive models, using the predictors obtained in the previous mentioned step, were built in order to predict frailty in patients. It was decided to derive a binary classifier which could predict the presence of frailty. The two classes are *non-frail* and *frail*. The classes *pre-frail* and *frail* from the original multiclass problem were fused to the class *frail* in order to work on a binary classification problem.

The main goal was to demonstrate that data mining and knowledge discovery

tools can be fruitfully applied in the frailty domain, which has been done in the scope of this thesis.

As a clear issue, the lack of enough data in order to build even more sophisticated and precise models remains.

7. Conclusions

In this thesis the feasibility of applying data mining techniques in order to extract models for frailty prediction using EHRs from patients some of which are frail, has been analysed.

From the work developed, it has been shown that in fact it is possible to extract meaningful patterns. Further, the importance of data preparation and data understanding for the successful extraction of predictive patterns has been demonstrated. Besides, it has been shown, that this is only feasible with the combined effort of the doctors and the data scientists.

Despite the importance of intelligent algorithms to extract the patterns, in this thesis we have additionally shown the paramount importance of pre-processing. Without a modest amount of effort in this phase, a reliable prediction model can not be built. Therefore, investing a lot of work in this phase proved to be highly beneficial in terms of accuracy and reliability of the obtained predictions.

Albeit the results seem to be very promising, for them to have more impact, it would be required to analyze a bigger cohort and to further validate the results with a different cohort of patients.

8. Future Work

This thesis has contributed towards the possibility of obtaining predictive models that can anticipate the onset of a disease. In particular, the problem of frailty has been analyzed in this work. However, for these models to be a reality, some work still needs to be done. This thesis opens new lines of research which will be reviewed in what follows.

8.1 Data View

Several issues make getting medical data still a hard task today. On the one hand, problems related to legal issues and all the issues concerning privacy and confidentiality and on the other hand, the problem of interoperability of systems make it difficult to have a complete view of the patient or to integrate data from different services at the hospital. Besides, one cannot forget the effort of obtaining a complete cohort of patients from which we can extract results. Consequently, in this thesis we would only analyze a cohort of 474 patients for which 284 variables were available. It would be desirable to have a bigger sample, so that results would become more significant and validations would be possible in different cohorts.

8.2 Technical View

Another future goal could be the automatic imputation of missing values in the EHRs, as it is a crucial but very time-consuming and complex part. It would also be interesting to determine and to analyse the best algorithm depending on the size of the data set. In this thesis the main focus was to show that data analysis is possible rather than showing which methods are the most efficient. Consequently,

in future work the feature selection process should be repeated once data of more patients is available.

All in all, one remaining task is removing step by step the expert from the deep processes of the data analysis pipeline by further developing the autonomy of the system. Another remaining task could be building a multi-class classification model for all 3 Fried stages (non-frail, pre-frail and frail) as in this work only the binary classification problem (frail/non-frail) was considered.

8.3 Medical View

Among the 284 features which were analyzed in this thesis, some potential predictor variables were not considered. In particular, the available information about medication intake (types of drugs, combination, etc.), which was not used in this work, could also be included in further investigations. As there are more parameters from the 2nd and 3rd clinical study waves on the way, future work could also focus on temporal analysis in order to be able to predict the evolution of patients regarding the frailty syndrome. All the available variables could be further, even more exhaustively, analyzed regarding their predictive potential. Moreover, the currently available data set may be enriched by nutritional and urinary data, as they potentially contain biomarkers of interest.

A. Appendix

A.1 Data Understanding

A.1.1 Tables

Attribute name	Values expected	Description	Type	How and when was it recorded?
ppeso	0,1	Fried criterium: weight loss >10 lbs. in past yr	categorical	calculated by hospital 2008
exhaustion	0,1	Fried criterium: exhaustion >=3days in past week	categorical	calculated by hospital 2008
pasefrag	0,1	Fried criterium: PASE <=20 percentile	categorical	calculated by hospital 2008
marchafragil	0,1	Fried criterium: time to walk >=80th percentile	categorical	calculated by hospital 2008
fuerzafragil	0,1	Fried criterium: grip strength <=20th percentile	categorical	calculated by hospital 2008
fragil	0,1,2	Frail status according to *Fried* scale	categorical	calculated by hospital 2008
ppeso_2013	0,1	Fried criterium: weight loss >10 lbs. in past yr	categorical	calculated by hospital 2013
exhaustion_2013	0,1	Fried criterium: exhaustion >=3days in past week	categorical	calculated by hospital 2013
pasefrag_2013	0,1	Fried criterium: PASE <=20 percentile	categorical	calculated by hospital 2013
marchafragil_2013	0,1	Fried criterium: time to walk >=80th percentile	categorical	calculated by hospital 2013
fuerzafragil_2013	0,1	Fried criterium: grip strength <=20th percentile	categorical	calculated by hospital 2013
fragil2013	0,1,2	Frail status according to *Fried* scale	categorical	calculated by hospital 2013

Table A.1: Features from the data set related to the *Fried* questions for determining the frailty status.

Attribute name	Values expected	Description	Type	How and when was it recorded?
ys1	yes, no	GDS1:Are you basically satisfied with your life?	binary	questionnaire answered by patient 2008
ys2	yes, no	GDS2:Have you dropped many of your activities and interests?	binary	questionnaire answered by patient 2008
ys3	yes, no	GDS3:Do you feel that your life is empty?	binary	questionnaire answered by patient 2008
ys4	yes, no	GDS4:Do you often get bored?	binary	questionnaire answered by patient 2008
ys5	yes, no	GDS5:Are you in good spirits most of the time?	binary	questionnaire answered by patient 2008
ys6	yes, no	GDS6:Are you afraid that something bad is going to happen to you?	binary	questionnaire answered by patient 2008
ys7	yes, no	GDS7:Do you feel happy most of the time?	binary	questionnaire answered by patient 2008
ys8	yes, no	GDS8:Do you often feel helpless?	binary	questionnaire answered by patient 2008
ys9	yes, no	GDS9:Do you prefer to stay at home, rather than going out and doing new things?	binary	questionnaire answered by patient 2008
ys10	yes, no	GDS10:Do you feel you have more problems with memory than most?	binary	questionnaire answered by patient 2008
ys11	yes, no	GDS11:Do you think it is wonderful to be alive now?	binary	questionnaire answered by patient 2008
ys12	yes, no	GDS12:Do you feel pretty worthless the way you are now?	binary	questionnaire answered by patient 2008
ys13	yes, no	GDS13:Do you feel full of energy?	binary	questionnaire answered by patient 2008
ys14	yes, no	GDS14:Do you feel that your situation is hopeless?	binary	questionnaire answered by patient 2008
ys15	yes, no	GDS15:Do you think that most people are better off than you are?	binary	questionnaire answered by patient 2008
gdstotal	0-15	GDS: Total Score	numeric	calculated by hospital 2008
depression	yes,no	gdstotal>=5	binary	calculated by hospital 2008

Table A.2: Features from the data set related to the Geriatric Depression Scale (*GDS*) questionnaire.

Attribute name	Values expected	Description	Type	How and when was it recorded?
drug_n_comercial_name	Name	Drug n commercial name	text	hospital 2008
drug_n_pa	Drug name	Drug n Active drug	text	hospital 2008
drug_n_atc	Code	Drug n ATC code	text	hospital 2008
drug_na	1,2,3,NAN	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	categorical	hospital 2008
drug_nb	1,2,3,NAN	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	categorical	hospital 2008

Table A.3: Medication related features from the data set: there are 11 ($n = \{1...11\}$) different drug attribute sets. All have the same format as in this table.

Attribute name	Values expected	Description	Type	How and when was it recorded?	Metabolic system
p1leu	4.5-11	Leucocytes [x10^9/L]	numeric	laboratory 2008	immune
p2hema	4-5	Erythrocytes [x10^{12}/L]	numeric	laboratory 2008	erythrocytes
p3hgb	12-15	Hemoglobyn [g/dL]	numeric	laboratory 2008	erythrocytes
p4hct	37-47	Hematocrit [%]	numeric	laboratory 2008	erythrocytes
p5vcm	80-99	Mean Corpuscular Volume (MCV) [fL]	numeric	laboratory 2008	erythrocytes
p6hcm	27-31	Mean Corpuscular Haemoglobin (MCH) [pg]	numeric	laboratory 2008	erythrocytes
p7chcm	33-37	Mean Corpuscular Haemoglobin Concentration (CHCM) [g/dL]	numeric	laboratory 2008	erythrocytes
p8ade	11.5-14.5	Red Cell Distribution Width (RDW) [%]	numeric	laboratory 2008	erythrocytes
p9lin	1-5	Lymphocytes [x10^9/L]	numeric	laboratory 2008	immune
p10mono	0.4-1.3	Monocytes [x10^9/L]	numeric	laboratory 2008	immune
p13eos	0.02-0.6	Eosinophiles [x10^9/L]	numeric	laboratory 2008	immune
p14baso	0-0.2	Basophiles [x10^9/L]	numeric	laboratory 2008	immune
p15dd	<500	D Dimer [μg/L]	numeric	laboratory 2008	coagulation
p16plaq	120-400	Platelets [x10^9/L]	numeric	laboratory 2008	coagulation
p17vpm	7-12	Mean Platelet Volume (MPV) [fl]	numeric	laboratory 2008	coagulation
p23glu	60-100	Glucose [mg/dL]	numeric	laboratory 2008	sugars
p24urea	10-71	Urea [mg/dL]	numeric	laboratory 2008	nephritic
p25acur	2.4-5.7	Uric acid [mg/dL]	numeric	laboratory 2008	nephritic
p26crea	0.5-0.9	Creatinine [mg/dL]	numeric	laboratory 2008	nephritic
p27prot	6.4-8.3	Protein [g/dL]	numeric	laboratory 2008	proteins
p28albu	3.4-4.8	Albumin [g/dL]	numeric	laboratory 2008	proteins
p30chol	110-230	Cholesterin [mg/dL]	numeric	laboratory 2008	fats
p31trig	60-200	Triglycerides [mg/dL]	numeric	laboratory 2008	fats
p32ca	8.4-10.2	Calcium (Ca) [mg/dL]	numeric	laboratory 2008	minerals
p33p	2.7-4.5	Phosphorus (P) [mg/dL]	numeric	laboratory 2008	minerals
p34na	132-146	Sodium (Na) [mEq/L]	numeric	laboratory 2008	minerals
p35k	3.7-5.4	Potassium (K) [mEq/L]	numeric	laboratory 2008	minerals
p36cl	94-110	Chloride (Cl) [mEq/L]	numeric	laboratory 2008	minerals
p37got	5-37	Glutamic-Oxaloacetic Transaminase (GOT) [U/L]	numeric	laboratory 2008	hepatic
p38gpt	5-40	Glutamic-Pyruvic Transaminase (GPT) [U/L]	numeric	laboratory 2008	hepatic
p39ggt	5-39	Gamma-Glutamyl Transferase (GGT) [U/L]	numeric	laboratory 2008	hepatic
p40falc	35-104	Alkaline phosphatase [U/L]	numeric	laboratory 2008	hepatic / nephritic
p41ldh	230-530	Lactate dehydrogenase (LDH) [U/L]	numeric	laboratory 2008	general
p42fe	40-145	Iron (FE) [μg/dL]	numeric	laboratory 2008	minerals
p43tfrr	200-360	Transferrin [mg/dL]	numeric	laboratory 2008	general
p44pcrh	<9	High-sensitivity C-reactive protein (hs-CRP) [mg/L]	numeric	laboratory 2008	cardiac
IGF1	50-300	Insulin like growth factor 1 (IGF1) [ng/mL]	numeric	laboratory 2008	growth
E2	0-200	17β-estradiol (E2) [pmol/L]	numeric	laboratory 2008	hormones
Dheas	0-200	Dehydroepiandrosterone sulfate (DHEA-S) [μg/dL]	numeric	laboratory 2008	homones
Dhea	0-10	Dehydroepiandrosterone (DHEA) [ng/mL]	numeric	laboratory 2008	homones
HDL	0-200	High-density lipoprotein (HDL) [mg/dL]	numeric	laboratory 2008	fats
LDL	0-200	Low-density lipoprotein (LDL) [mg/dL]	numeric	laboratory 2008	fats
INSULINA	0-2000	Insulin [U/mL]	numeric	laboratory 2008	sugars
ADMA	50-150	Asymmetric dimethylarginine (ADMA) [μmol/L]	numeric	laboratory 2008	proteins
TESTOTOTAL	0-1000	Total testosterone [ng/dL]	numeric	laboratory 2008	homones
TESTOLIBRE	0-10	Free testosterone [ng/dL]	numeric	laboratory 2008	homones

Table A.4: Blood related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
k1	111,222,333	WHO activity 6: Any difficulty washing face and arms?	categorical	questionnaire answered by patient 2008
k2	111,222,333	WHO activity 8: Any difficulty dressing and undressing?	categorical	questionnaire answered by patient 2008
k3	111,222,333	WHO activity 11: Any difficulty using the toilet?	categorical	questionnaire answered by patient 2008
k4	111,222,333	WHO activity 12: Any difficulty getting in and out of bed?	categorical	questionnaire answered by patient 2008
k5	111,222,333	WHO activity 19: Any difficulty controlling urination and bowel movements?	categorical	questionnaire answered by patient 2008
k6	111,222,333	WHO activity 9: Any difficulty eating (e.g.,holding a fork, cutting food, drinking from a glass)?	categorical	questionnaire answered by patient 2008
katz2008	0-6	Number of ADL abilities	numeric	calculated by hospital 2008
k1_2013	1,2,3	WHO activity 6: Any difficulty washing face and arms?	categorical	questionnaire answered by patient 2013
k2_2013	1,2,3	WHO activity 8: Any difficulty dressing and undressing?	categorical	questionnaire answered by patient 2013
k3_2013	1,2,3	WHO activity 11: Any difficulty using the toilet?	categorical	questionnaire answered by patient 2013
k4_2013	1,2,3	WHO activity 12: Any difficulty getting in and out of bed?	categorical	questionnaire answered by patient 2013
k5_2013	1,2,3	WHO activity 19: Any difficulty controlling urination and bowel movements?	categorical	questionnaire answered by patient 2013
k6_2013	1,2,3	WHO activity 9: Any difficulty eating (e.g.,holding a fork, cutting food, drinking from a glass)?	categorical	questionnaire answered by patient 2013
katz2013	0-6	Number of ADL abilities	categorical	calculated by hospital 2013

Table A.5: Activities of Daily Living questionnaire (ADL) related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
lw1	111,222,333,444	WHO activity 20: Any difficulty using the telephone?	categorical	questionnaire answered by patient 2008
lw2	111,222,333,444	WHO activity 5: Any difficulty shopping daily for basic necessities?	categorical	questionnaire answered by patient 2008
lw3	111,222,333,444	WHO activity 10: Any difficulty cooking a simple meal?	categorical	questionnaire answered by patient 2008
lw4	111,222,333,444,555	WHO activity 13: Any difficulty doing light housework (e.g., doing dishes, light cleaning)?	categorical	questionnaire answered by patient 2008
lw5	111,222,333	WHO activity 14: Any difficulty doing heavy housework (e.g., washing windows, floor)?	categorical	questionnaire answered by patient 2008
lw6	111,222,333,444,555	WHO activity 22: Any difficulty using public transportation?	categorical	questionnaire answered by patient 2008
lw7	111,222,333	WHO activity 23: Any difficulty taking medications correctly?	categorical	questionnaire answered by patient 2008
lw8	111,222,333	WHO activity 24: Any difficulty managing home finances?	categorical	questionnaire answered by patient 2008
lawton2008	0-8	Number of IADL abilities (0-8)	numeric	calculated by hospital 2008
lw1_2013	1,2,3,4	WHO activity 20: Any difficulty using the telephone?	categorical	questionnaire answered by patient 2013
lw2_2013	1,2,3,4	WHO activity 5: Any difficulty shopping daily for basic necessities?	categorical	questionnaire answered by patient 2013
lw3_2013	1,2,3,4	WHO activity 10: Any difficulty cooking a simple meal?	categorical	questionnaire answered by patient 2013
lw4_2013	1,2,3,4,5	WHO activity 13: Any difficulty doing light housework (e.g., doing dishes, light cleaning)?	categorical	questionnaire answered by patient 2013
lw5_2013	1,2,3	WHO activity 14: Any difficulty doing heavy housework (e.g., washing windows, floor)?	categorical	questionnaire answered by patient 2013
lw6_2013	1,2,3,4,5	WHO activity 22: Any difficulty using public transportation?	categorical	questionnaire answered by patient 2013
lw7_2013	1,2,3	WHO activity 23: Any difficulty taking medications correctly?	categorical	questionnaire answered by patient 2013
lw8_2013	1,2,3	WHO activity 24: Any difficulty managing home finances?	categorical	questionnaire answered by patient 2013
lawton2013	0-8	Number of IADL abilities	categorical	physician 2013

Table A.6: Instrumental Activities of Daily Living ($IADL$) questionnaire related features from the data set

Attribute name	Values expected	Description	Type	How and when was it recorded?
alch1	0-8	How many drinks do you have?	categorical	questionnaire answered by patient 2008
alch1a1	$\mathbb{N}+$	how many glasses of wine do you drink daily?	numeric	questionnaire answered by patient 2008
alch1a2	$\mathbb{N}+$	how many glasses of beer do you drink daily?	numeric	questionnaire answered by patient 2008
alch1a3	$\mathbb{N}+$	how many glasses of spirits do you drink daily?	numeric	questionnaire answered by patient 2008
alch1b	0-age	For how many years?	numeric	questionnaire answered by patient 2008
alch2	yes,no	did you drink previously?	binary	questionnaire answered by patient 2008
alch2a	1-5	Kind of drinker	categorical	questionnaire answered by patient 2008
alch2b	1-8	Starting age	categorical	questionnaire answered by patient 2008
alch2c	1-8	Ending age	categorical	questionnaire answered by patient 2008
tab1	1-4	Have you smoked at least 100 cigarettes in your entire life?	categorical	questionnaire answered by patient 2008
tab1a	1-3	If yes, Did you smoke cigarettes daily, occasionally, or not at all?	categorical	questionnaire answered by patient 2008
tab1a1	1-3	Do you smoke actually?	categorical	questionnaire answered by patient 2008
tab1a1a	1-8	If not, How many time have you stopped smoking?	categorical	questionnaire answered by patient 2008
tab1a3	0-age	For how many years did you smoke?	numeric	questionnaire answered by patient 2008
@1_year_smoker	yes, no	smoker for at least one year	binary	physician 2008
current_smoker	yes, no	current smoker	binary	physician 2008

Table A.7: Consumption related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
fuerza1a	0-60	Muscle strength (upper) with dynamometer: hand grip dominant limb (kg)	numeric	physician 2008
peso1	30-200	Weight (kg)	numeric	physician 2008
altura1	120-220	Height (cm)	numeric	physician 2008
ppca	20-200	Anthropometry: hip perimeter (cm)	numeric	physician 2008
pasetotal	0-400+	Physical activity scale for elderly score	numeric	questionnaire answered by patient 2008
codigo01	alive,death	Dead at follow up?	binary	physician 2008

Table A.8: Physique related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
ccv1	yes,no	Myocardial infarction / Heart attack (self reported)+E148	binary	questionnaire answered by patient 2008
ccv2	yes,no	Congestive heart failure (self reported)	binary	questionnaire answered by patient 2008
ccv4	yes,no	Angina pectoris (self reported)	binary	questionnaire answered by patient 2008
ccv6	yes,no	Hypertension (self-report,drugs,BP tests)	binary	questionnaire answered by patient 2008
ccv8	yes,no	Diabetes mellitus (self reported, drugs)	binary	questionnaire answered by patient 2008
cv1cv4	yes,no	Myocardial infarction / Heart attack (self reported)/angina pectoris	binary	questionnaire answered by patient 2008
charlsonindex	0-37	Charlson co-morbidity index	categorical	physician 2008

Table A.9: Comorbidity related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
enpot1	0,1,NaN	What day of the week is this? (MMSE question)	categorical	questionnaire answered by patient 2008
enpot2	0,1,NaN	What is today's date? (MMSE question)	categorical	questionnaire answered by patient 2008
enpot3	0,1,NaN	What month is this? (MMSE question)	categorical	questionnaire answered by patient 2008
enpot4	0,1,NaN	What year is this? (MMSE question)	categorical	questionnaire answered by patient 2008
enpot6	0,1,NaN	Which season is this? (MMSE question)	categorical	questionnaire answered by patient 2008
enpol1	0,1,NaN	IN HOME: What is the street address of this house? // IN FACILITY: What is the name of this building? (MMSE question)	categorical	questionnaire answered by patient 2008
enpol2	0,1,NaN	IN HOME: What room are we in? // IN FACILITY: What floor are we on? (MMSE question)	categorical	questionnaire answered by patient 2008
enpol3	0,1,NaN	What city/town are we in? (MMSE question)	categorical	questionnaire answered by patient 2008
enpol4	0,1,NaN	What province are we in? (MMSE question)	categorical	questionnaire answered by patient 2008
enpol5	0,1,NaN	What county are we in? (MMSE question)	categorical	questionnaire answered by patient 2008
enpmem1a	1,2,3,4,NaN	SAY: I am going to name three objects. When I am finished, I want you to repeat theM. Remember what they are because I am going to ask you to name them again in a few minutes. // Say the following words slowly at 1-second intervals - peseta (coin in spanish), caballo (horse in spanish), manzana (apple in spanish) (MMSE question)	categorical	questionnaire answered by patient 2008
enpat2	1,2,3,4,5,6,NaN	Spell the word MUNDO (world in spanish). Now spell it backwards.	categorical	questionnaire answered by patient 2008
enpat1	1,2,3,4,5,6,NaN	Count backwards by 7 starting from 100	categorical	questionnaire answered by patient 2008
enpmem2	1,2,3,4,NaN	Now what were the three objects I asked you to remember?	categorical	questionnaire answered by patient 2008
enpleng1	1,2,3,NaN	Show a wristchatch and a pencil. What are these called?	categorical	questionnaire answered by patient 2008
enpleng2	1,2,NaN	SAY: I would like you to repeat this phrase after me: Ni si, ni no, ni pero. (No ifs, ands or buts. In spanish)	categorical	questionnaire answered by patient 2008
enpleng4	1,2,NaN	SAY: Read the words on the page and then do what it says. Then hand the person the sheet with "Cierre los ojos" (close your eyes in spanish) on it. If the subject read and does not close their eyes, repeat yp to three times. Score only if subject closes eyes.	categorical	questionnaire answered by patient 2008
enpprx1	1,2,NaN	Hand the person a pencil and paper. SAY: write any complete sentence on that piece of paper. (Note: The sentence must make sense. Ignore spelling errors)	categorical	questionnaire answered by patient 2008
enpprx2	1,2,NaN	Place design, eraser and pencil in front of the person. SAY: copy this design please. // Allow multiple tries. Wait until person is finished and hands it back. Score only for correctly copied diagram with a 4-sided figure between two 5-sided figures.	categorical	questionnaire answered by patient 2008
enpleng3	1,2,3,4,NaN	Ask the person if he is right or left handed. Take a piece of paper and hold it up in front of the person. SAY: Take this paper in your right/left hand (whichever is non-dominant), fold the paper in half once with both hands and put the paper down on the floor. Score 1 point for each instruction executed correctly.	categorical	questionnaire answered by patient 2008
mmse2009	0-30	MMSE raw score	numeric	calculated by hospital 2008
cognitive _impairment _mmse _educative _level	yes,no	Has the patient a cognitive impairment?	binary	determined by physician 2008

Table A.10: Mini-Mental-State-Examination ($MMSE$) related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
cq8	yes, no, 88 = NA	Leukemia or Polycytemia	categorical	physician 2008
cq9	yes, no, 88 = NA	Lymphoma	categorical	physician 2008
cq10	yes, no, 88 = NA	Cancer (except Leukemia, polycythemia and lymphoma)	categorical	physician 2008
cq6	1,2.3,4	Did any doctor tell you that you had Alzheimer's disease, senile dementia or another dementia?	categorical	questionnaire answered by patient 2008
cq6a	1-10	What kind of dementia did your say doctor that you had?	categorical	questionnaire answered by patient 2008
reum1	1,2.3,4	Have you ever had any joint inflammated for more than 4 weeks in a row?	categorical	questionnaire answered by patient 2008
reum2	1,2.3,4	Have you ever felt pain in any joint for more than 4 weeks in a row?	categorical	questionnaire answered by patient 2008
reum3	1,2.3,4	Do you ever feel that you can't move or feel rigid for over half an hour during the morning?	categorical	questionnaire answered by patient 2008
reum4	1,2.3,4	Have you ever been told you have arthritis?	categorical	questionnaire answered by patient 2008
reum5	1-9	Please select in the mannequin the joints in which you have had or have now inflammation for more than 4 weeks in a row (note the location of the affected joints). SHOW CARD 2.	categorical	questionnaire answered by patient 2008
reum6	1,2.3,4	Do you feel pain or have inflammation in any joint?	categorical	questionnaire answered by patient 2008
reum6a	1-9	If yes,Please, show which joints. SHOW CARD 2:	categorical	questionnaire answered by patient 2008
reum7	1-6	Did any doctor tell you that you had arthritis or arthrosis in your..?	categorical	questionnaire answered by patient 2008
reum7a	1,2.3,4	if yes (1, 2 or 3)The doctor said that you had it after a hip or knee radiography, or both?	categorical	questionnaire answered by patient 2008
epoc1	1,2.3,4	Did any doctor tell you that you had a chronic obstructive pulmonary disease: emphysema or chronic bronchitis?	categorical	questionnaire answered by patient 2008
epoc2	1,2.3,4	Did any doctor say tell that you had asthma?	categorical	questionnaire answered by patient 2008
epoc3	1,2.3,4	Did any doctor tell you that you had any lung disease?	categorical	questionnaire answered by patient 2008
epoc4	1,2.3,4	Did any doctor tell you that you had had a pneumonía or bronchopneumonía?	categorical	questionnaire answered by patient 2008
epoc5	1,2.3,4	Did any doctor tell you that you had had an acute bronchitis?	categorical	questionnaire answered by patient 2008
epoc6	1,2.3,4	Have you ever been operated of your lung?	categorical	questionnaire answered by patient 2008
epoc7	1,2.3,4	Do you have any other lung disease?	categorical	questionnaire answered by patient 2008

Table A.11: Disease related features from the data set

Attribute name	Values expected	Description	Type	How and when was it recorded?
em1	yes,no	Are you able to walk at home?	binary	questionnaire answered by patient 2008
em1a	yes,no	If answered YES; Do you get tired when doing it?	binary	questionnaire answered by patient 2008
em1b	yes,no	If answered YES; Do you need help when doing it?	binary	questionnaire answered by patient 2008
em2	yes,no	Are you able to go out from home?	binary	questionnaire answered by patient 2008
em2a	yes,no	If answered YES; Do you get tired when doing it?	binary	questionnaire answered by patient 2008
em2b	yes,no	If answered YES; Do you need help when doing it?	binary	questionnaire answered by patient 2008
em3	yes,no	Are you able to climb stairs?	binary	questionnaire answered by patient 2008
em3a	yes,no	If answered YES; Do you get tired when doing it?	binary	questionnaire answered by patient 2008
em3b	yes,no	If answered YES; Do you need help when doing it?	binary	questionnaire answered by patient 2008
em4	yes,no	Are you able to walk outside (nice weather)?	binary	questionnaire answered by patient 2008
em4a	yes,no	If answered YES; Do you get tired when doing it?	binary	questionnaire answered by patient 2008
em4b	yes,no	If answered YES; Do you need help when doing it?	binary	questionnaire answered by patient 2008
em5	yes,no	Are you able to walk outside (bad weather)?	binary	questionnaire answered by patient 2008
em5a	yes,no	If answered YES; Do you get tired when doing it?	binary	questionnaire answered by patient 2008
em5b	yes,no	If answered YES; Do you need help when doing it?	binary	questionnaire answered by patient 2008

Table A.12: Mobility Scale (MS) related features from the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
hi1	7-8 digit number	ETES ID	numeric	assigned and recorded by the hospital
frailomic_code	"TO" + hi1	FRAILOMIC ID	2 constant characters + numeric	assigned and recorded by the hospital
Parma_serum_code	Code	Parma Serum code	text	assigned and recorded by the hospital
Parma_Edta_code	Code	Parma EDTA Code	text	assigned and recorded by the hospital
Jena_Edta_code	Code	Jena EDTA Code	text	assigned and recorded by the hospital
Evercyte_Edta_code	Code	Evercyte EDTA Code	text	assigned and recorded by the hospital
Cardiff_serum_code	Code	Cardiff Serum Code	text	assigned and recorded by the hospital
Cardiff_Edta_code	Code	Cardiff EDTA Code	text	assigned and recorded by the hospital
EV_Edta_code	Code	EV EDTA Code	text	assigned and recorded by the hospital

Table A.13: Codes and IDs of the hospital which appear in the data set.

Attribute name	Values expected	Description	Type	How and when was it recorded?
hi8	0-130	Age in years	numeric	physician 2008
hi11	male, female	Gender	binary	physician 2008
individualincome	1-12	Income of the individual	categorical	physician 2008
householdincome	1-15	Income of the household in which the individual lives	categorical	physician 2008
numpersonsfamilyunit	1-10	Number of persons in the family	categorical	physician 2008

Table A.14: Features related to demographic properties of the patients.

Attribute name	Values expected	Description	Type	How and when was it recorded?
ekg1	40-200	EKG: Heart rate (beats/minute)	numeric	physician 2008
tadd	40-140	Pressure arterial. Diastolic	numeric	physician 2008
tads	80-260	Pressure arterial. Systolic	numeric	physician 2008

Table A.15: Features related to cardiac properties of the patients.

Attribute name	Values expected	Description	Type	How and when was it recorded?
ps1	1-6	How would you evaluate your current health? How do you feel now?	categorical	questionnaire answered by patient 2008
ps2	1-6	How is your health compared to 1 yr ago?	categorical	questionnaire answered by patient 2008
ps3	1-6	How would you judge your health compared to other people of your same age?	categorical	questionnaire answered by patient 2008

Table A.16: Features related to self reported health status of the patients.

A.1.2 Statistical Analysis

Table A.17: Description of $Frailomic\_code$

$Frailomic\_code$	
Meaning	This feature contains the Frailomic Code for each patient and has no relevance for the data analysis as it was assigned from the hospital for organizational purposes.

Table A.18: Description of $hi1$

$hi1$	
Meaning	This feature contains the ETES ID for each patient and has no relevance for the data analysis as it was assigned from the hospital for organizational purposes.

Table A.19: Description of *hi*8

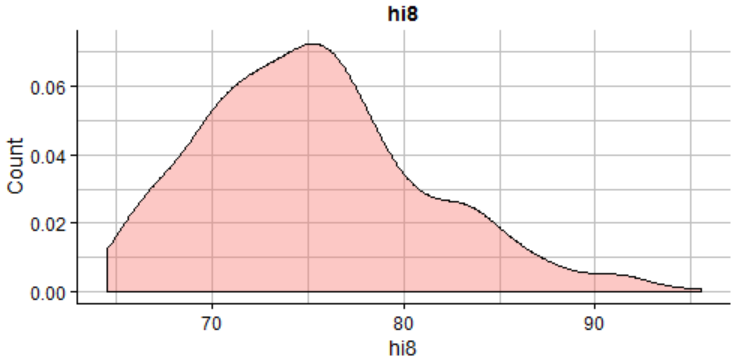
*hi*8	
Meaning	This feature represents the age in years for each patient. For the study only participants with age 65+ were used.
Statistics	min: 65.00; max: 95.00; average: 75.29; median: 75.00; σ^2: 33.44; σ: 5.78; # missings: 0.00; % missings: 0.00
Distribution	

Table A.20: Description of $hi11$

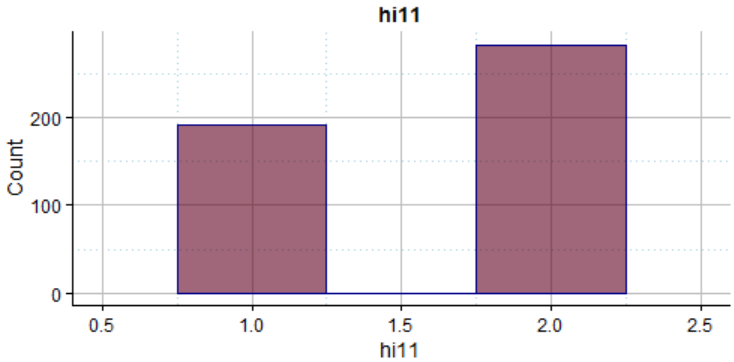
$hi11$	
Meaning	This feature gives binary information about the gender of the patients.
Statistics	mode: 2.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	1: male 2: female

Table A.21: Description of $ps1$

$ps1$	
Meaning	This feature gives categorical information about the current health status of the patient in his view. Asked question: "How would you evaluate your current health? How do you feel now?"
Statistics	mode: 2.00; levels: 7.00; # missings: 2.00; % missings: 0.42

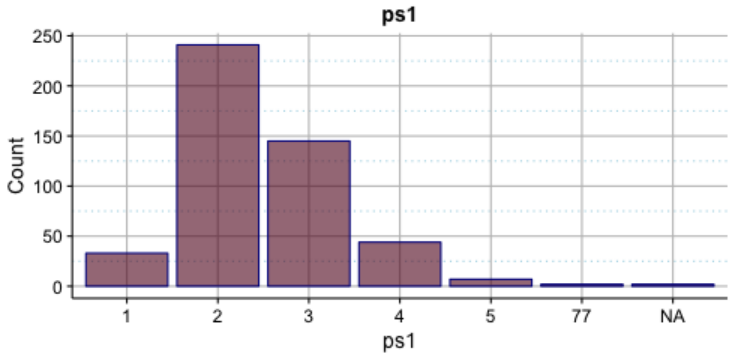
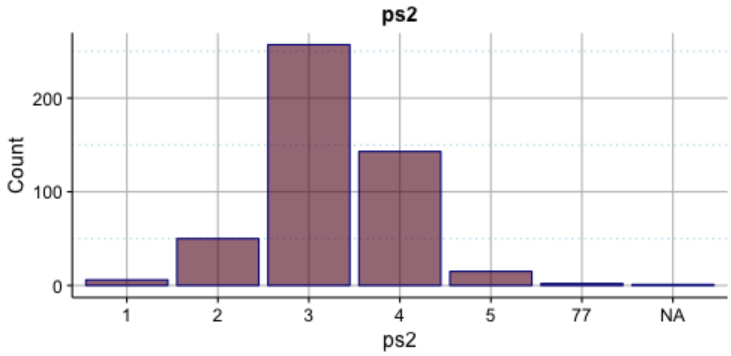
Distribution	
Discretization & Semantic scales	1: Very good 2: Good 3: Fair (so-so) 4: Poor 5: Very poor 6: Undetermined 77: not available (?)

Table A.22: Description of *ps*2

*ps*2	
Meaning	This feature gives categorical information about the current health status compared to one year ago in the view of the patient. Asked question: "How is your health compared to 1 year ago?"
Statistics	mode: 3.00; levels: 7.00; # missings: 1.00; % missings: 0.21
Distribution	

Discretization & Semantic scales	1: Much better 2: Better 3: The same 4: Slightly worse 5: Much worse 6: Undetermined 77: not available (?)

Table A.23: Description of $ps3$

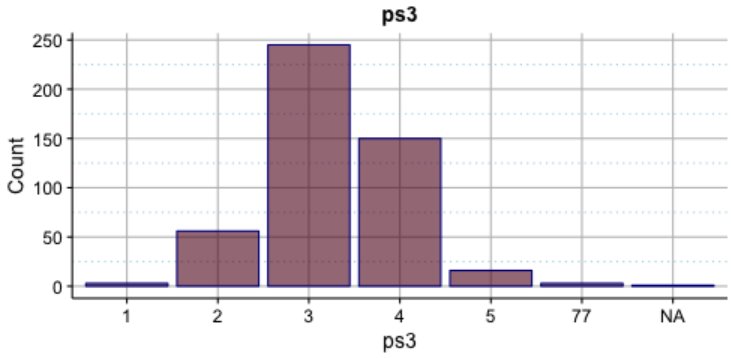
$ps3$	
Meaning	This feature gives categorical information about the current health status of the patient compared to other people with the same age in the view of the patient. Asked question: "How would you judge your health compared to other people of your same age?"
Statistics	mode: 3.00; levels: 7.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Much worse 2: Sligthly worse 3: The same 4: Better 5: Much better 6: Undetermined 77: not available (?)

Table A.24: Description of $ps4$

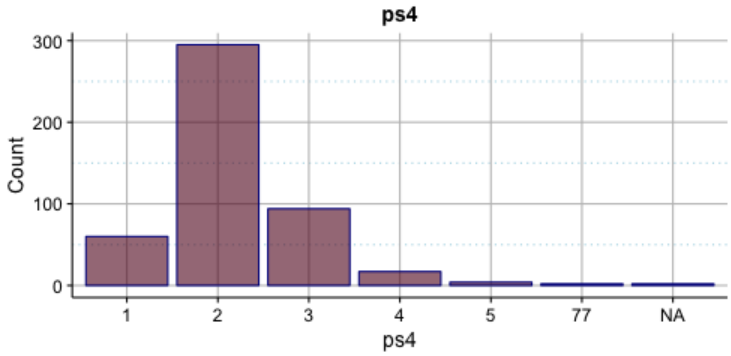
$ps4$	
Meaning	This feature gives categorical information about the current state of happiness of the patient compared to other people with the same age in the view of the patient. Asked question: "Are you happy in general?"
Statistics	mode: 2.00; levels: 7.00; # missings: 2.00; % missings: 0.42
Distribution	
Discretization & Semantic scales	1: Very happy 2: Happy 3: Not happy nor unhappy 4: Unhappy 5: Very unhappy 6: Undetermined/Not applicable 77: not available (?)

Table A.25: Description of $ps5$

$ps5$	
Meaning	This feature gives categorical information about the current state of satisfaction of the patient. Asked question: "If you are thinking about you life till now, how satisfied are you?"
Statistics	mode: 2.00; levels: 7.00; # missings: 2.00; % missings: 0.42

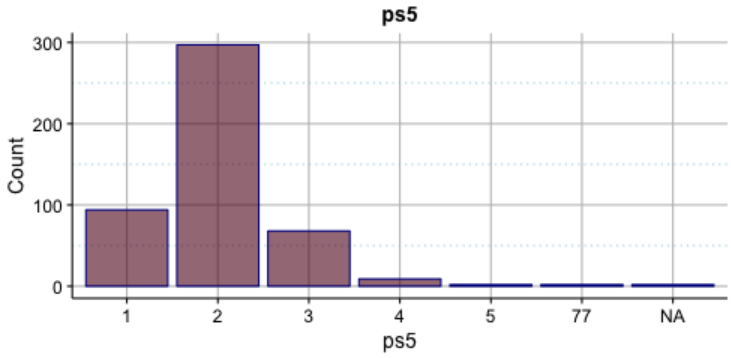
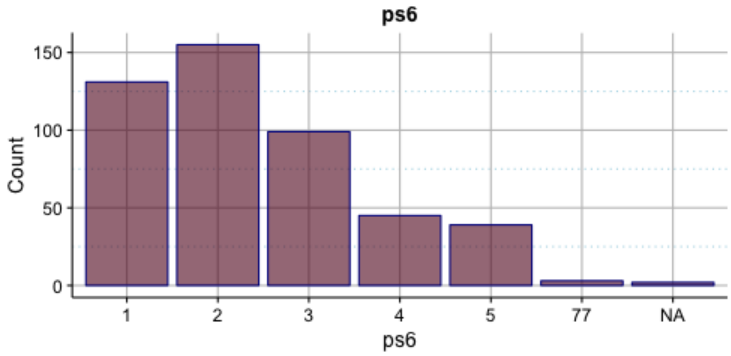
Distribution	
Discretization & Semantic scales	1: Very satisfied 2: Satisfied 3: Not satisfied nor unsatisfied 4: Unsatisfied 5: Very unsatisfied 6: Undetermined/Not applicable 77: not available (?)

Table A.26: Description of *ps*6

*ps*6	
Meaning	This feature gives categorical information about the capacity of dealing with problems of the patient. Asked question: "Are you feeling incapable of tackling problems in your life?".
Statistics	mode: 2.00; levels: 7.00; # missings: 2.00; % missings: 0.42
Distribution	

Discretization & Semantic scales	1: Never 2: Almost never 3: Sometimes 4: Many times (frequently) 5: Often (very frequently) 6: Undetermined/Not applicable 77: not available (?)

Table A.27: Description of *ps7*

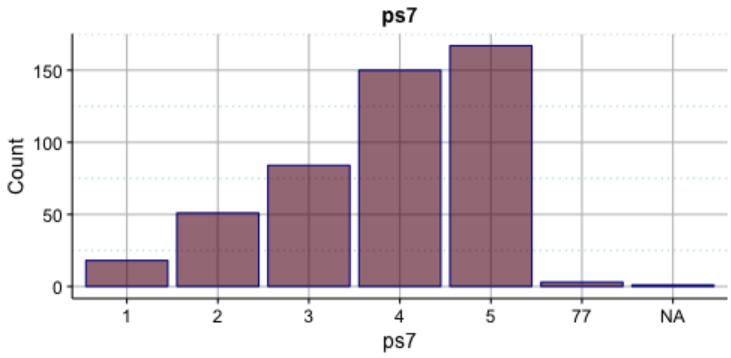
ps7	
Meaning	This feature gives categorical information about the capacity of dealing with tasks of the patient. Asked question: "Are you feel capable of tackling every task you would like to?".
Statistics	mode: 5.00; levels: 6.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Never 2: Almost never 3: Sometimes 4: Many times (frequently) 5: Often (very frequently) 6: Undetermined/Not applicable 77: not available (?)

Table A.28: Description of *ps*8

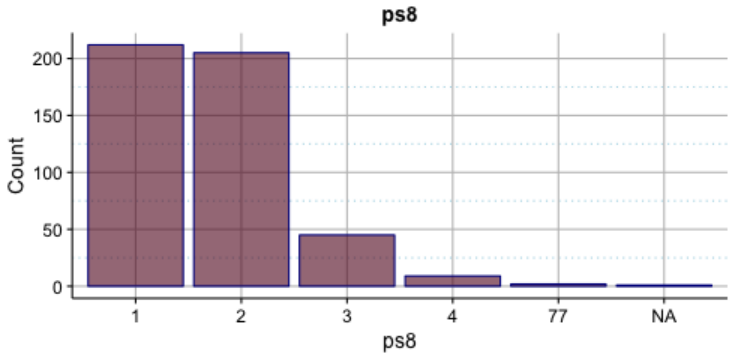
*ps*8	
Meaning	This feature gives categorical information about the felt pain in the last week of the patient. Asked question: "During the last week, did you feel physical pain?".
Statistics	mode: 1.00; levels: 5.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: No pain 2: Slight pain, did not influence daily tasks 3: Pain which interfered with daily tasks 4: Heavy pain, which forced me to stay in bed or seated 5: Undetermined/Not applicable 77: not available (?)

Table A.29: Description of *ps*9

*ps*9	
Meaning	This feature gives categorical information about the frequency of visiting the general practitioner. Asked question: "During the last month, how many times did you visit the general practitioner because of being sick?".
Statistics	mode: 1.00; levels: 6.00; # missings: 3.00; % missings: 0.63

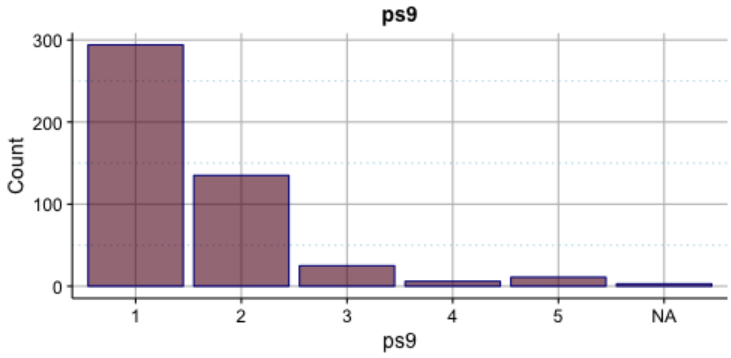
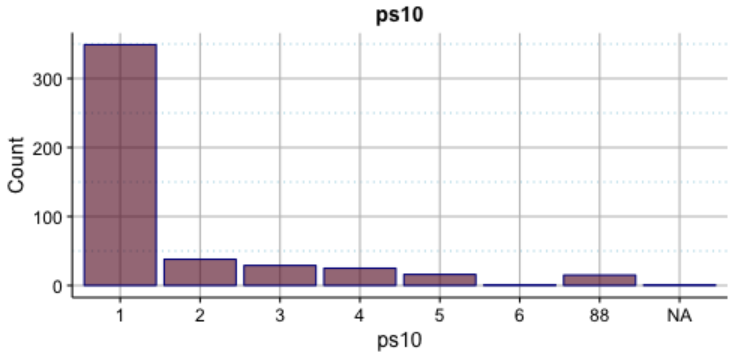
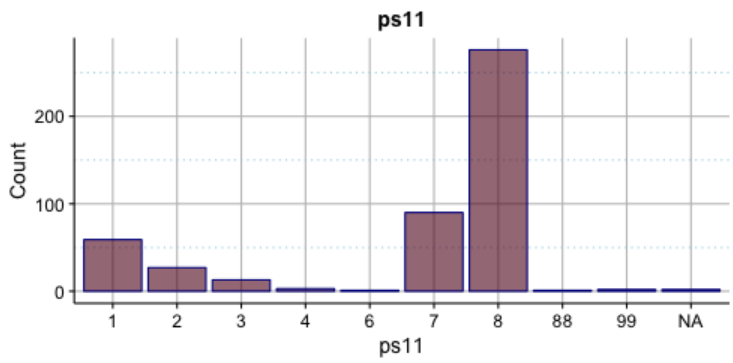
Distribution	
Discretization & Semantic scales	1: Never 2: One time 3: Two times 4: Three times 5: Four or more times 77: not available (?)

Table A.30: Description of $ps10$

$ps10$	
Meaning	This feature gives categorical information about the time which has past since the patient has spoken to a medical professional about his/her health. Asked question: "When was the last time that you visited a medical doctor or another medical professional in order to speak about your health?".
Statistics	mode: 1.00; levels: 7.00; # missings: 1.00; % missings: 0.21
Distribution	

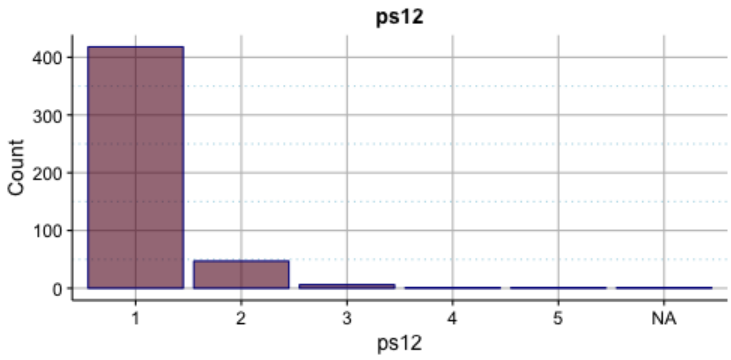
Discretization & Semantic scales	1: Three month or less 2: More than three months but less than 6 months 3: More than 6 months but less than 12 months 4: More than one year but less then 3 years 5: More than 3 years 6: Never 7: NS 8: NC

Table A.31: Description of $ps11$

$ps11$	
Meaning	This feature gives categorical information about the time which the patient needed to process a trauma which has happened in his life. Asked question: "Think about the most painful/woebegone event which has happened in the last ten years. How much time did you need to recover from it?".
Statistics	mode: 8.00; levels: 10.00; # missings: 2.00; % missings: 0.42
Distribution	

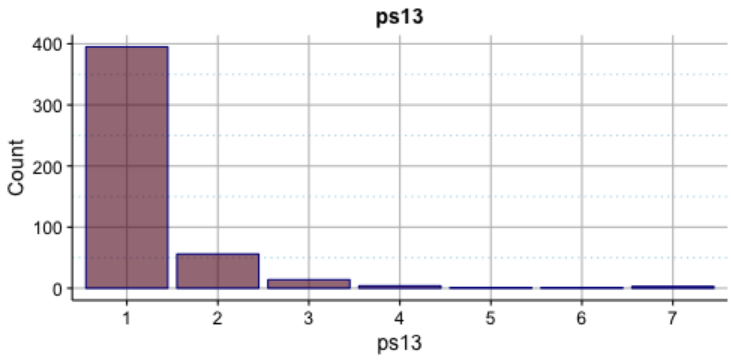
Discretization & Semantic scales	1: Less than 6 months 2: Between 6 and 12 months 3: Between 1 and 2 years 4: Between 2 and 4 years 5: Between 4 and 6 years 6: More than 6 years 7: Not recovered yet 8: Doesn't know 9: Doesn't respond

Table A.32: Description of $ps12$

$ps12$	
Meaning	This feature gives categorical information about how often the patient was hospitalized in the last year. Asked question: "During the last 12 months, how many times have you been hospitalized (over night)?".
Statistics	mode: 1.00; levels: 6.00; # missings: 1.00; % missings: 0.21
Distribution	

Discretization & Semantic scales	1: Never 2: Once 3: Twice 4: 3 times 5: 4 or more times 6: Doesn't know 7: Doesn't respond

Table A.33: Description of $ps13$

$ps13$	
Meaning	This feature gives categorical information about how often the patient visited the hospital in the last year because of a case of need. Asked question: "During the last 12 months, how many times did you visit the hospital because of an emergency (without spending the night)?".
Statistics	mode: 1.00; levels: 8.00; # missings: 0.00; % missings: 0.00
Distribution	

Discretization & Semantic scales	1: Never 2: Once 3: Twice 4: 3 times 5: 4 times 6: 5 times 7: 6 or more times 8: Doesn't know 9: Doesn't respond

Table A.34: Description of $ps14$

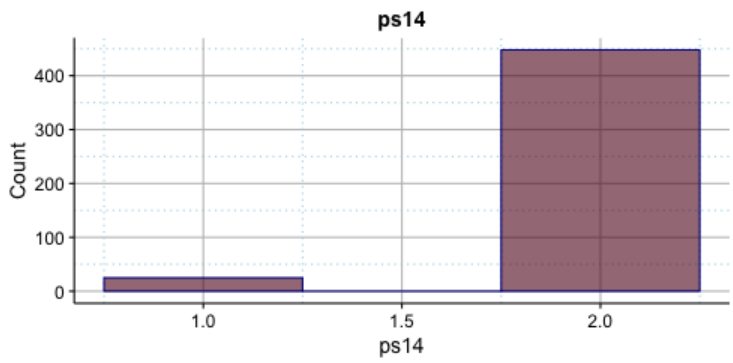
$ps14$	
Meaning	This feature gives binary information about if the patient visited an institution for rehabilitation. Asked question: "During the last 12 months, where you patient in a rehabilitation center (with spending the night)?".
Statistics	mode: 2.00; levels: 3.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Yes 2: No 7: Doesn't know 8: Doesn't respond

Table A.35: Description of $ps14a$

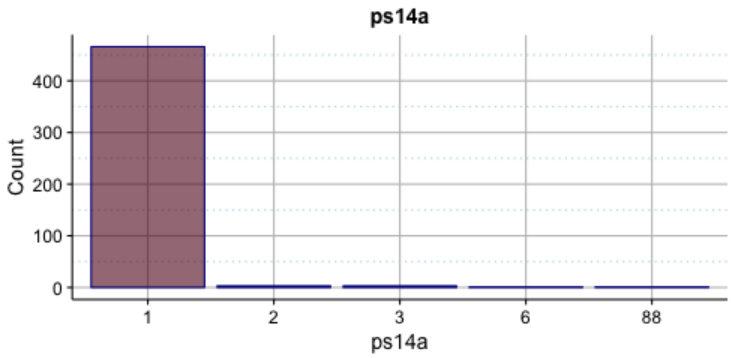
$ps14a$	
Meaning	This feature gives categorical information about for how much time the patient visited an institution for rehabilitation. Asked question: "During the last 12 months, how much time did you spend in an institution for physical therapy (with spending the night)?".
Statistics	mode: 1.00; levels: 5.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	1: Never 2: Less than 15 days 3: Between 15 and 30 days 4: Between 30 and 60 days 5: Between 60 and 90 days 6: More than 90 days 7: Doesn't know 8: Doesn't respond

Table A.36: Description of $ccv1$

$ccv1$	
Meaning	This feature gives binary information about if the patient had a Myocardial infarction or a Heart attack (self reported).

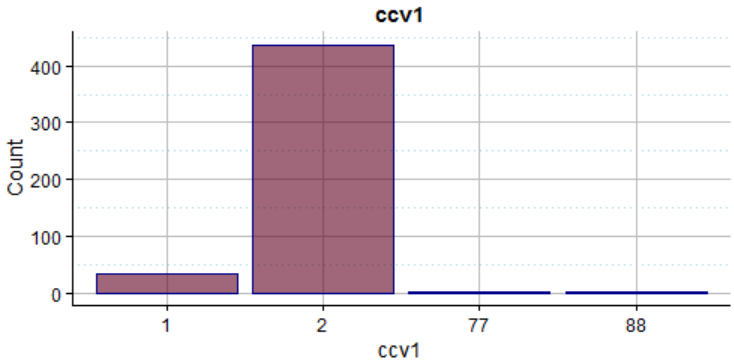
Statistics	mode	levels	# missings	% missings
	2.00	4.00	0.00	0.00
Distribution				
Discretization & Semantic scales	1: Yes 2: No			

Table A.37: Description of *ccv*2

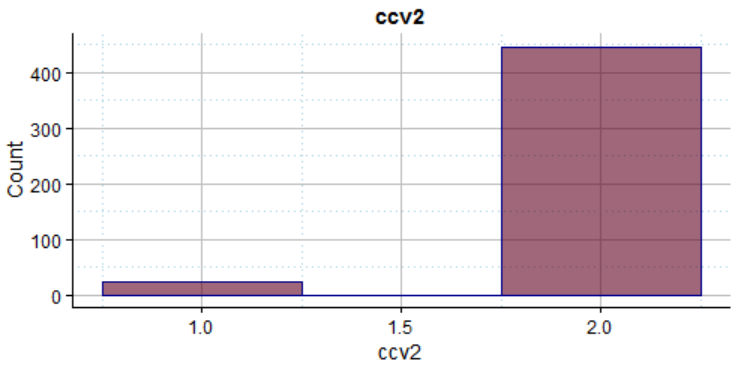
*ccv*2				
Meaning	This feature gives binary information about Congestive heart failure (self reported)			
Statistics	mode	levels	# missings	% missings
	2.00	3.00	5.00	0.01
Distribution				
Discretization & Semantic scales	1: Yes 2: No			

Table A.38: Description of *ccv*4

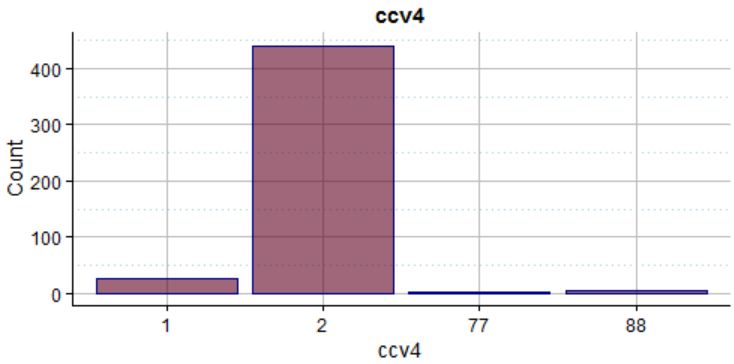
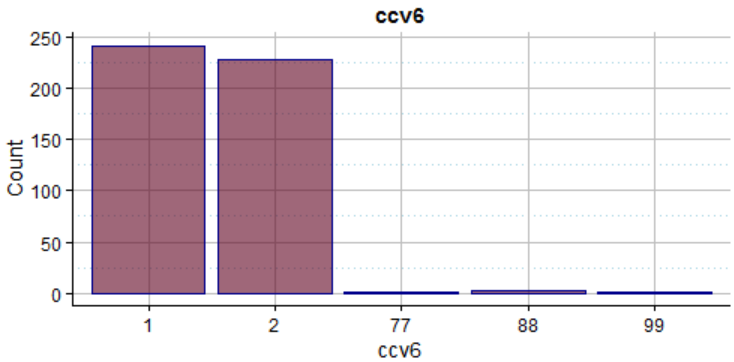
*ccv*4	
Meaning	This feature gives binary information about the presence of the disease Angina pectoris (self-reported by the patient).
Statistics	mode: 2.00; levels: 4.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	1: Yes 2: No

Table A.39: Description of *ccv*6

*ccv*6	
Meaning	This feature gives binary information about the presence of Hypertension (self-reported by the patient).
Statistics	mode: 1.00; levels: 5.00; # missings: 0.00; % missings: 0.00
Distribution	

Discretization & Semantic scales	1: Yes 2: No

Table A.40: Description of *ccv*8

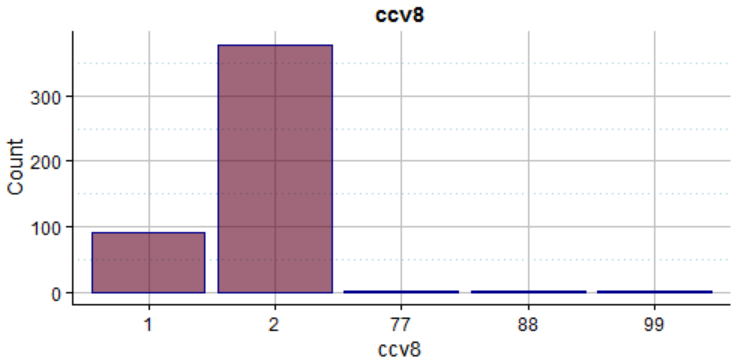
*ccv*8				
Meaning	This feature gives binary information about the presence of the disease Diabetes mellitus (self-reported by the patient).			
Statistics	mode	levels	# missings	% missings
	2.00	5.00	0.00	0.00
Distribution				
Discretization & Semantic scales	1: Yes 2: No			

Table A.41: Description of *tab*1

*tab*1				
Meaning	This feature gives binary information about the tobacco consumption respectively smoking behaviour of the patient. Asked question: "Have you smoked at least 100 cigarettes in your entire life?"			
Statistics	mode	levels	# missings	% missings
	2.00	5.00	2.00	0.00

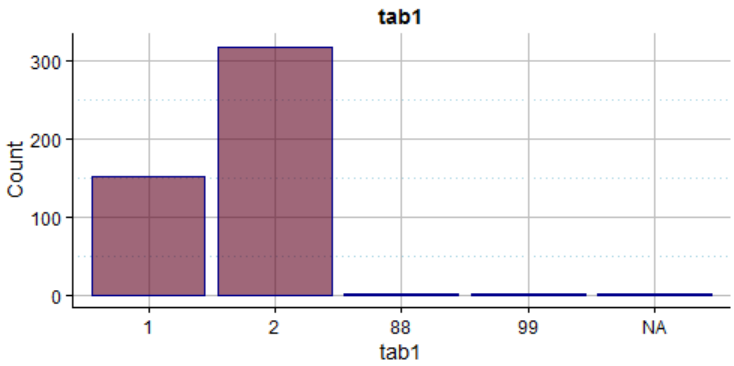
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Unknown 4: NA

Table A.42: Description of *tab1a*

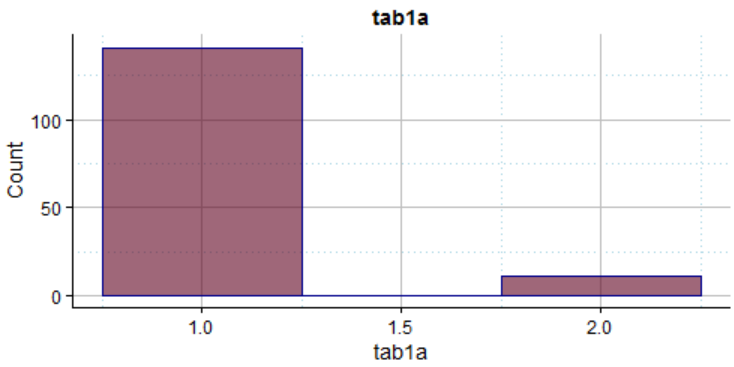
tab1a	
Meaning	This feature gives categorical information about the smoking behaviour of the patient and is the follow up question when *tab1* was answered positively with 1 (Yes). Asked question: "If yes, Did you smoke cigarettes daily, occasionally, or not at all?"
Statistics	mode: 1.00; levels: 3.00; # missings: 322.00; % missings: 0.68
Distribution	
Discretization & Semantic scales	1: Daily 2: Ocassionally 3: Undecided

Table A.43: Description of *tab1a1*

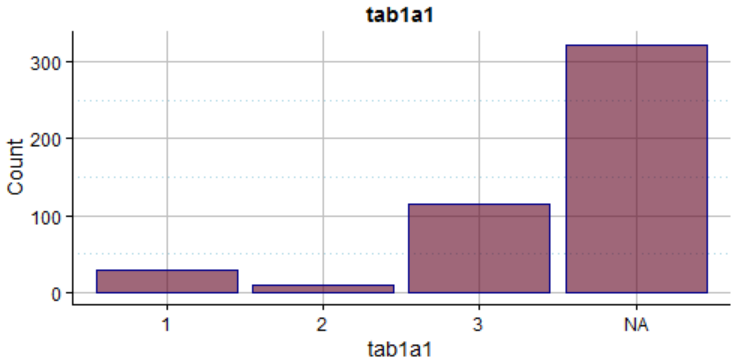
tab1a1	
Meaning	This feature gives categorical information about the current smoking behaviour of the patient. Asked question: "Do you smoke currently?"
Statistics	mode: 3.00; levels: 4.00; # missings: 322.00; % missings: 0.68
Distribution	
Discretization & Semantic scales	1: Yes, daily 2: Yes, occasionally 3: No

Table A.44: Description of *tab1a1a*

tab1a1a	
Meaning	This feature gives categorical information about the number of times the patient has quit smoking. Asked question: "If not, when have you stopped smoking?"
Statistics	mode: 8.00; levels: 8.00; # missings: 356.00; % missings: 0.75

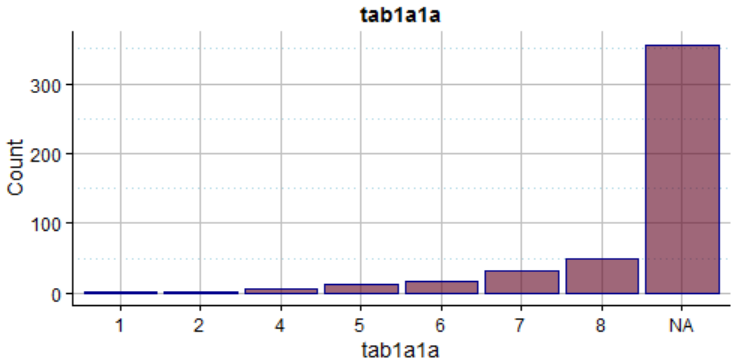
Distribution	
Discretization & Semantic scales	1: Yesterday 2: 2-6 days ago 3: 7-30 days ago 4: 1-12 months ago 5: 1-5 years ago 6: 6-10 years ago 7: 11-20 years ago 8: more than 20 years ago

Table A.45: Description of *tab1a3*

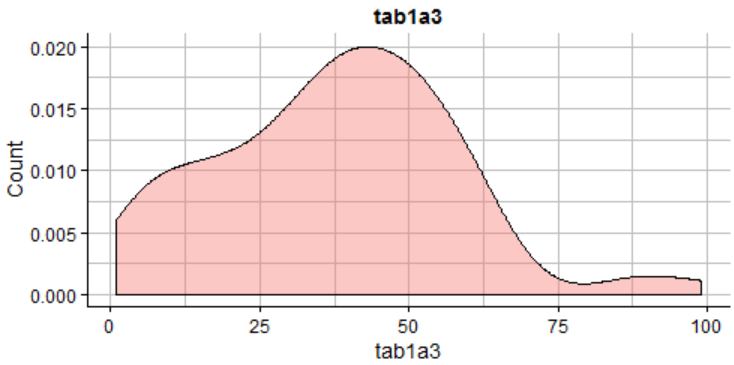
tab1a3	
Meaning	This feature gives numeric information about the time in years the patient has smoked. Asked question: "For how many years did you smoke?"
Statistics	min: 1.00; max: 99.00; average: 37.75; median: 40.00; σ^2: 402.15; σ: 20.05; # missings: 324.00; % missings: 0.68
Distribution	

Table A.46: Description of *alch*1

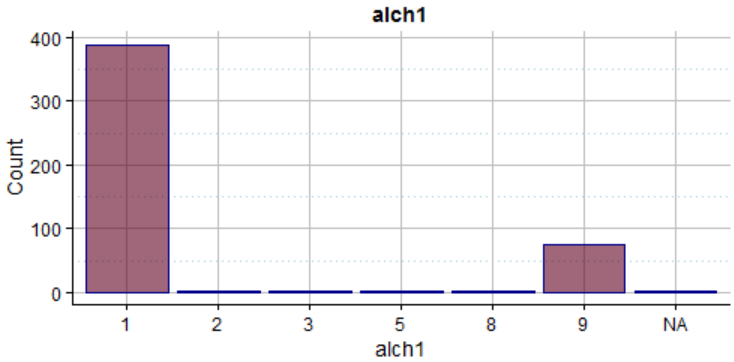
*alch*1	
Meaning	This feature gives categorical information about the alcohol consumption of the patient. Question asked: "How many drinks do you have?".
Statistics	mode: 1.00; levels: 7.00; # missings: 3.00; % missings: 0.01
Distribution	
Discretization & Semantic scales	0: Never (in the last year) 1: One or less per month 2: from 2 to 4 per month 3: Twice per week 4: 3 Times per week 5: 4 Times per week 6: 5 Times per week 7: 6 Times per week 8: Daily

Table A.47: Description of *alch*1*a*1

*alch*1*a*1	
Meaning	This feature gives numerical information about the wine consumption in glasses per day. Question asked: "How many glasses of wine do you drink daily?"

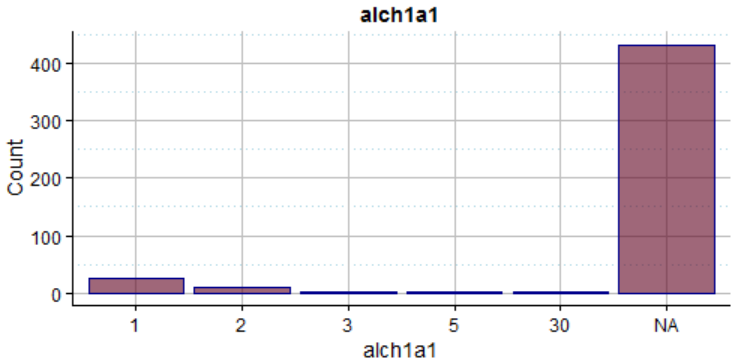
Statistics	mode	levels	# missings	% missings
	1.00	6.00	432.00	0.91
Distribution				

Table A.48: Description of *alch*1*a*2

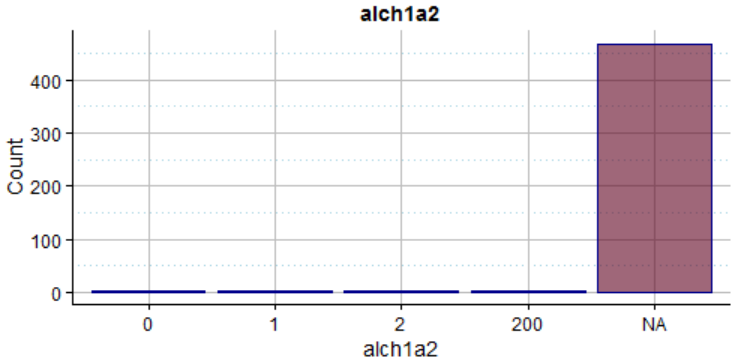
*alch*1*a*2	
Meaning	This feature gives numerical information about the beer consumption in glasses per day. Question asked: "How many glasses of beer do you drink daily?"
Statistics	mode levels # missings % missings 1.00 5.00 468.00 0.99
Distribution	

Table A.49: Description of *alch*1*a*3

*alch*1*a*3	
Meaning	This feature gives numerical information about the consumption of spirits in glasses per day. Question asked: "How many glasses of spirits do you drink daily?"

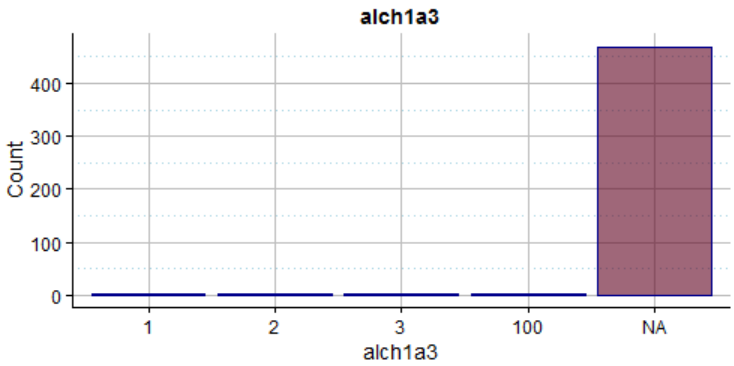
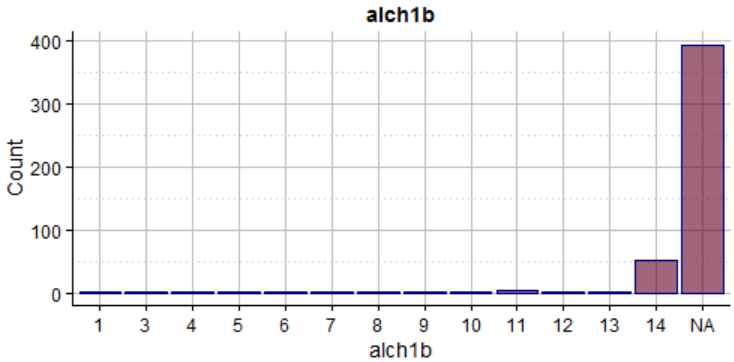
Statistics	mode	levels	# missings	% missings
	3.00	5.00	469.00	0.99
Distribution				

Table A.50: Description of *alch*1*b*

*alch*1*b*	
Meaning	This feature gives numeric information about the years the patients' drinking behaviour is like described in Variable *alch*1 (table A.46). Follow up question asked: "For how many years?".
Statistics	mode: 14.00; levels: 14.00; # missings: 393.00; % missings: 0.83
Distribution	

Discretization & Semantic scales	1: $\leqslant$ 1 year 2: 2 years 3: 3 years 4: 4 years 5: 5 years 6: 6 years 7: 7 years 8: 8 years 9: 9 years 10: 10 years 11: 10-15 years 11: 15-20 years 11: 20-30 years 11: >30 years

Table A.51: Description of *alch*2

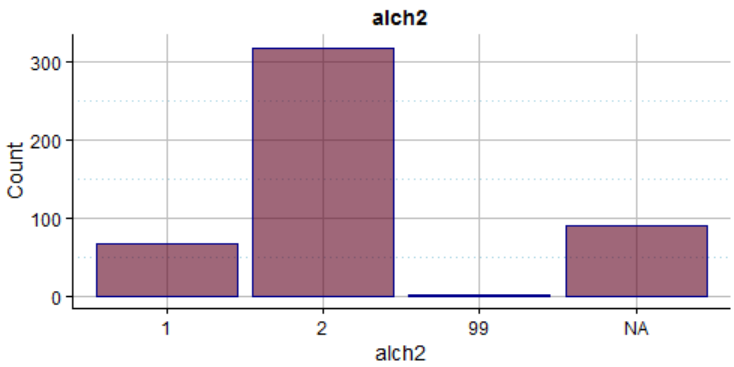
*alch*2	
Meaning	This feature gives binary information about if the patient consumed alcohol previously in life or not. Question asked: "Did you drink previously?".
Statistics	mode: 2.00; levels: 4.00; # missings: 90.00; % missings: 0.19
Distribution	
Discretization & Semantic scales	1: Yes 2: No

Table A.52: Description of *alch2a*

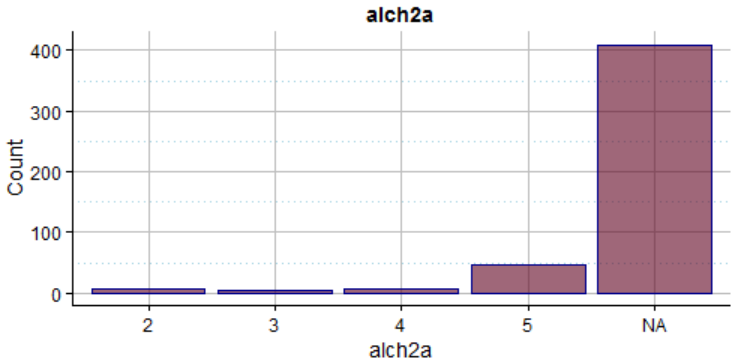
alch2a	
Meaning	This feature gives category information about the kind of drinker the patient is.
Statistics	mode: 5.00; levels: 5.00; # missings: 409.00; % missings: 0.86
Distribution	
Discretization & Semantic scales	1: male, more as 12,female, more as 8 2: M=9-12, W=7-8 3: M=7-8, W=5-6 4: M=3-6, W=3-4 5: M=1-2, W=1-2 (units of alcohol/day)

Table A.53: Description of *alch2b*

alch2b	
Meaning	This feature gives categorical information about the drinking starting age.
Statistics	mode: 2.00; levels: 7.00; # missings: 410.00; % missings: 0.86

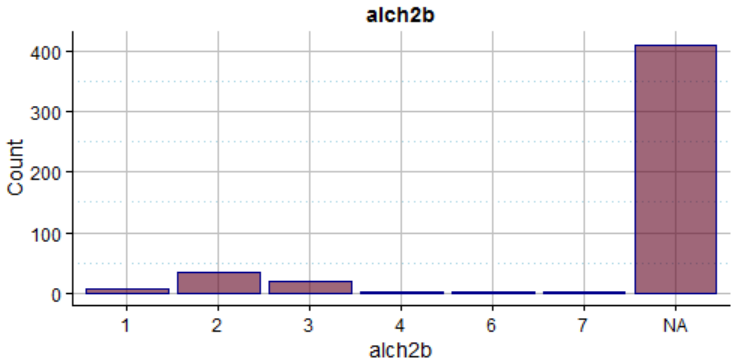
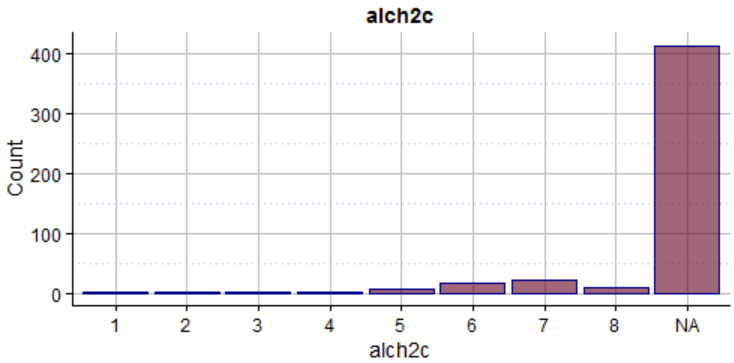
Distribution	
Discretization & Semantic scales	1: <15 2: 15-20 3: 21-30 4: 31-40 5: 41-50 6: 51-60 7: 61-70 8: 71-80 9 >80

Table A.54: Description of *alch2c*

alch2c				
Meaning	This feature gives categorical information about the drinking ending age.			
Statistics	mode	levels	# missings	% missings
	7.00	9.00	412.00	0.87
Distribution				

Discretization & Semantic scales	1: <15 2: 15-20 3: 21-30 4: 31-40 5: 41-50 6: 51-60 7: 61-70 8: 71-80 9 >80

Table A.55: Description of $k1$

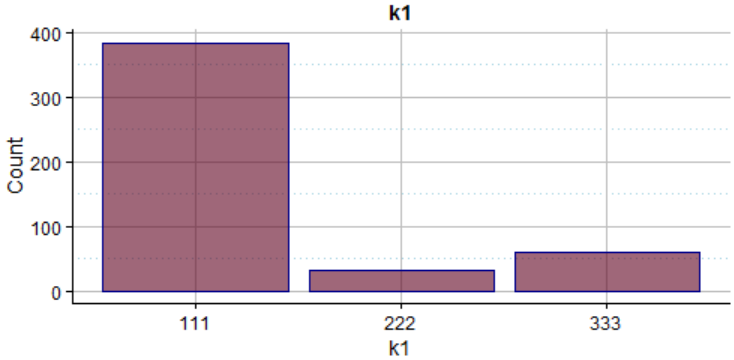
$k1$	
Meaning	This feature gives categorical information about the WHO activity 6: "Any difficulty washing face and arms?". This feature is associated with the ADL test, and represents question 1.
Statistics	mode: 111.00; levels: 3.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (independent, score=1) 333: Unable to do it (dependent, score=0)

Table A.56: Description of $k2$

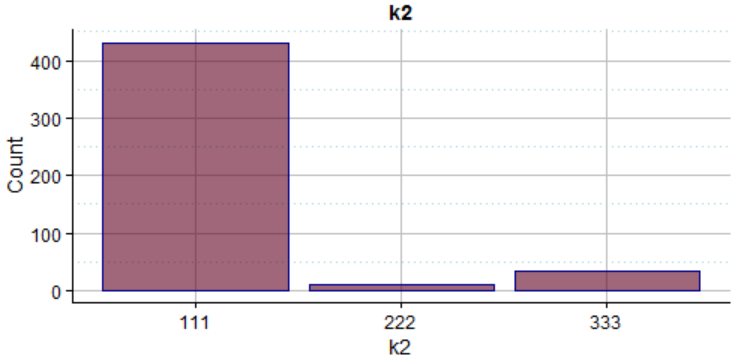
$k2$	
Meaning	This feature gives categorical information about the WHO activity 8: "Any difficulty dressing and undressing?". This feature is associated with the ADL test, and represents question 2.
Statistics	mode: 111.00; levels: 3.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (independent, score=1) 333: Unable to do it (dependent, score=0)

Table A.57: Description of $k3$

$k3$	
Meaning	This feature gives categorical information about the WHO activity 11: "Any difficulty using the toilet?". This feature is associated with the ADL test, and represents question 3.
Statistics	mode: 111.00; levels: 3.00; # missings: 1.00; % missings: 0.21

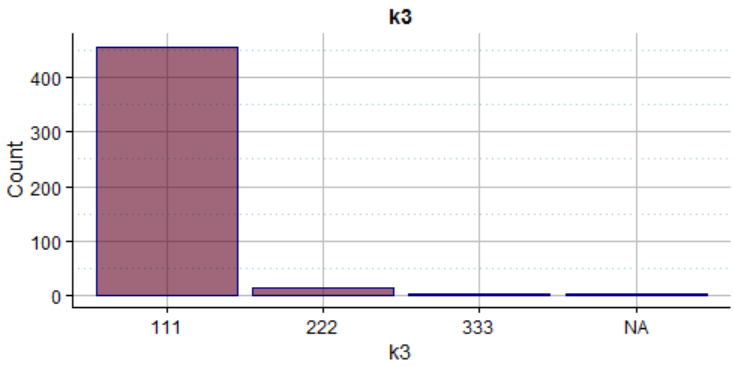
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (dependent, score=0) 333: Unable to do it (dependent, score=0)

Table A.58: Description of $k4$

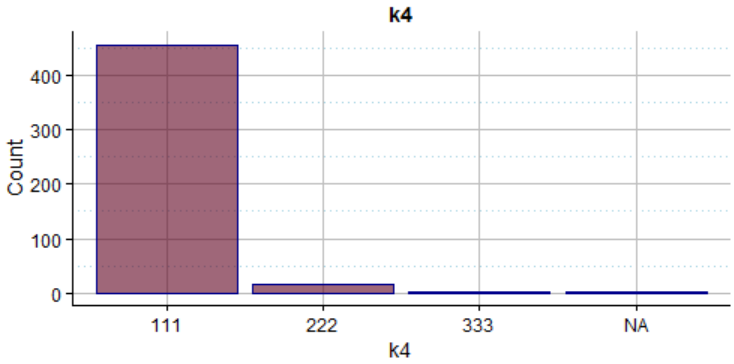
$k4$	
Meaning	This feature gives categorical information about the WHO activity 12: "Any difficulty getting in and out of bed?". This feature is associated with the ADL test, and represents question 4.
Statistics	mode: 111.00; levels: 3.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (dependent, score=0) 333: Unable to do it (dependent, score=0)

Table A.59: Description of $k5$

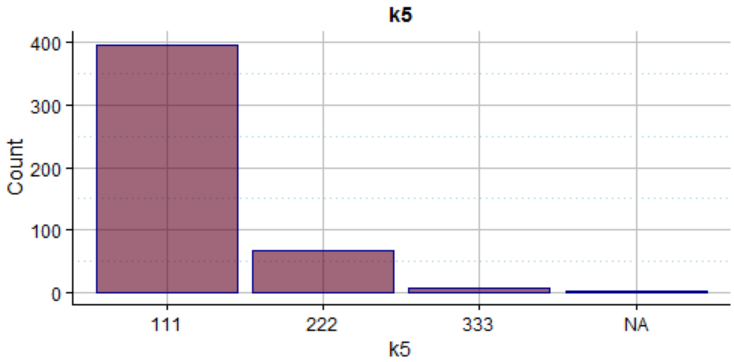
$k5$	
Meaning	This feature gives categorical information about the WHO activity 19: "Any difficulty controlling urination and bowel movements?".This feature is associated with the ADL test, and represents question 5.
Statistics	mode: 111.00; levels: 3.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (dependent, score=0) 333: Unable to do it (dependent, score=0)

Table A.60: Description of $k6$

$k6$	
Meaning	This feature gives categorical information about the WHO activity 9: "Any difficulty eating (e.g.,holding a fork, cutting food, drinking from a glass)?". This feature is associated with the ADL test, and represents question 6.
Statistics	mode: 111.00; levels: 3.00; # missings: 3.00; % missings: 0.63

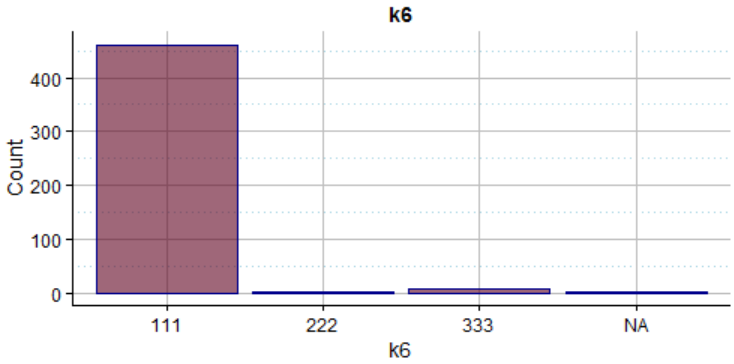
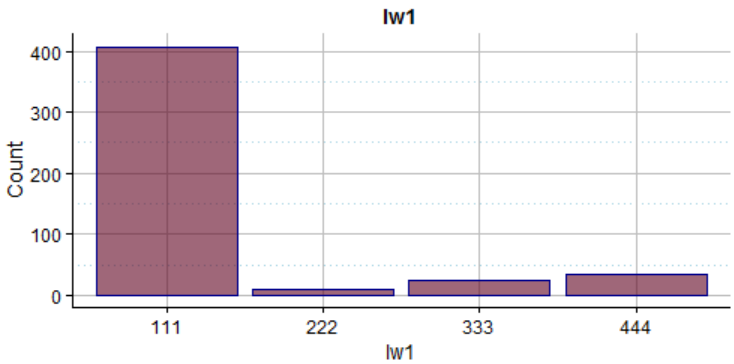
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (independent, score=1) 333: Unable to do it (dependent, score=0)

Table A.61: Description of $lw1$

$lw1$	
Meaning	This feature is associated with the IADL test, and represents question 1.
Statistics	mode: 111.00; levels: 4.00; # missings: 0.00; % missings: 0.00
Distribution	

Discretization & Semantic scales	111: Operates telephone on own initiative(independent, score=1); looks up and dials numbers, etc. 222: Dials a few well-known numbers (independent, score=1) 333: Answers telephone but does not dial (independent, score=1) 444: Does not use telephone at all (dependent, score=0)

Table A.62: Description of $lw2$

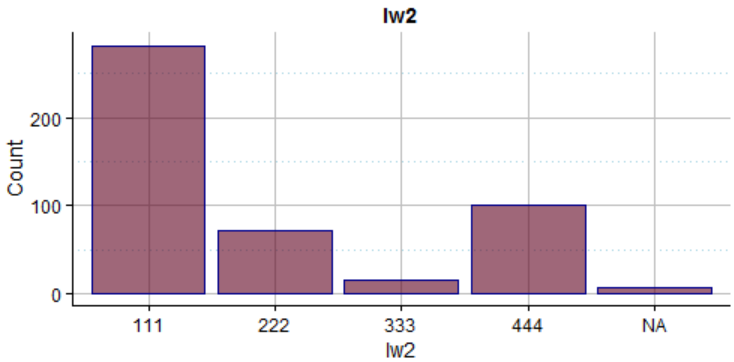
$lw2$	
Meaning	This feature is associated with the IADL test, and represents question 2.
Statistics	mode: 111.00; levels: 4.00; # missings: 6.00; % missings: 1.27
Distribution	
Discretization & Semantic scales	111: Takes care of all shopping needs independently (independent, score=1) 222: Shops independently for small purchases (dependent, score=0) 333: Needs to be accompanied on any shopping trip (dependent, score=0) 444: Completely unable to shop (dependent, score=0)

Table A.63: Description of $lw3$

lw3

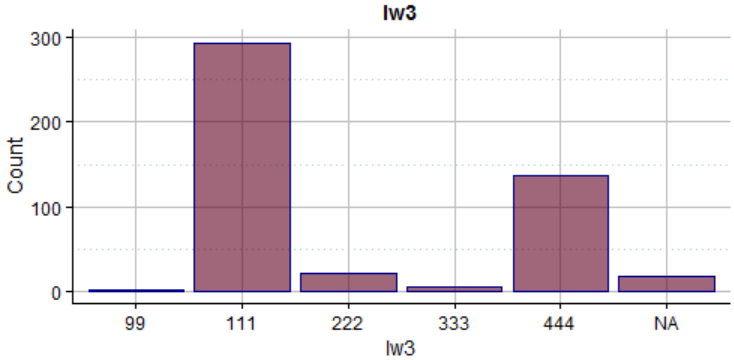
Meaning	This feature is associated with the IADL test, and represents question 3.
Statistics	mode: 111.00; levels: 5.00; # missings: 17.00; % missings: 3.59
Distribution	
Discretization & Semantic scales	111: Plans, prepares, and serves adequate meals independently(independent, score=1) 222: Prepares adequate meals if supplied with ingredients (dependent, score=0) 333: Heats and serves prepared meals, or prepares meals but does not maintain adequate diet (dependent, score=0) 444: Needs to have meals prepared and served (dependent, score=0)

Table A.64: Description of *lw4*

lw4	
Meaning	This feature is associated with the IADL test, and represents question 4.
Statistics	mode: 111.00; levels: 4.00; # missings: 12.00; % missings: 2.53

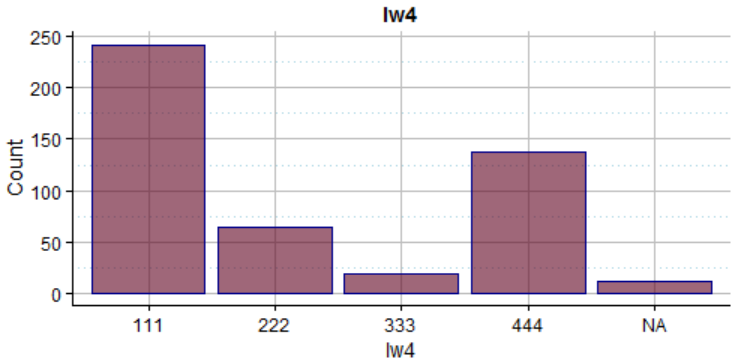
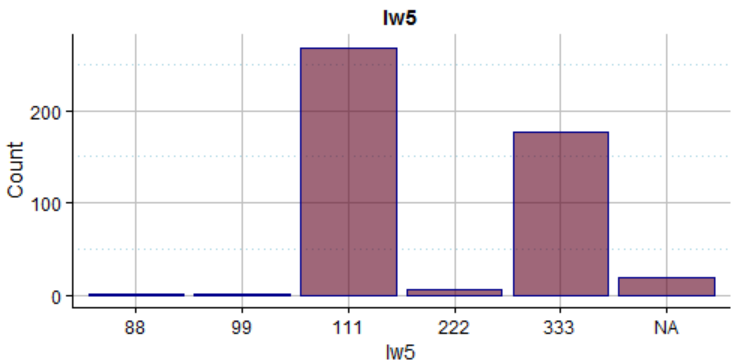
Distribution	
Discretization & Semantic scales	111: Maintains house alone or with occasional assistance (e.g., "heavy work domestic help")(independent, score=1) 222: Performs light daily tasks such as dishwashing, bed making(independent, score=1) 333: Performs light daily tasks but cannot maintain acceptable level of cleanliness (dependent, score=0) 444: Needs help with all home maintenance tasks (dependent, score=0) 555: Does not participate in any housekeeping tasks (dependent, score=0)

Table A.65: Description of $lw5$

$lw5$	
Meaning	This feature is associated with the IADL test, and represents question 5.
Statistics	mode: 111.00; levels: 5.00; # missings: 20.00; % missings: 4.22
Distribution	

Discretization & Semantic scales	111: Does personal laundry completely (independent, score=1) 222: Launders small items; rinses stockings, etc. (dependent, score=0)) 333: All laundry must be done by others (dependent, score=0)

Table A.66: Description of *lw*6

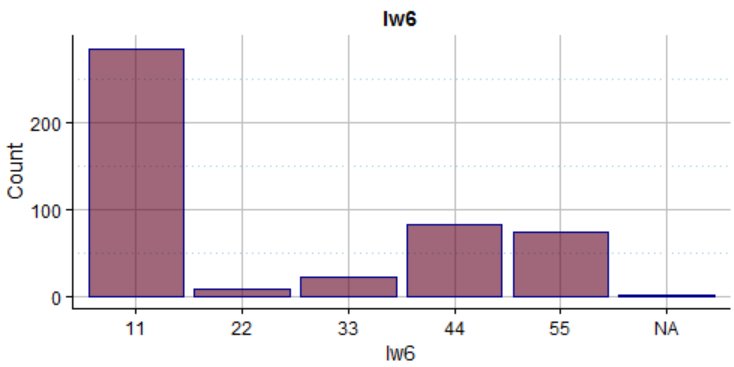
*lw*6	
Meaning	This feature is associated with the IADL test, and represents question 6.
Statistics	mode: 11.00; levels: 5.00; # missings: 2.00; % missings: 0.42
Distribution	
Discretization & Semantic scales	111: Travels independently on public transportation or drives own car(independent, score=1) 222: Arranges own travel via taxi, but does not otherwise use public transportation (independent, score=1) 333: Travels on public transportation when assisted or accompanied by another (dependent, score=0) 444: Travel limited to taxi or automobile with assistance of another (dependent, score=0) 555: Does not travel at all (dependent, score=0)
Note	The values should be "111,222,333,444,555" instead of "11,22,33,44,55".

Table A.67: Description of $lw7$

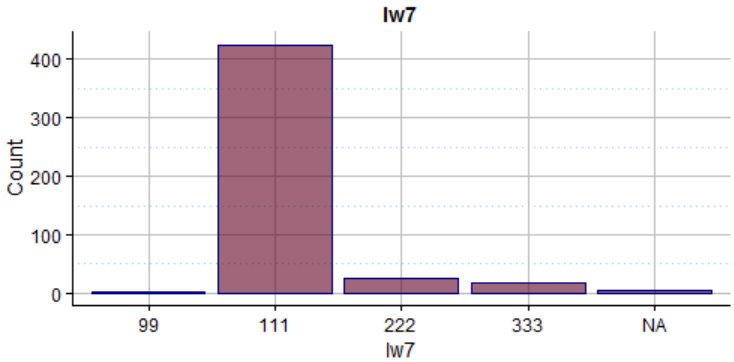
$lw7$	
Meaning	This feature is associated with the IADL test, and represents question 7.
Statistics	mode: 111.00; levels: 4.00; # missings: 4.00; % missings: 0.84
Distribution	
Discretization & Semantic scales	111: Is responsible for taking medication in correct dosages at correct time(independent, score=1) 222: Takes responsibility if medication is prepared in advance in separate dosages (dependent, score=0) 333: Is not capable of dispensing own medication (dependent, score=0)

Table A.68: Description of $lw8$

$lw8$	
Meaning	This feature is associated with the IADL test, and represents question 9.
Statistics	mode: 111.00; levels: 3.00; # missings: 0.00; % missings: 0.00

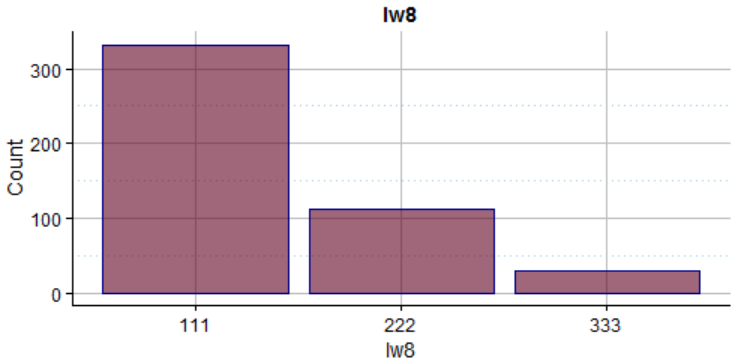
Distribution	
Discretization & Semantic scales	111: Manages financial matters independently (budgets, writes checks, pays rent and bills, goes to bank), collects and keeps track of income (independent, score=1) 222: Manages day-to-day purchases, but needs help with banking, major purchases, etc: (independent, score=1) 333: Incapable of handling money (dependent, score=0)

Table A.69: Description of $ys1$

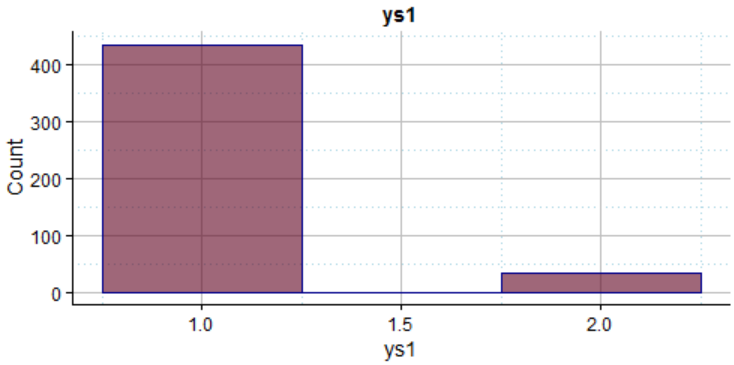
$ys1$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS1: "Are you basically satisfied with your life?".
Statistics	mode: 1.00; levels: 2.00; # missings: 3.00; % missings: 0.63
Distribution	
Semantic scales	1: Yes (score 0) 2: No (score 1)

Table A.70: Description of $ys2$

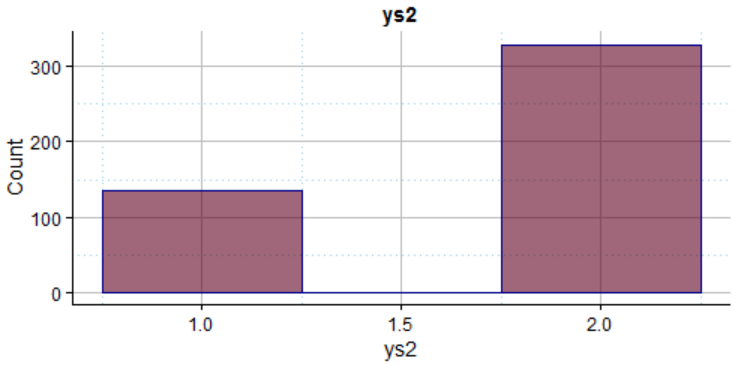
$ys2$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS2: "Have you dropped many of your activities and interests?".
Statistics	mode: 2.00; levels: 2.00; # missings: 9.00; % missings: 1.90
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.71: Description of $ys3$

$ys3$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS3: "Do you feel that your life is empty?".
Statistics	mode: 2.00; levels: 2.00; # missings: 9.00; % missings: 1.90

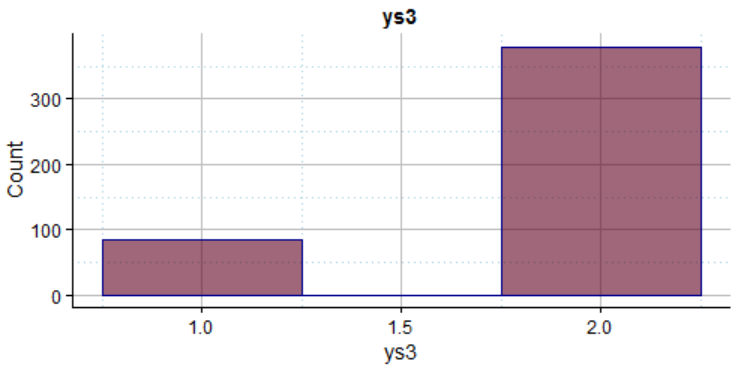
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.72: Description of $ys4$

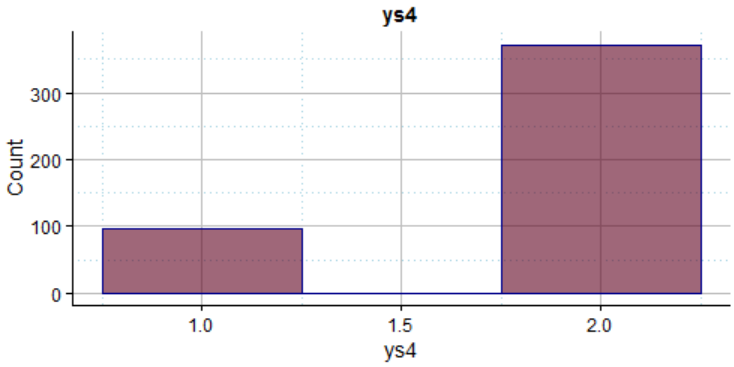
$ys4$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS4: "Do you often get bored?".
Statistics	mode: 2.00; levels: 2.00; # missings: 6.00; % missings: 1.27
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.73: Description of $ys5$

$ys5$

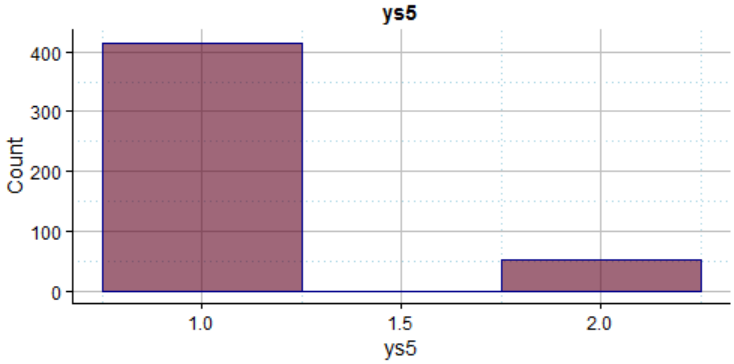
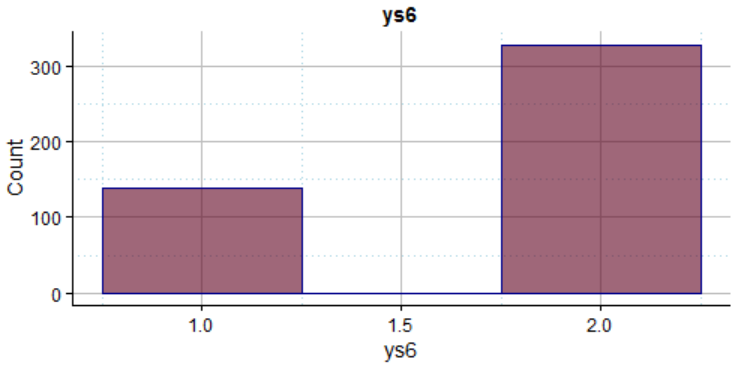
Meaning	This feature gives binary information about the the geriatric depression scale question GDS5: "Are you in good spirits most of the time?".
Statistics	mode: 1.00; levels: 2.00; # missings: 6.00; % missings: 1.27
Distribution	
Discretization & Semantic scales	1: Yes (score 0) 2: No (score 1)

Table A.74: Description of $ys6$

$ys6$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS6: "Are you afraid that something bad is going to happen to you?".
Statistics	mode: 2.00; levels: 2.00; # missings: 8.00; % missings: 1.69
Distribution	

Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.75: Description of $ys7$

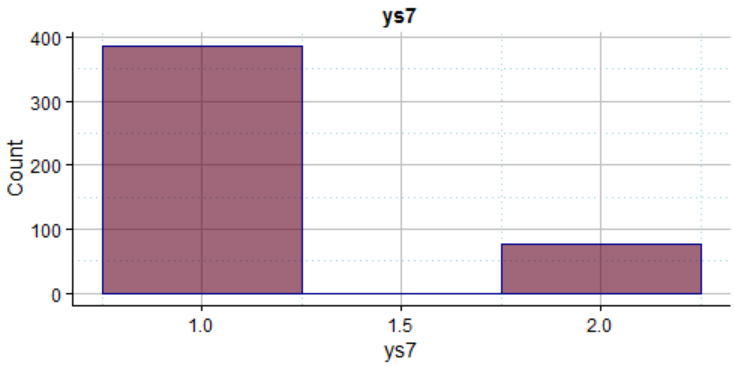
$ys7$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS7: "Do you feel happy most of the time?".
Statistics	mode: 1.00; levels: 2.00; # missings: 9.00; % missings: 1.90
Distribution	
Discretization & Semantic scales	1: Yes (score 0) 2: No (score 1)

Table A.76: Description of $ys8$

$ys8$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS8: "Do you often feel helpless?".
Statistics	mode: 2.00; levels: 2.00; # missings: 6.00; % missings: 1.27

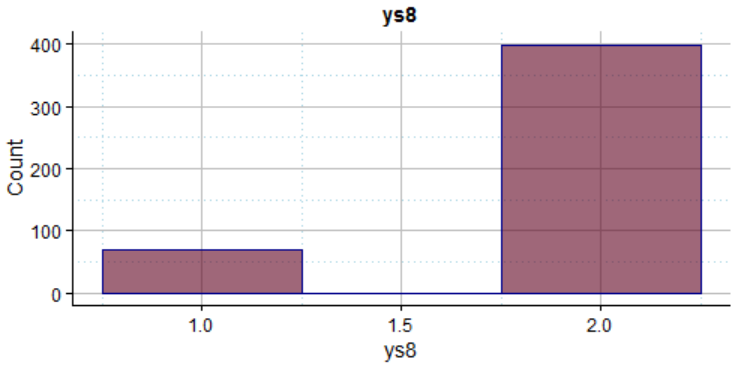
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.77: Description of $ys9$

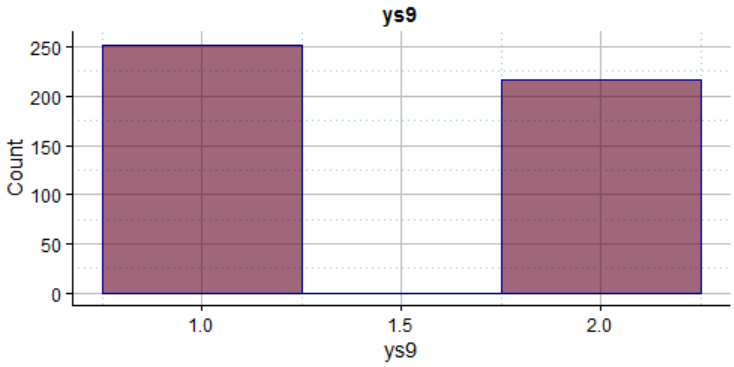
$ys9$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS9: "Do you prefer to stay at home, rather than going out and doing new things?".
Statistics	mode: 1.00; levels: 2.00; # missings: 5.00; % missings: 1.05
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.78: Description of $ys10$

$ys10$

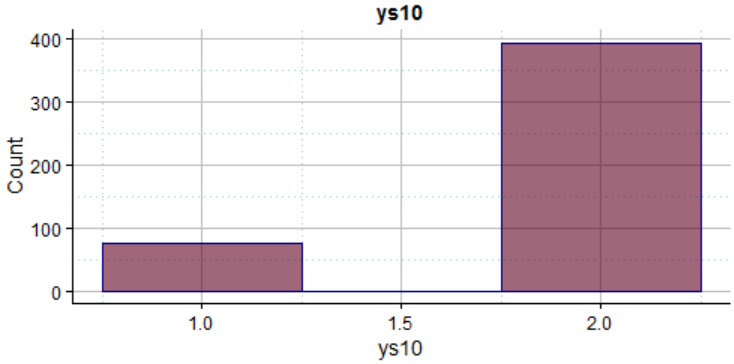
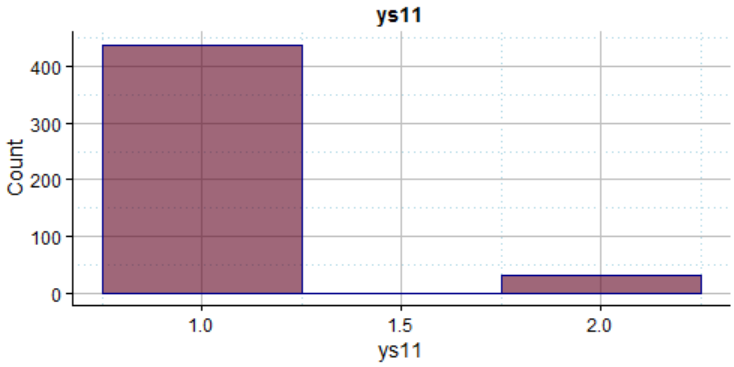
Meaning	This feature gives binary information about the the geriatric depression scale question GDS10: "Do you feel you have more problems with memory than most?".
Statistics	mode: 2.00; levels: 2.00; # missings: 5.00; % missings: 1.05
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.79: Description of $ys11$

$ys11$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS11: "Do you think it is wonderful to be alive now?".
Statistics	mode: 1.00; levels: 2.00; # missings: 5.00; % missings: 1.05
Distribution	

Discretization & Semantic scales	1: Yes (score 0) 2: No (score 1)

Table A.80: Description of $ys12$

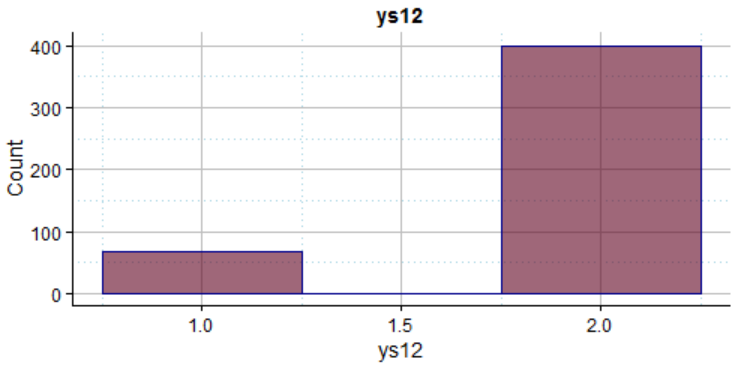
$ys12$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS12: "Do you feel pretty worthless the way you are now?".
Statistics	mode: 2.00; levels: 2.00; # missings: 7.00; % missings: 1.48
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.81: Description of $ys13$

$ys13$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS13: "Do you feel full of energy?".
Statistics	mode: 2.00; levels: 2.00; # missings: 4.00; % missings: 0.84

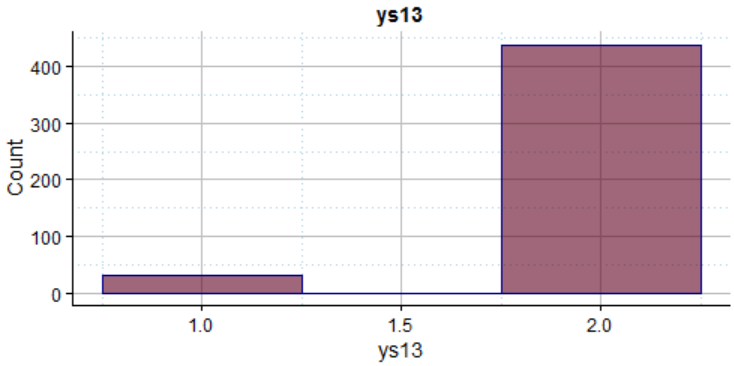
Distribution	
Discretization & Semantic scales	1: Yes (score 0) 2: No (score 1)

Table A.82: Description of $ys14$

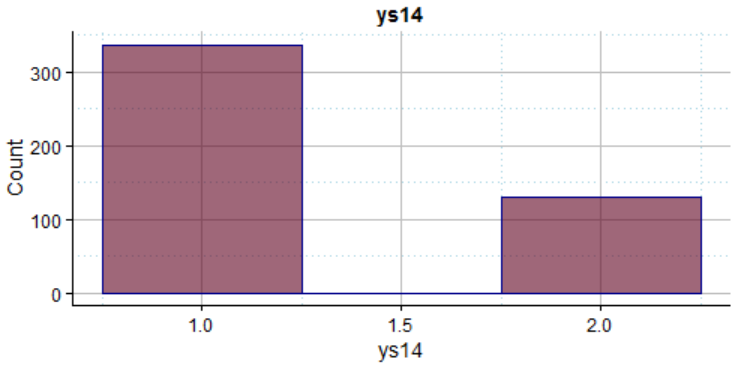
$ys14$	
Meaning	This feature gives binary information about the the geriatric depression scale question GDS14: "Do you feel that your situation is hopeless?".
Statistics	mode: 1.00; levels: 2.00; # missings: 7.00; % missings: 1.48
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.83: Description of $ys15$

$ys15$

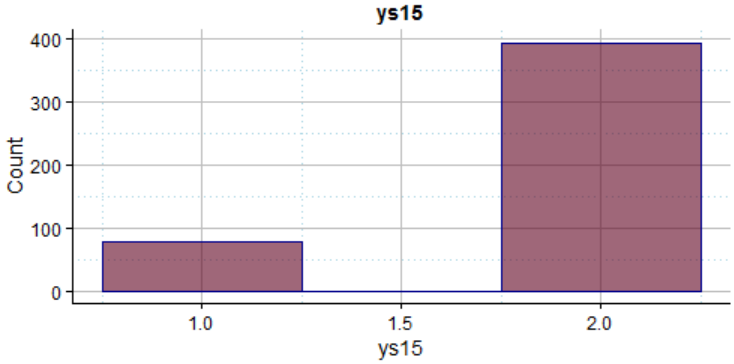
Meaning	This feature gives binary information about the the geriatric depression scale question GDS15: "Do you think that most people are better off than you are?".
Statistics	mode: 2.00; levels: 2.00; # missings: 4.00; % missings: 0.84
Distribution	
Discretization & Semantic scales	1: Yes (score 1) 2: No (score 0)

Table A.84: Description of *altura*1

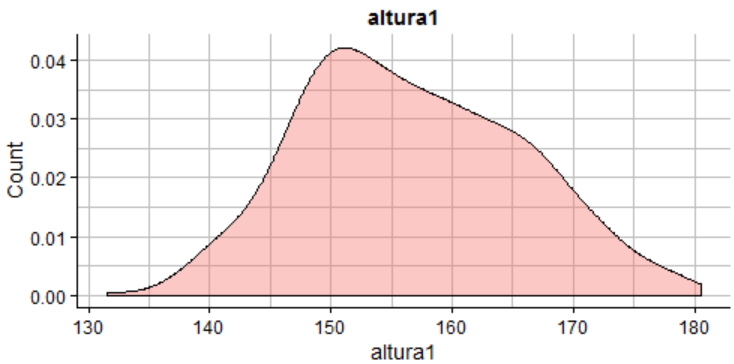
*altura*1	
Meaning	This feature gives numeric information about the height of the patient in centimeters.
Statistics	min: 132.00; max: 180.00; average: 156.67; median: 156.00; σ^2: 82.41; σ: 9.08; # missings: 0.00; % missings: 0.00
Distribution	

Table A.85: Description of *peso*1

peso1

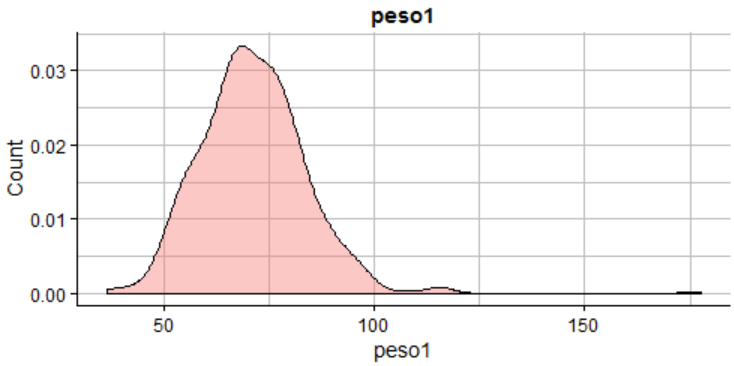
Meaning	This feature gives numeric information about the weight of the patient in kilograms.
Statistics	min: 37.00; max: 177.00; average: 71.35; median: 70.50; σ^2: 167.11; σ: 12.93; # missings: 0.00; % missings: 0.00
Distribution	

Table A.86: Description of *ppca*

ppca

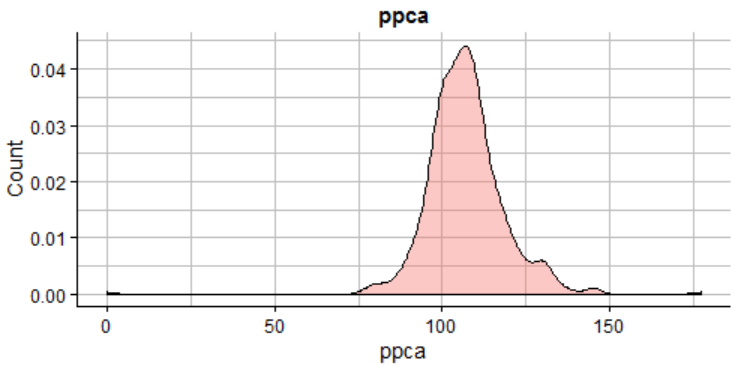
Meaning	This feature gives numeric information about the hip perimeter of the patient in centimeters.
Statistics	min: 0.00; max: 177.00; average: 106.96; median: 106.00; σ^2: 142.88; σ: 11.95; # missings: 1.00; % missings: 0.21
Distribution	

Table A.87: Description of *ppci*

ppci

Meaning	This feature gives numeric information about the waist perimeter of the patient in centimeters.
Statistics	

min	max	average	median	σ^2	σ	# missings	% missings
0.00	130.00	99.41	100.00	144.94	12.04	1.00	0.21

Distribution

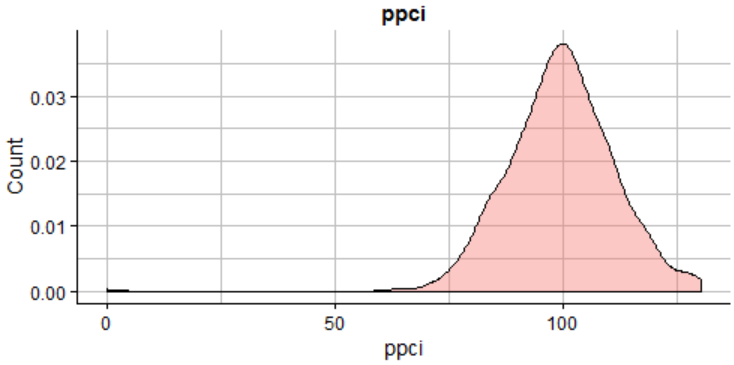


Table A.88: Description of *ekg*1

*ekg*1	
Meaning	This feature gives numeric information about the heart rate in beats per minute.
Statistics	

mode	levels	# missings	% missings
2.00	2.00	466.00	98.31

Distribution

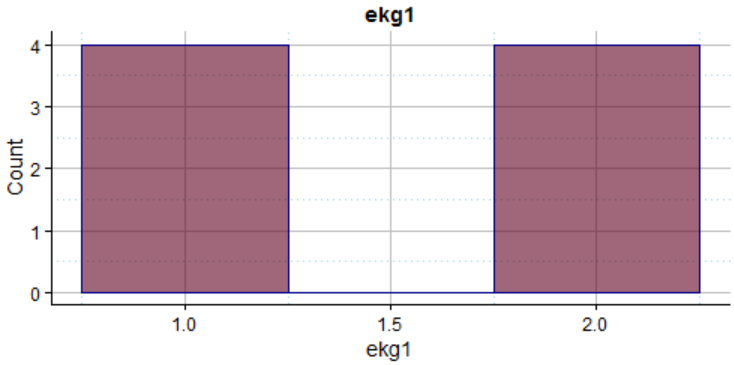


Table A.89: Description of *silla*

silla

Meaning	This feature gives numeric information about the number of times the patient is able to stand up from the chair in a time of 30 seconds.
Statistics	

min	max	average	median	σ^2	σ	# missings	% missings
0.00	24.00	9.28	10.00	27.27	5.22	15.00	3.16

Distribution

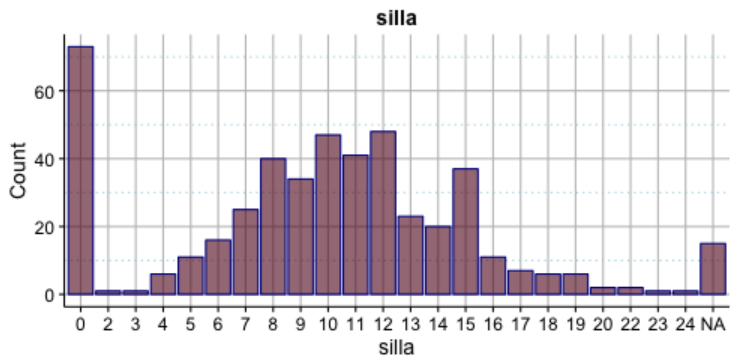


Table A.90: Description of *marcha*

marcha	
Meaning	This feature gives numeric information about the time in seconds it takes for the patient to walk 3 meter.
Statistics	

min	max	average	median	σ^2	σ	# missings	% missings
2.00	70.00	7.22	6.00	29.67	5.45	0.00	0.00

Distribution

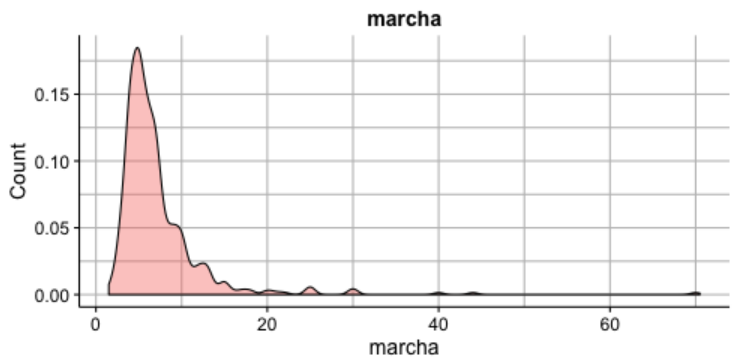


Table A.91: Description of $fuerza1a$

$fuerza1a$

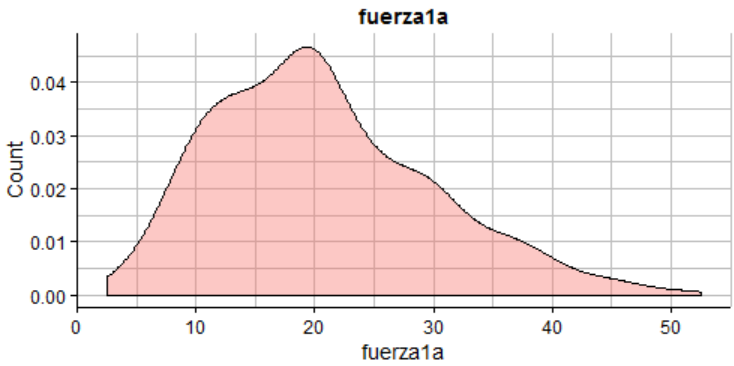
Meaning	This feature gives numeric information about the upper muscle strength measured with a dynamometer (hand grip of the dominant limb in kilograms).
Statistics	mode: 20.00; levels: 46.00; # missings: 0.00; % missings: 0.00
Distribution	

Table A.92: Description of *p1leu*

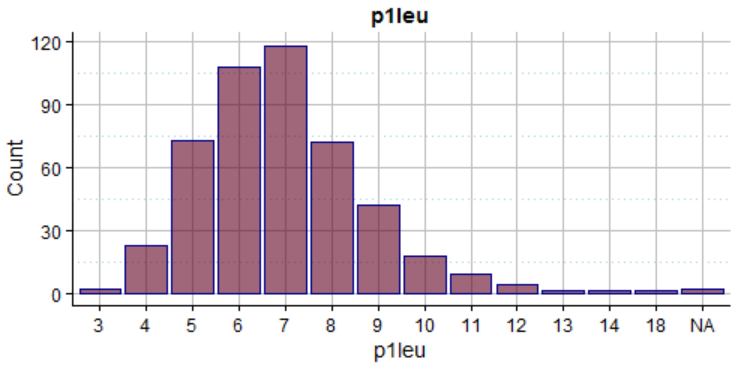
p1leu	
Meaning	This feature gives numeric information about the leukocyte count.
Statistics	min: 3.00; max: 18.00; average: 6.91; median: 7.00; σ^2: 3.16; σ: 1.78; # missings: 2.00; % missings: 0.42
Distribution	

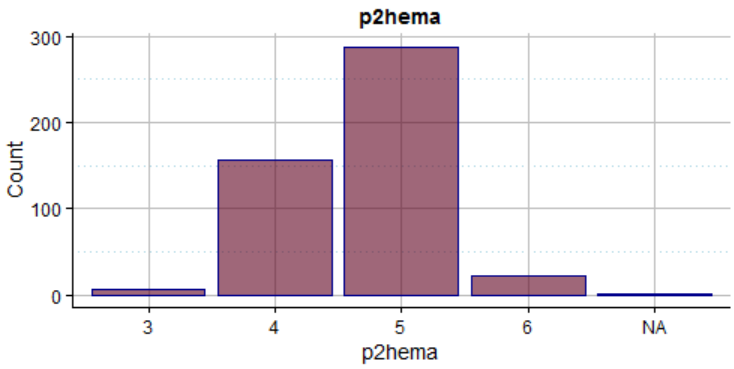
Table A.93: Description of *p2hema*

p2hema

Meaning	This feature gives numeric information about the erythrocyte count.
Statistics	

mode	levels	# missings	% missings
5.00	5.00	2.00	0.42

Distribution



Note	Due to the fact that the feature has only 5 levels, it is visualised and measured like a categorical feature.

Table A.94: Description of *p3hgb*

p3hgb

Meaning	This feature gives numeric information about the hemoglobin count.
Statistics	

min	max	average	median	σ^2	σ	# missings	% missings
8.00	18.00	14.11	14.00	2.60	1.61	2.00	0.42

Distribution

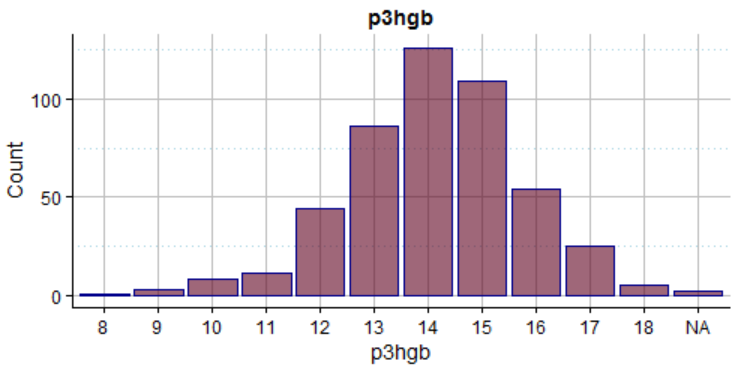


Table A.95: Description of *p4hct*

p4hct

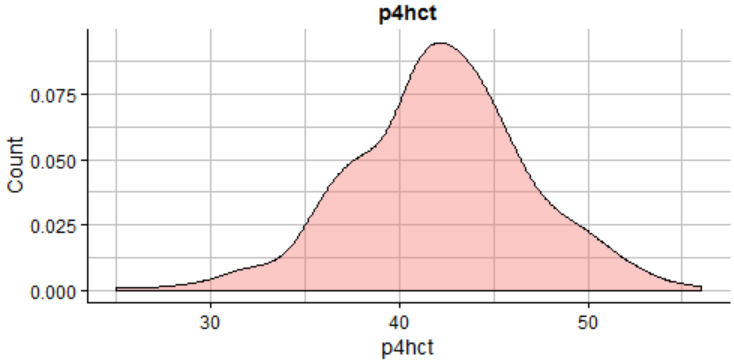
Meaning	This feature gives numeric information about the hematocrit.
Statistics	min: 25.00; max: 56.00; average: 42.18; median: 42.00; σ^2: 22.13; σ: 4.70; # missings: 2.00; % missings: 0.42
Distribution	

Table A.96: Description of *p5vcm*

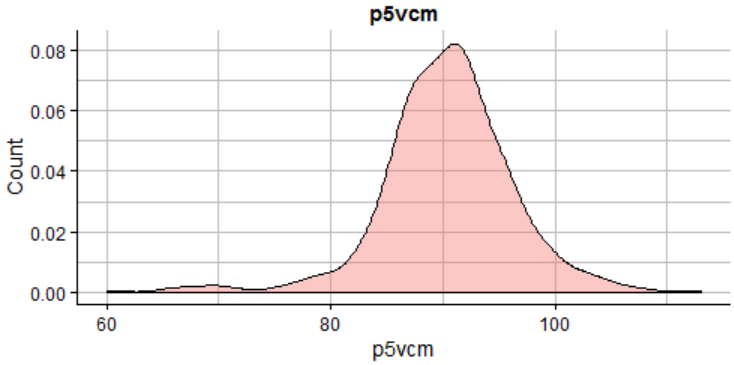
p5vcm	
Meaning	This feature gives numeric information about the Mean Corpuscular Volume (MCV) in fL.
Statistics	min: 60.00; max: 113.00; average: 90.42; median: 91.00; σ^2: 35.47; σ: 5.96; # missings: 2.00; % missings: 0.42
Distribution	

Table A.97: Description of *p6hcm*

p6hcm	
Meaning	This feature gives numeric information about the Mean Corpuscular Haemoglobin (MCH) in pg.

Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	19.00	39.00	30.29	31.00	4.63	2.15	2.00	0.42

Distribution

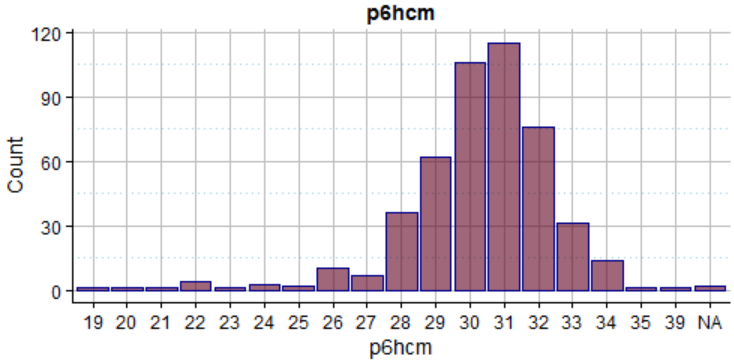


Table A.98: Description of *p7chcm*

p7chcm								
Meaning	This feature gives numeric information about the Mean Corpuscular Haemoglobin Concentration (CHCM) in g/dL.							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	29.00	37.00	33.44	34.00	1.33	1.15	2.00	0.42

Distribution

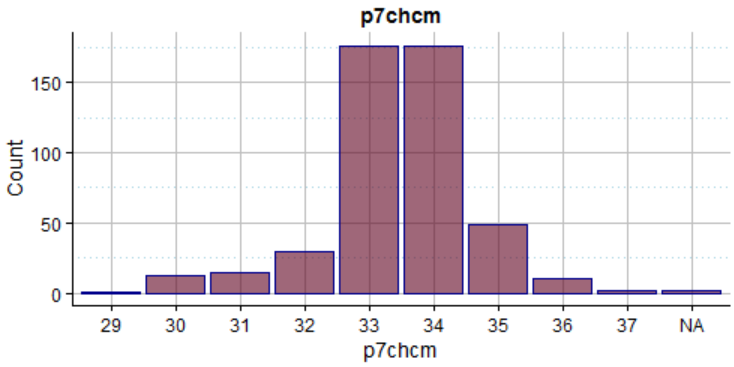


Table A.99: Description of *p8ade*

p8ade								
Meaning	This feature gives numeric information about the Red Cell Distribution Width (RDW) in percent.							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	12.00	27.00	14.09	14.00	2.69	1.64	2.00	0.42

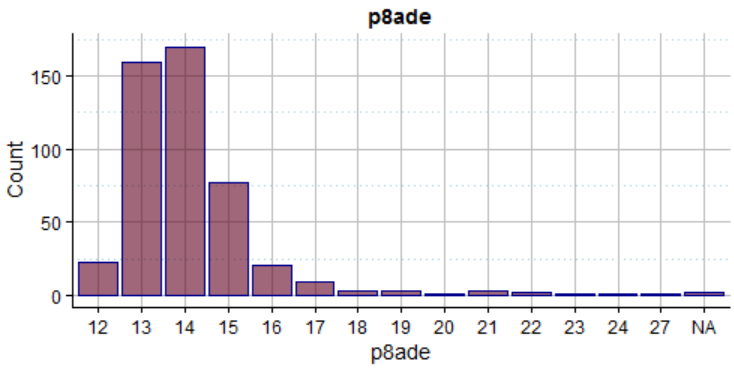
Distribution	

Table A.100: Description of *p9lin*

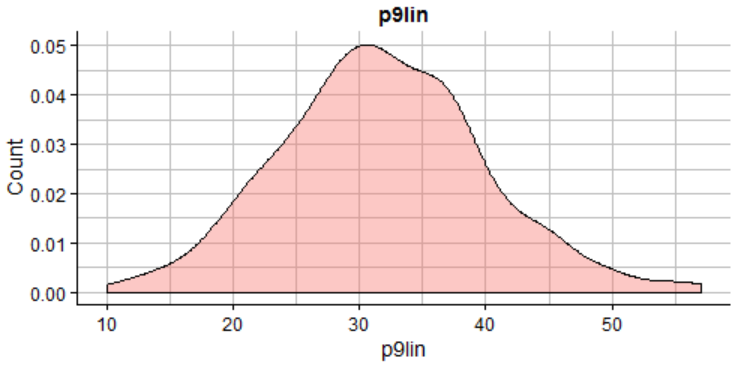
p9lin								
Meaning	This feature gives numeric information about the lymphocyte count in $x10^9$/L.							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	10.00	57.00	31.81	32.00	66.42	8.15	3.00	0.63
Distribution								

Table A.101: Description of *p10mono*

p10mono								
Meaning	This feature gives numeric information about the monocyte count in $x10^9$/L.							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	4.00	18.00	8.01	8.00	3.76	1.94	3.00	0.63

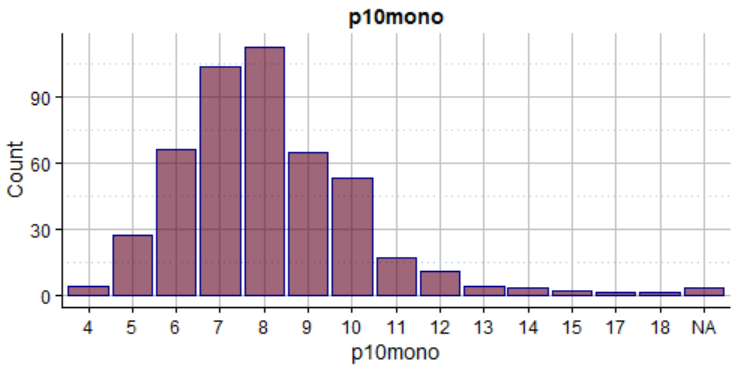
Distribution	

Table A.102: Description of *p13eos*

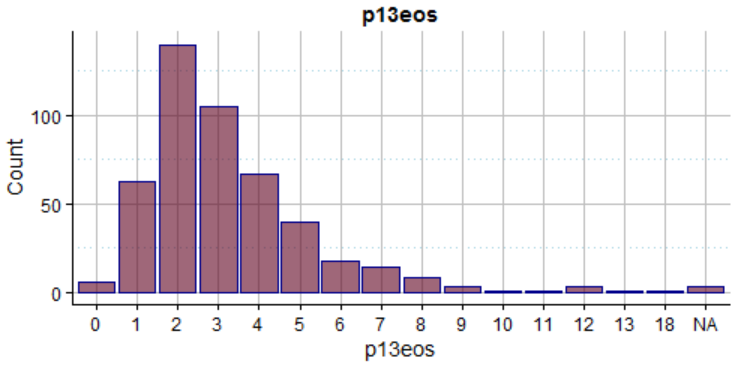
p13eos	
Meaning	This feature gives numeric information about the eosinophiles count in $x10^9$/L.
Statistics	min: 0.00; max: 18.00; average: 3.21; median: 3.00; σ^2: 4.31; σ: 2.08; # missings: 3.00; % missings: 0.63
Distribution	

Table A.103: Description of *p14baso*

p14baso	
Meaning	This feature gives numeric information about the basophiles count in $x10^9$/L.
Statistics	mode: 1.00; levels: 4.00; # missings: 3.00; % missings: 0.01

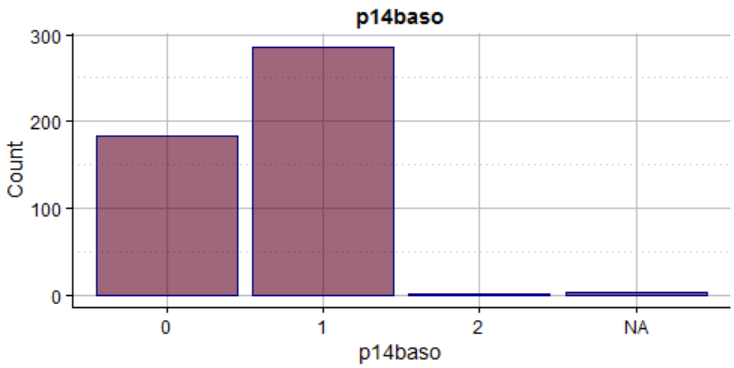
Distribution	
Note	Due to the fact that the feature has only 4 levels, it is visualised and measured like a categorical feature.

Table A.104: Description of $p15dd$

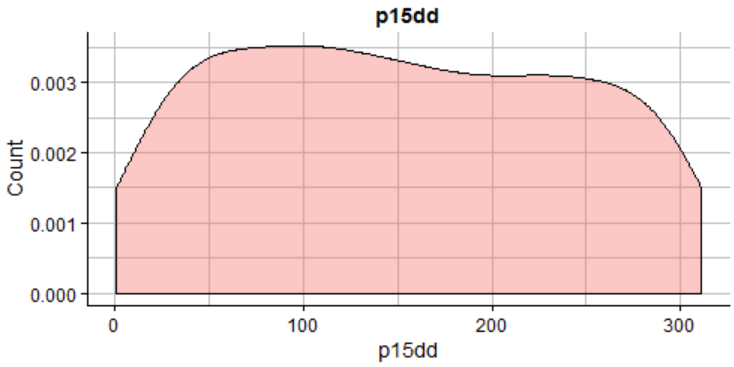
$p15dd$	
Meaning	This feature gives information about the D-Dimer concentration in ğ/L.
Statistics	min: 1.00; max: 311.00; average: 153.04; median: 150.50; σ^2: 7783.92; σ: 88.23; # missings: 84.00; % missings: 17.72
Distribution	

Table A.105: Description of $p16plaq$

$p16plaq$	
Meaning	This feature gives numeric information about the Platelets in $x10^9$/L.
Statistics	min: 98.00; max: 468.00; average: 233.38; median: 226.00; σ^2: 3713.82; σ: 60.94; # missings: 2.00; % missings: 0.42

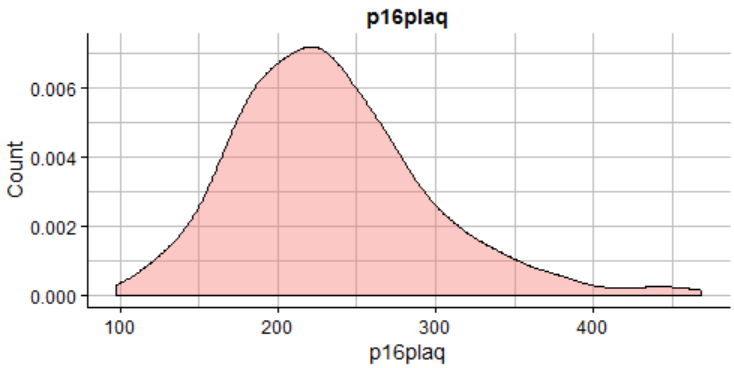
Distribution	

Table A.106: Description of *p17vpm*

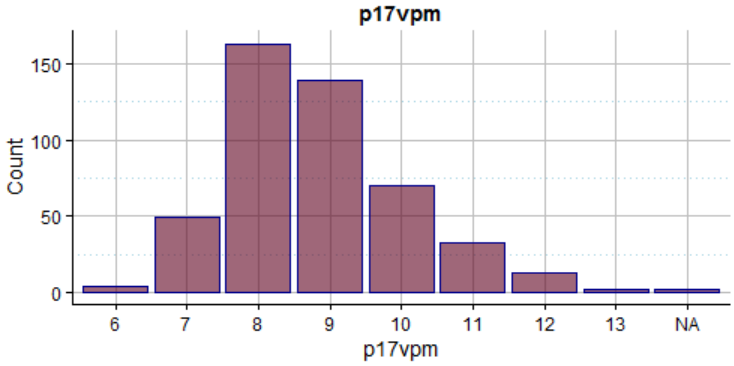
p17vpm	
Meaning	This feature gives numeric information about the Mean Platelet Volume (MPV) in fL.
Statistics	mode: 8.00; levels: 8.00; # missings: 2.00; % missings: 0.42
Distribution	
Note	Due to the fact that the feature has only 8 levels, it is visualised and measured like a categorical feature.

Table A.107: Description of *p23glu*

p23glu	
Meaning	This feature gives numeric information about the blood glucose in mg/dL.
Statistics	min: 72.00; max: 268.00; average: 106.72; median: 100.00; σ^2: 685.44; σ: 26.18; # missings: 0.00; % missings: 0.00

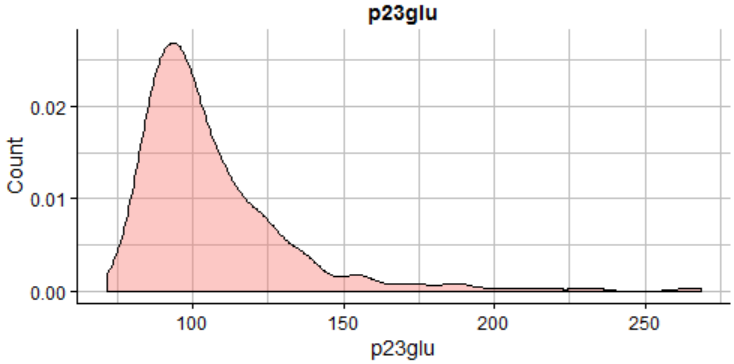
Distribution	

Table A.108: Description of *p24urea*

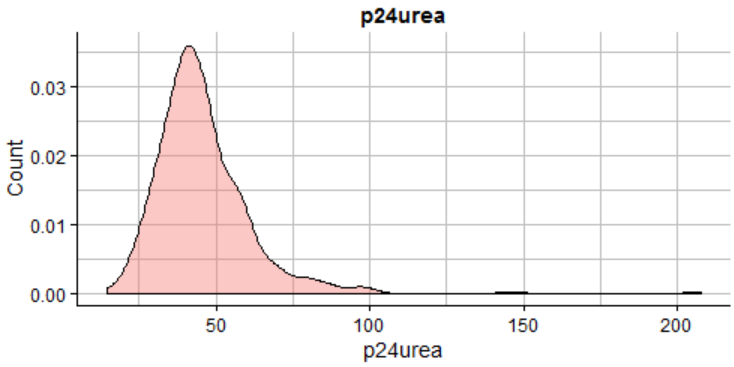
p24urea								
Meaning	This feature gives numeric information about the urea in mg/dL.							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	15.00	207.00	45.74	43.00	262.99	16.22	0.00	0.00
Distribution								

Table A.109: Description of *p25acur*

p25acur								
Meaning	This feature gives numeric information about uric acid [mg/dL].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	1.00	12.00	5.27	5.00	2.24	1.50	1.00	0.21

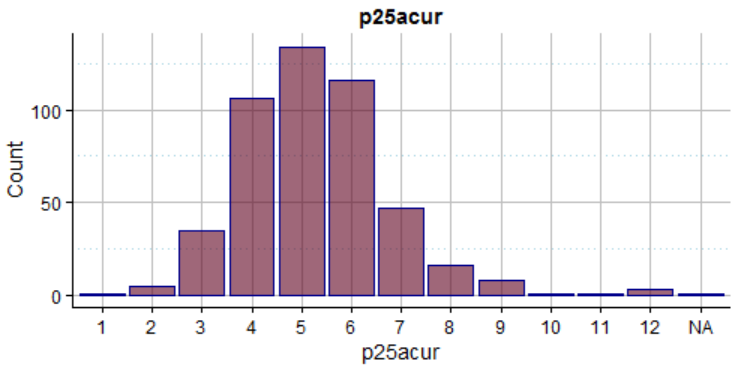
Distribution	

Table A.110: Description of *p26crea*

p26crea	
Meaning	This feature gives numeric information about creatinine [mg/dL].
Statistics	mode: 1.00; levels: 5.00; # missings: 0.00; % missings: 0.00
Distribution	
Note	Due to the fact that the feature has only 5 levels, it is visualised and measured like a categorical feature.

Table A.111: Description of *p27prot*

p27prot	
Meaning	This feature gives numeric information about blood protein [g/dL].
Statistics	mode: 7.00; levels: 4.00; # missings: 0.00; % missings: 0.00

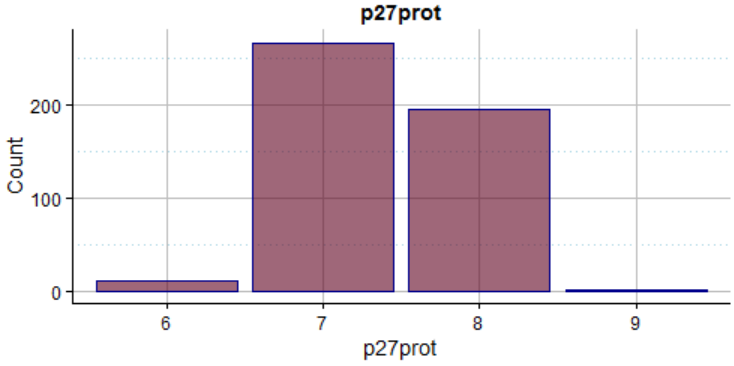
Distribution	
Note	Due to the fact that the feature has only 5 levels, it is visualised and measured like a categorical feature.

Table A.112: Description of *p28albu*

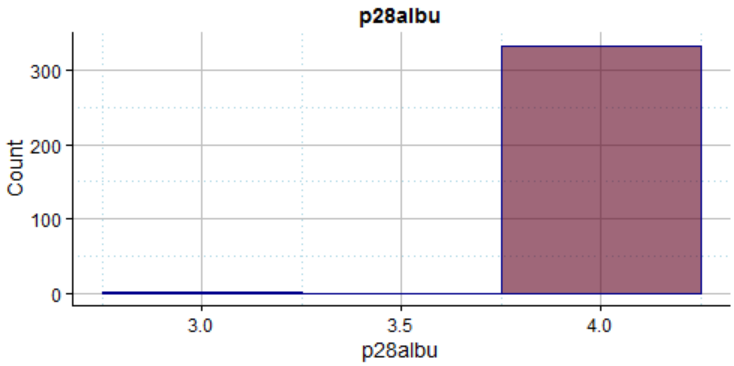
p28albu				
Meaning	This feature gives numeric information about albumin [g/dL].			
Statistics	mode	levels	# missings	% missings
	4.00	2.00	140.00	29.54
Distribution				
Note	Due to the fact that the feature has only 2 levels, it is visualised and measured like a categorical feature.			

Table A.113: Description of *p30chol*

p30chol	
Meaning	This feature gives numeric information about cholesterin [mg/dL].

Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	89.00	320.00	191.96	192.00	1372.25	37.04	0.00	0.00

Distribution

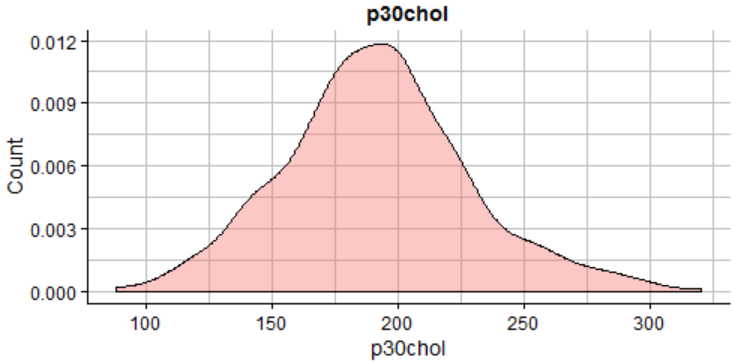


Table A.114: Description of *p31trig*

p31trig	
Meaning	This feature gives numeric information about triglycerides [mg/dL].

Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	41.00	1207.00	116.78	98.50	6194.42	78.70	2.00	0.42

Distribution

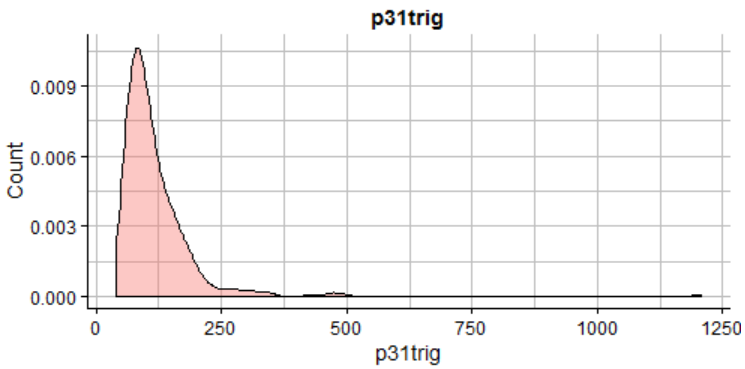


Table A.115: Description of *p32ca*

p32ca	
Meaning	This feature gives numeric information about calcium (Ca) [mg/dL]

Statistics	mode	levels	# missings	% missings
	9.00	5.00	0.00	0.00

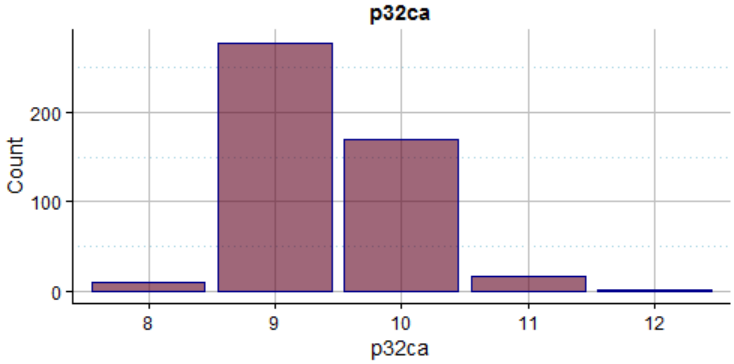
Distribution	
Note	Due to the fact that the feature has only 5 levels, it is visualised and measured like a categorical feature.

Table A.116: Description of *p33p*

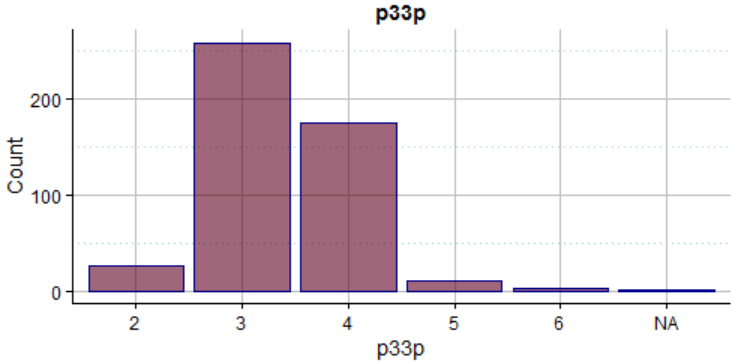
p33p	
Meaning	This feature gives numeric information about phosphorus (P) [mg/dL].
Statistics	mode: 3.00; levels: 5.00; # missings: 2.00; % missings: 0.42
Distribution	

Table A.117: Description of *p34na*

p34na	
Meaning	This feature gives numeric information about sodium (Na) [mEq/L].

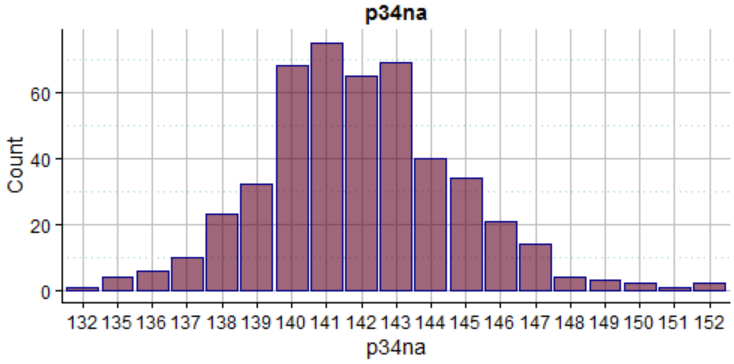
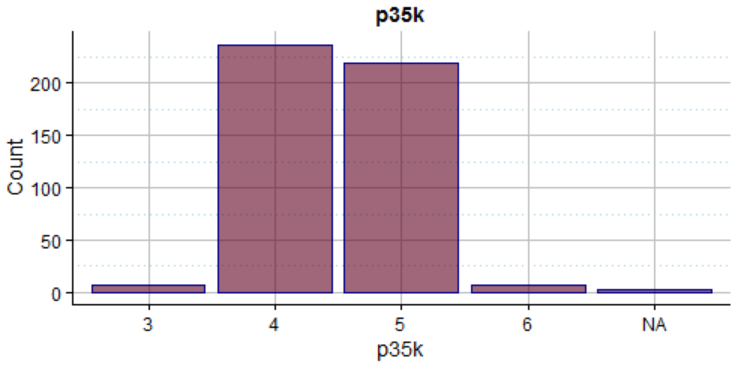
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	132.00	152.00	141.94	142.00	7.79	2.79	0.00	0.00
Distribution								

Table A.118: Description of *p35k*

p35k				
Meaning	This feature gives numeric information about potassium (K) [mEq/L].			
Statistics	mode	levels	# missings	% missings
	4.00	4.00	3.00	0.63
Distribution				

Note	Due to the fact that the feature has only 4 levels, it is visualised and measured like a categorical feature.

Table A.119: Description of *p36cl*

p36cl	
Meaning	This feature gives numeric information about chloride (Cl) [mEq/L].

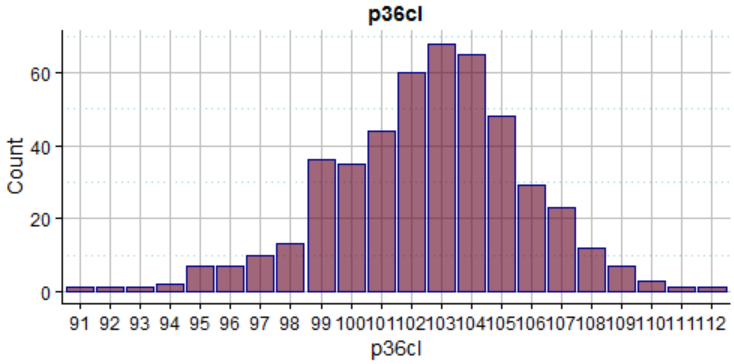
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	91.00	112.00	102.58	103.00	10.10	3.18	0.00	0.00
Distribution								

Table A.120: Description of *p37got*

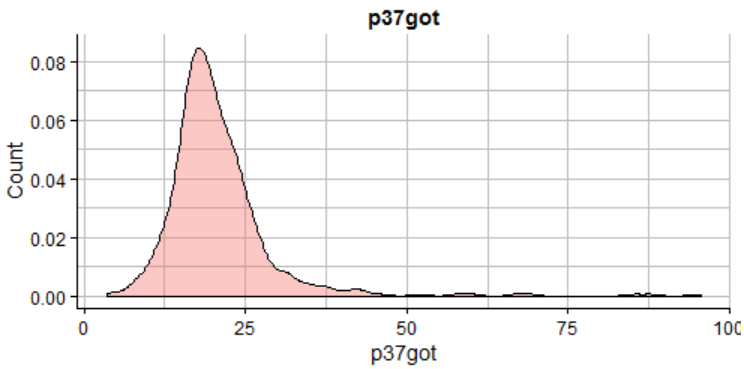
p37got								
Meaning	This feature gives numeric information about Glutamic-Oxaloacetic Transaminase (GOT) [U/L].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	4.00	95.00	20.85	19.00	82.60	9.09	0.00	0.00
Distribution								

Table A.121: Description of *p38gpt*

p38gpt								
Meaning	This feature gives numeric information about Glutamic-Pyruvic Transaminase (GPT) [U/L].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	5.00	153.00	19.94	18.00	140.88	11.87	0.00	0.00

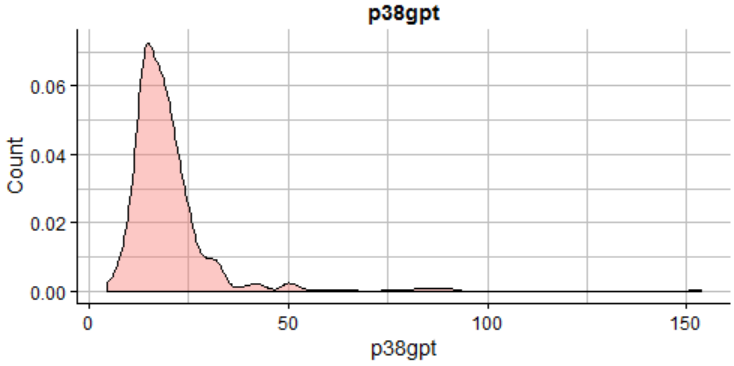
Distribution	

Table A.122: Description of $p39ggt$

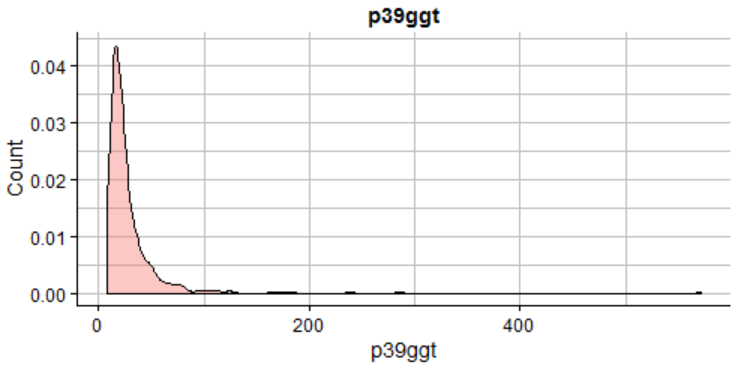
$p39ggt$								
Meaning	This feature gives numeric information about Gamma-Glutamyl Transferase (GGT) [U/L].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	8.00	571.00	29.50	21.00	1311.48	36.21	0.00	0.00
Distribution								

Table A.123: Description of $p40falc$

$p40falc$								
Meaning	This feature gives numeric information about Alkaline phosphatase [U/L].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	28.00	341.00	79.65	74.00	744.70	27.29	1.00	0.21

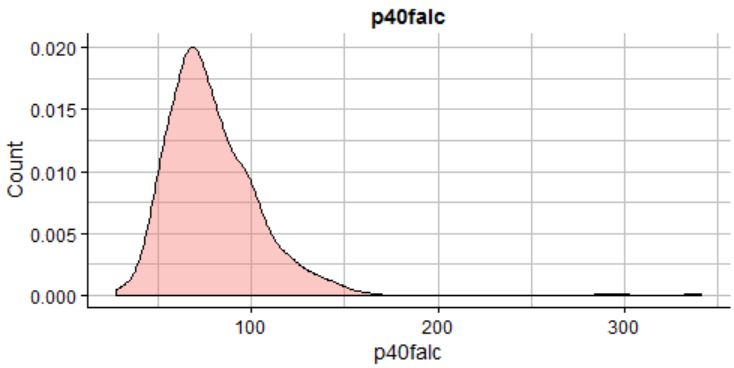
Distribution	

Table A.124: Description of $p41ldh$

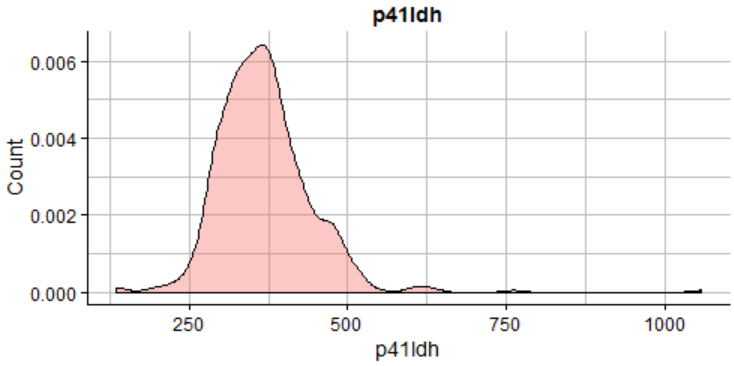
$p41ldh$								
Meaning	This feature gives numeric information about Lactate dehydrogenase (LDH) [U/L].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	135.00	1058.00	368.58	362.00	5784.88	76.06	2.00	0.42
Distribution								

Table A.125: Description of $p42fe$

$p42fe$								
Meaning	This feature gives numeric information about Iron (FE) [µg/dL].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	17.00	202.00	88.63	87.00	916.67	30.28	0.00	0.00

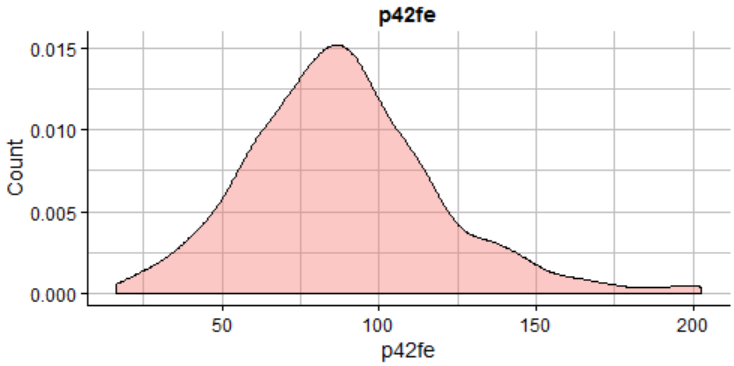
Distribution	

Table A.126: Description of $p43tfrr$

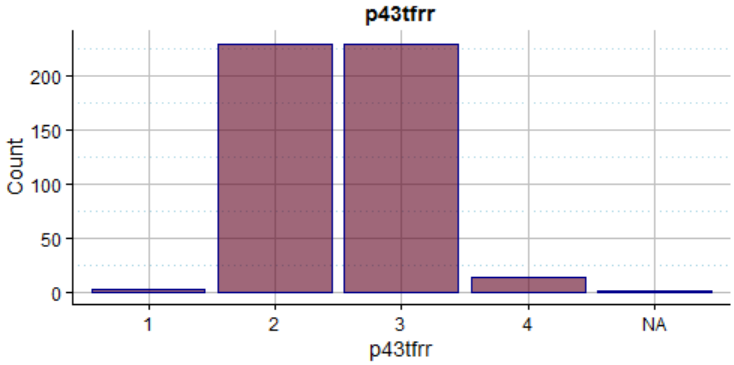
$p43tfrr$	
Meaning	This feature gives numeric information about Transferrin [mg/dL].
Statistics	mode: 2.00; levels: 4.00; # missings: 1.00; % missings: 0.21
Distribution	
Note	Due to the fact that the feature has only 4 levels, it is visualised and measured like a categorical feature.

Table A.127: Description of $p44pcrh$

$p44pcrh$	
Meaning	This feature gives numeric information about High-sensitivity C-reactive protein (hs-CRP) [mg/L].
Statistics	min: 0.00; max: 70.00; average: 5.61; median: 3.00; σ^2: 57.33; σ: 7.57; # missings: 71.00; % missings: 14.98

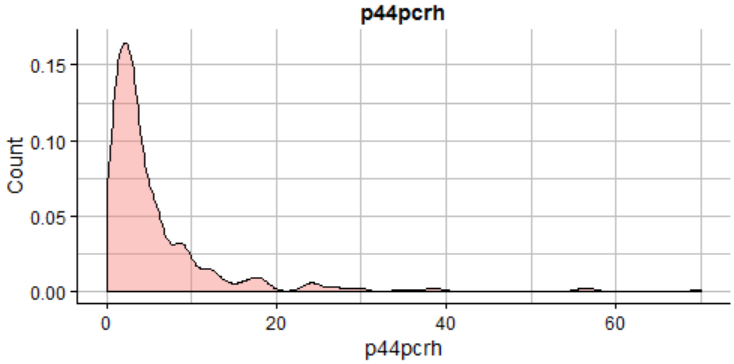
Distribution	

Table A.128: Description of *pasetotal*

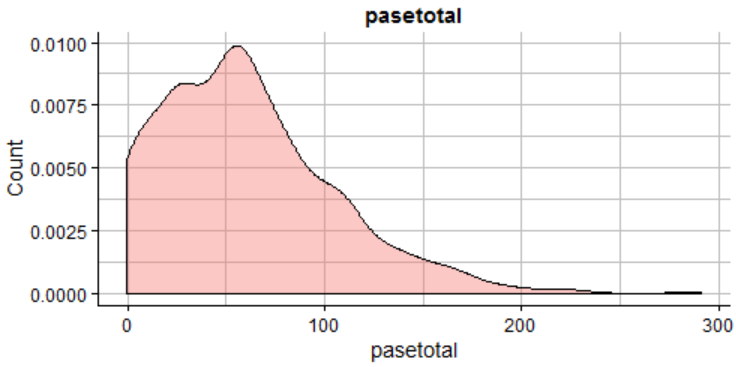
pasetotal								
Meaning	This feature gives binary information about the Physical activity scale for elderly score.							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	0.00	291.00	60.27	53.50	2012.90	44.87	0.00	0.00
Distribution								

Table A.129: Description of *ppeso*

ppeso				
Meaning	This feature gives binary information about the Fried criterion: "weight loss >10 lbs. in past year".			
Statistics	mode	levels	# missings	% missings
	0.00	2.00	0.00	0.00

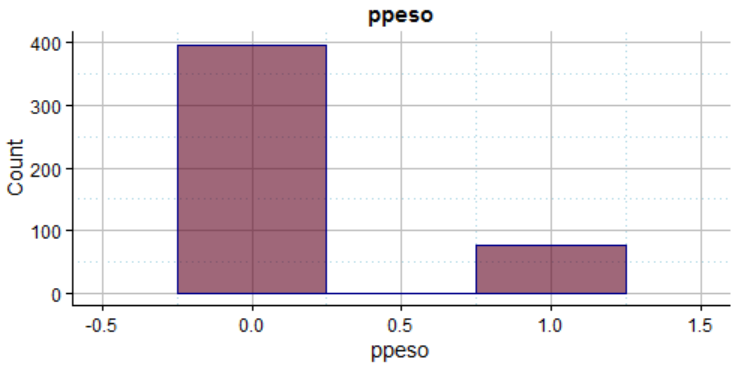
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.130: Description of *exhaution*

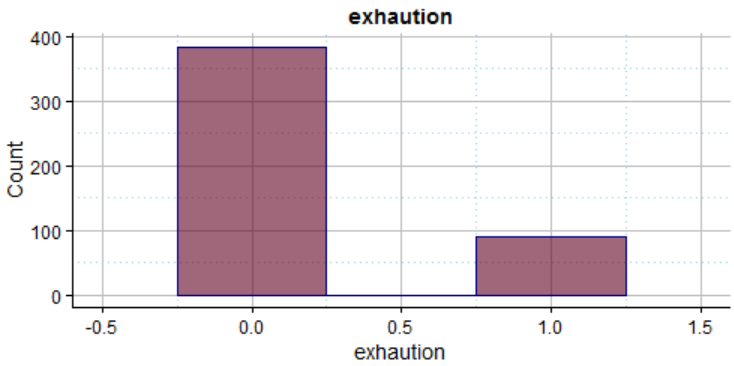
exhaution	
Meaning	This feature gives binary information about the Fried criterion: "exhaustion >=3days in past week".
Statistics	mode: 0.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.131: Description of *FRAGIL*

FRAGIL	
Meaning	This feature gives categorical information about Frail status according to Fried scale.

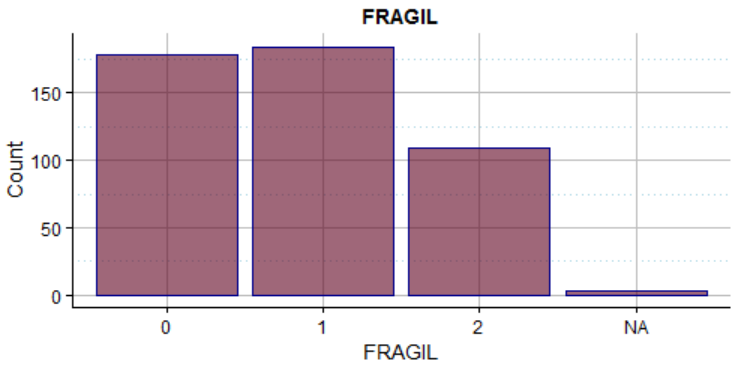
Statistics	mode	levels	# missings	% missings
	1.00	3.00	3.00	0.63
Distribution				
Discretization & Semantic scales	0: non-frail 1: pre-frail 2: frail			

Table A.132: Description of *ktaz*2008

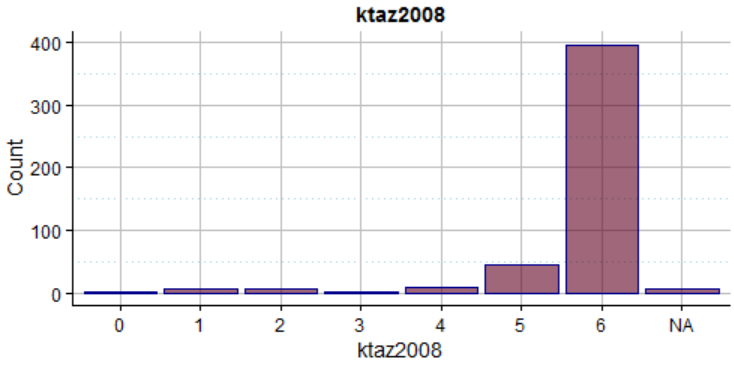
*ktaz*2008				
Meaning	This feature gives numeric information about Number of ADL abilities.			
Statistics	mode	levels	# missings	% missings
	6.00	7.00	6.00	1.27
Distribution				

Table A.133: Description of *lawton*2008

*lawton*2008

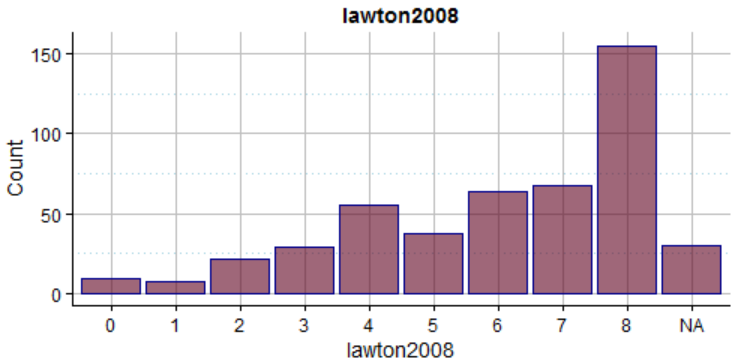
Meaning	This feature gives numeric information about the Number of IADL abilities.
Statistics	min: 0.00; max: 8.00; average: 5.93; median: 6.50; σ^2: 4.60; σ: 2.14; # missings: 30.00; % missings: 6.33
Distribution	

Table A.134: Description of *mmse*2008

*mmse*2008

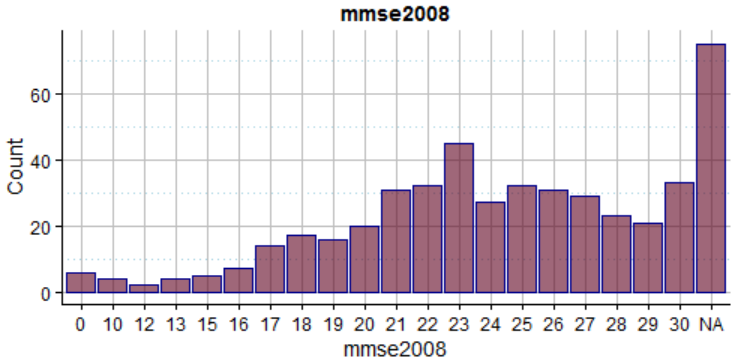
Meaning	This feature represents the raw MMSE score.
Statistics	min: 0.00; max: 30.00; average: 23.11; median: 23.00; σ^2: 26.42; σ: 5.14; # missings: 75.00; % missings: 15.82
Distribution	

Table A.135: Description of *pasefrag*

pasefrag

Meaning	This feature gives binary information about the Fried criterion: "pase score $\leqslant 20^{\text{th}}$ percentile"

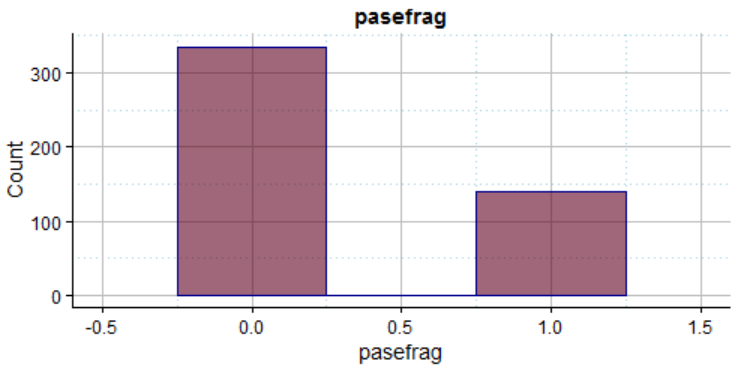
Statistics	mode	levels	# missings	% missings
	0.00	2.00	0.00	0.00
Distribution				
Discretization & Semantic scales	0: not true 1: true			

Table A.136: Description of *gdstotal*

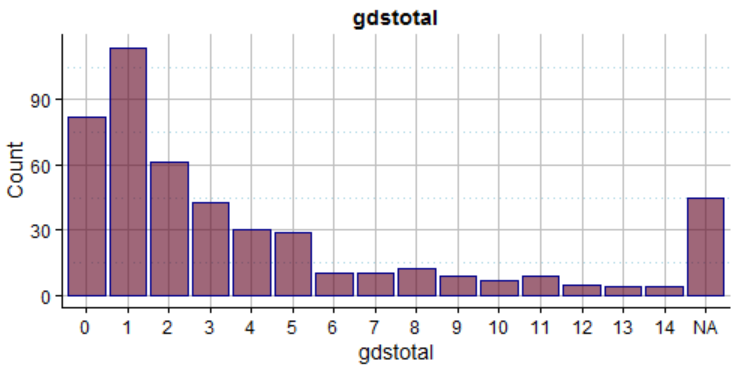
gdstotal								
Meaning	This feature gives numeric information about the total Geriatric Depression Score (GDS).							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	0.00	14.00	2.97	2.00	10.47	3.24	45.00	9.49
Distribution								

Table A.137: Description of *Depression*

Depression

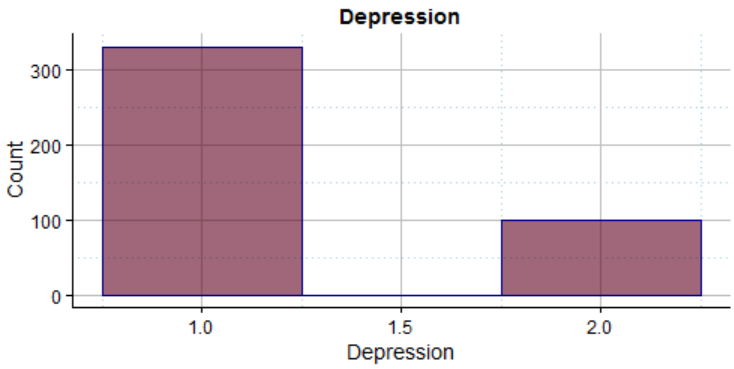
Meaning	This feature gives binary information about the presence of depression.
Statistics	mode: 1.00; levels: 2.00; # missings: 45.00; % missings: 9.49
Distribution	
Discretization & Semantic scales	0: GDS $<$ 5, therefore no depression present 1: GDS $\geqslant$ 5, therefore depression present

Table A.138: Description of *fuerzafragil*

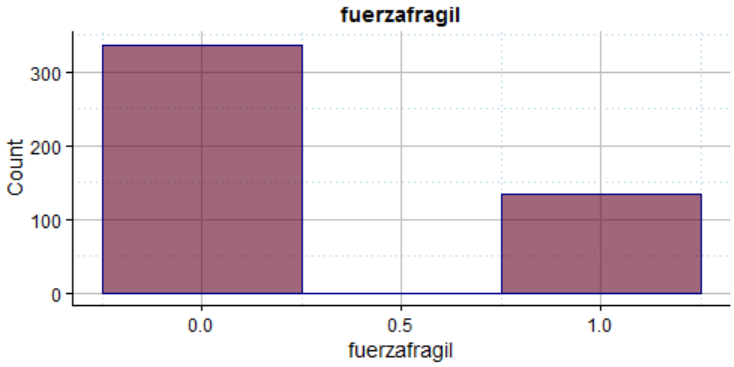
fuerzafragil	
Meaning	This feature gives binary information about the Fried criterion: "grip strength $\leqslant$ 20th percentile".
Statistics	mode: 0.00; levels: 2.00; # missings: 3.00; % missings: 0.63
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.139: Description of *marchafragil*

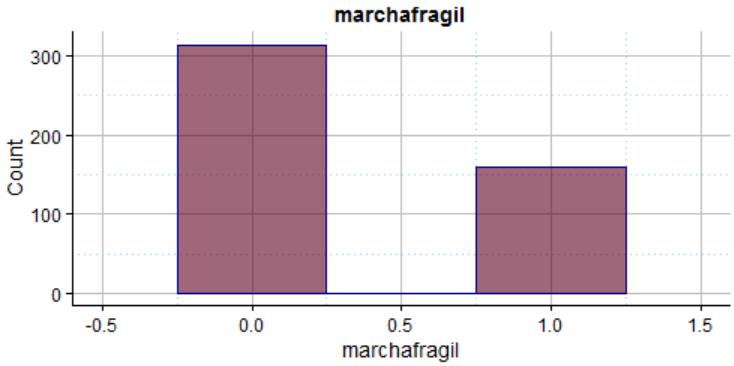
marchafragil	
Meaning	This feature gives binary information about the Fried criterion: "time to walk $\geqslant 80^{\text{th}}$ percentile".
Statistics	mode: 0.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.140: Description of *INSULINA*

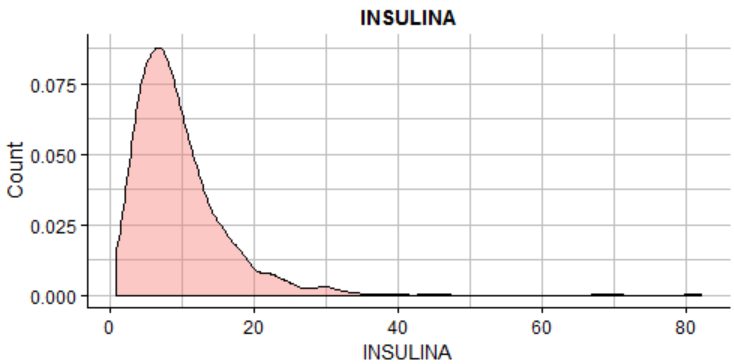
INSULINA	
Meaning	This feature gives numeric information about the blood insulin [U/mL].
Statistics	min: 1.00; max: 82.00; average: 9.91; median: 8.00; σ^2: 58.04; σ: 7.62; # missings: 55.00; % missings: 11.60
Distribution	

Table A.141: Description of HDL

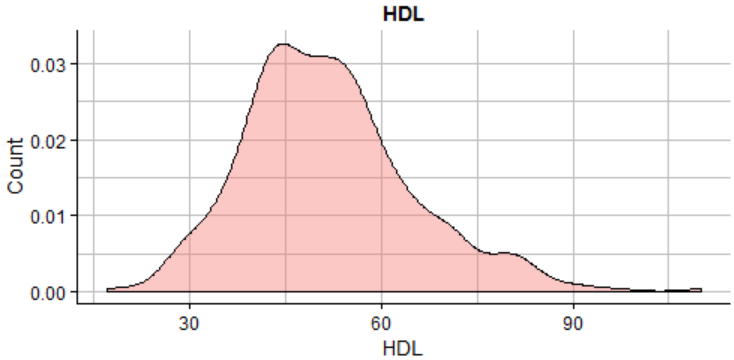
HDL	
Meaning	This feature gives numeric information about high-density lipoprotein (HDL) [mg/dL].
Statistics	min: 17.00; max: 110.00; average: 51.67; median: 50.00; σ^2: 176.16; σ: 13.27; # missings: 43.00; % missings: 9.07
Distribution	

Table A.142: Description of LDL

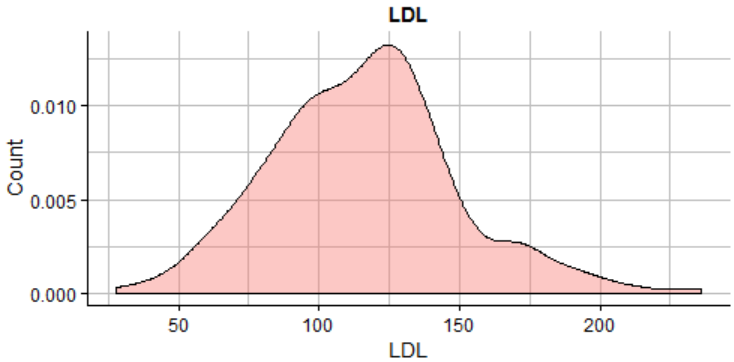
LDL	
Meaning	This feature gives binary information about the low-density lipoprotein (LDL) [mg/dL].
Statistics	min: 28.00; max: 236.00; average: 115.99; median: 117.00; σ^2: 1099.30; σ: 33.16; # missings: 43.00; % missings: 9.07
Distribution	

Table A.143: Description of $TESTOTOTAL$

$TESTOTOTAL$

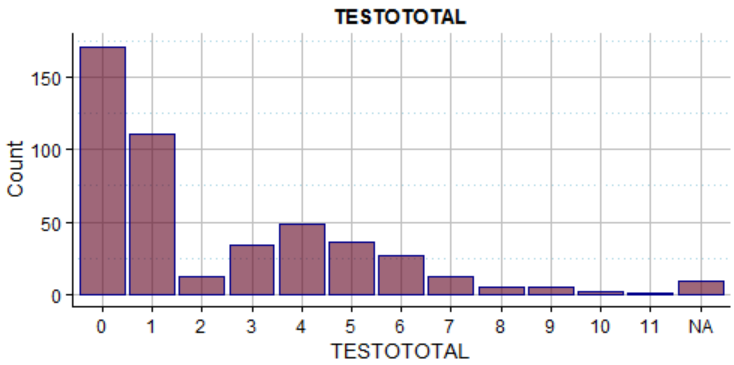
Meaning	This feature gives numeric information about the total testosterone in the blood [ng/dL].
Statistics	min: 0.00; max: 11.00; average: 2.10; median: 1.00; σ^2: 5.84; σ: 2.42; # missings: 9.00; % missings: 1.90
Distribution	

Table A.144: Description of *TESTOLIBRE*

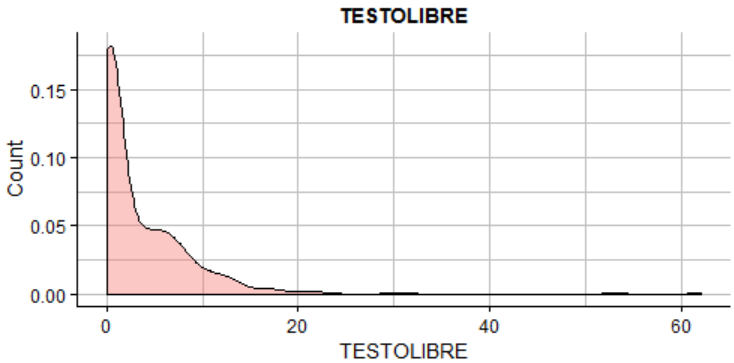
TESTOLIBRE	
Meaning	This feature gives numeric information about free testosterone in the blood [ng/dL].
Statistics	min: 0.00; max: 62.00; average: 3.56; median: 1.00; σ^2: 34.10; σ: 5.84; # missings: 9.00; % missings: 1.90
Distribution	

Table A.145: Description of *codigo*01

*codigo*01	
Meaning	This feature gives binary information about the question "Is the patient dead at follow up?".

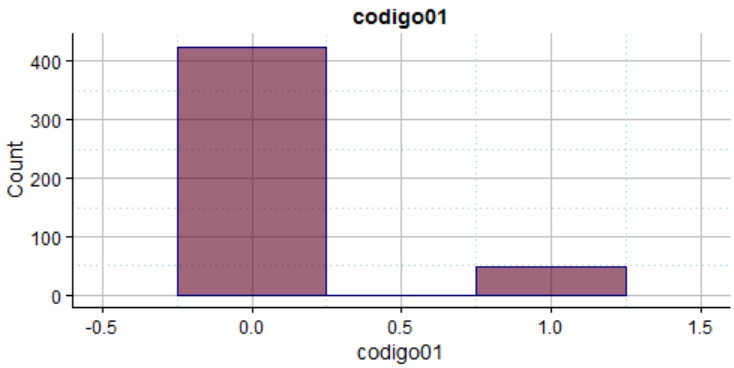
Statistics	mode	levels	# missings	% missings
	0.00	2.00	0.00	0.00
Distribution				
Discretization & Semantic scales	0: alive 1: dead			

Table A.146: Description of *ADMA*

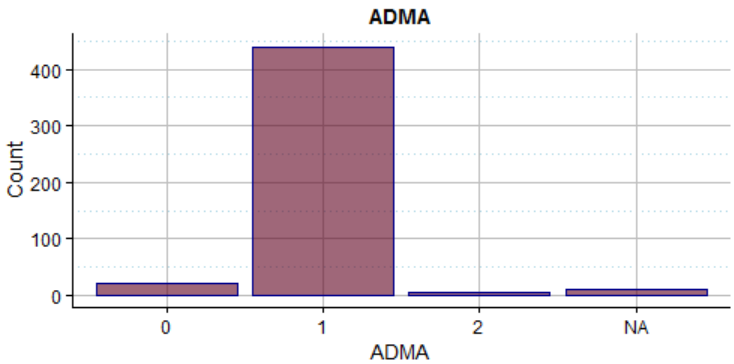
ADMA				
Meaning	This feature gives numeric information about asymmetric dimethylarginine (ADMA) [µmol/L].			
Statistics	mode	levels	# missings	% missings
	1.00	3.00	9.00	1.90
Distribution				
Note	Due to the fact that the feature has only 3 levels, it is visualised and measured like a categorical feature.			

Table A.147: Description of *lawton*2013

*lawton*2013

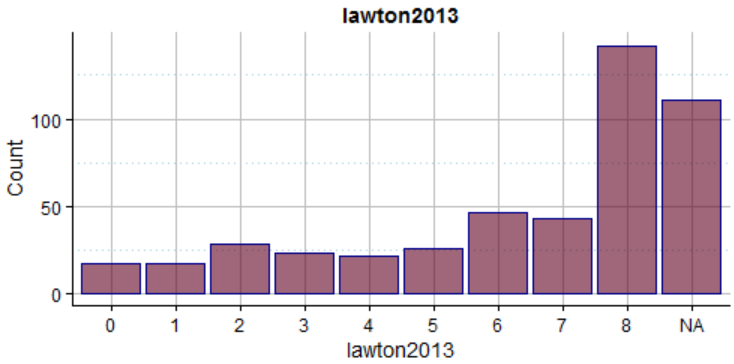
Meaning	This feature gives numeric information about the number of IADL abilities on the follow up.
Statistics	min: 0.00; max: 8.00; average: 5.70; median: 7.00; σ^2: 6.58; σ: 2.56; # missings: 111.00; % missings: 23.42
Distribution	

Table A.148: Description of *katz*_2013

*katz*_2013

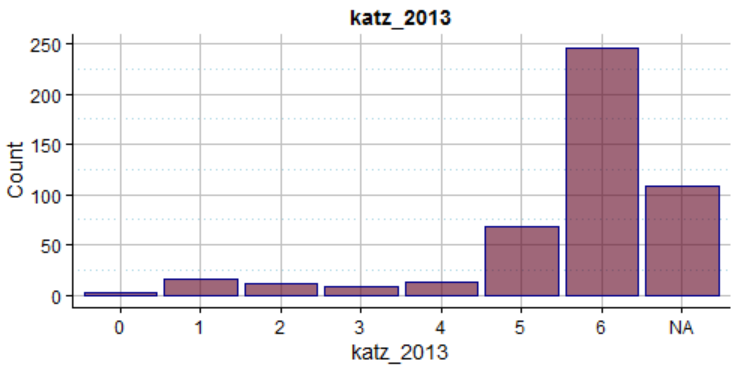
Meaning	This feature gives numeric information about the number of ADL abilities on the follow up.
Statistics	mode: 6.00; levels: 7.00; # missings: 108.00; % missings: 22.78
Distribution	

Table A.149: Description of *FRAGIL*_2013

*FRAGIL*_2013

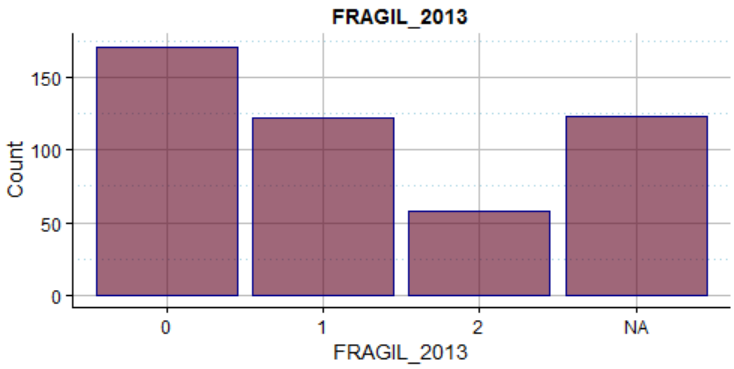
Meaning	This feature gives categorical information about the frailty status according to Fried scale on the follow up.
Statistics	mode: 0.00; levels: 3.00; # missings: 123.00; % missings: 25.95
Distribution	
Discretization & Semantic scales	0: non-frail 1: pre-frail 2: frail

Table A.150: Description of $em1$

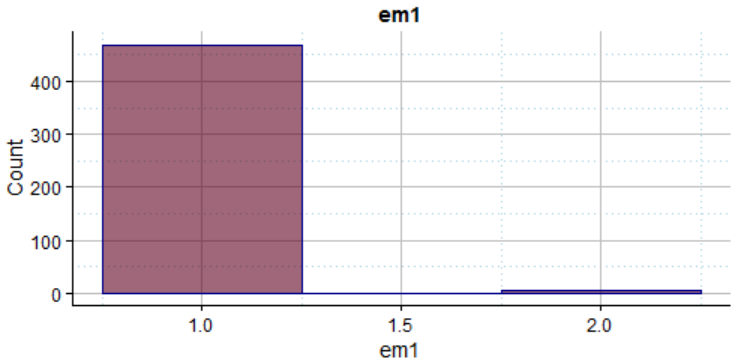
$em1$	
Meaning	This Mobility Score (MS) related features gives binary information about the question "Are you able to walk at home?".
Statistics	mode: 1.00; levels: 2.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.151: Description of *em1a*

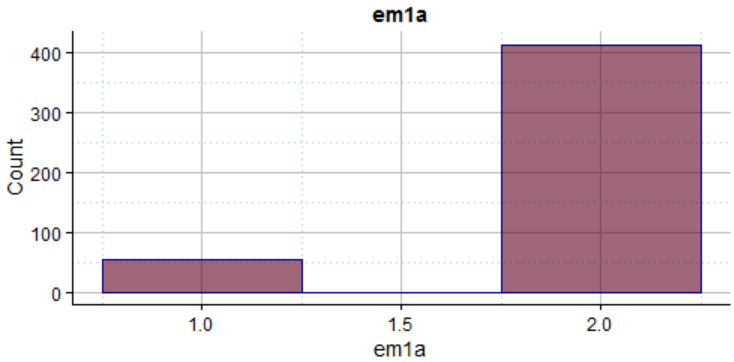
em1a	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you get tired when doing it?".
Statistics	mode: 2.00; levels: 2.00; # missings: 5.00; % missings: 1.05
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.152: Description of *em1b*

em1b	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you need help when doing it?".
Statistics	mode: 2.00; levels: 2.00; # missings: 6.00; % missings: 1.27

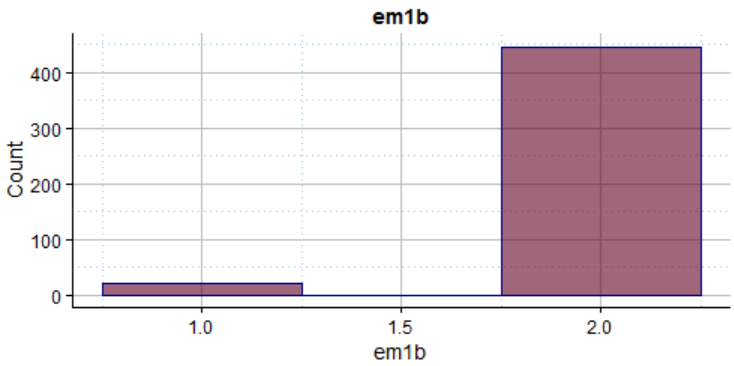
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.153: Description of $em2$

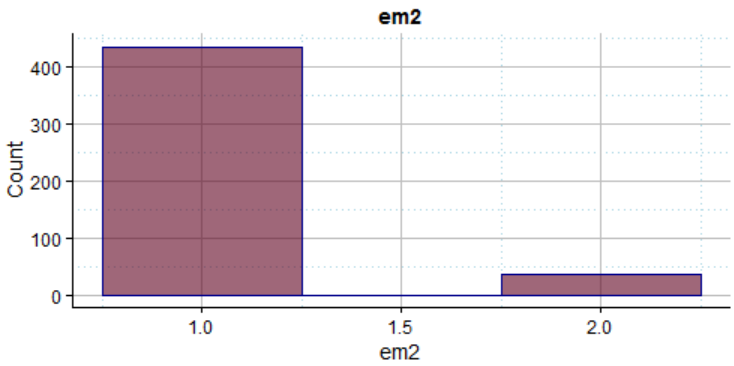
$em2$	
Meaning	This Mobility Score (MS) related features gives binary information about the question "Are you able to go out from home?".
Statistics	mode: 1.00; levels: 2.00; # missings: 2.00; % missings: 0.42
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.154: Description of $em2a$

$em2a$

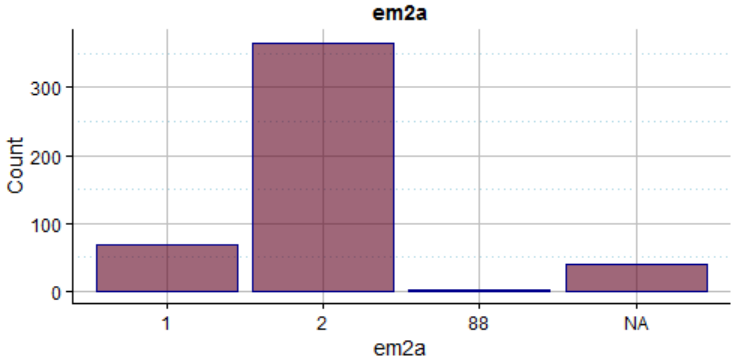
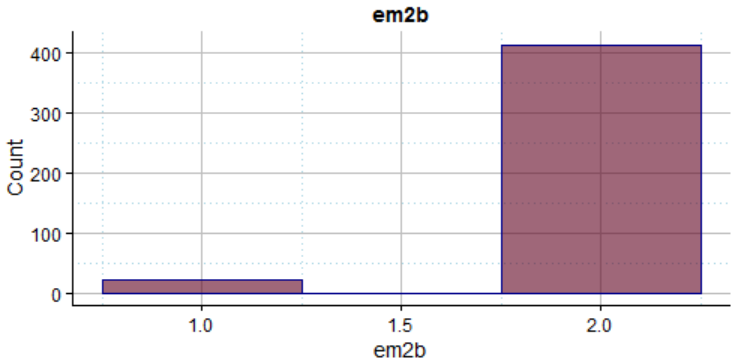
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you get tired when doing it?".
Statistics	mode: 2.00; levels: 3.00; # missings: 39.00; % missings: 8.23
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.155: Description of *em2b*

em2b	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you need help when doing it?".
Statistics	mode: 2.00; levels: 2.00; # missings: 40.00; % missings: 8.44
Distribution	

Discretization & Semantic scales	1: yes 2: no

Table A.156: Description of $em3$

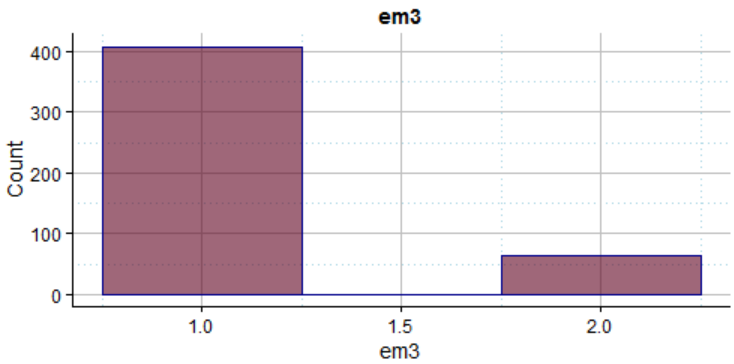
$em3$	
Meaning	This Mobility Score (MS) related features gives binary information about the question "Are you able to climb stairs?".
Statistics	mode: 1.00; levels: 2.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.157: Description of $em3a$

$em3a$	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you get tired when doing it?".
Statistics	mode: 2.00; levels: 3.00; # missings: 67.00; % missings: 14.13

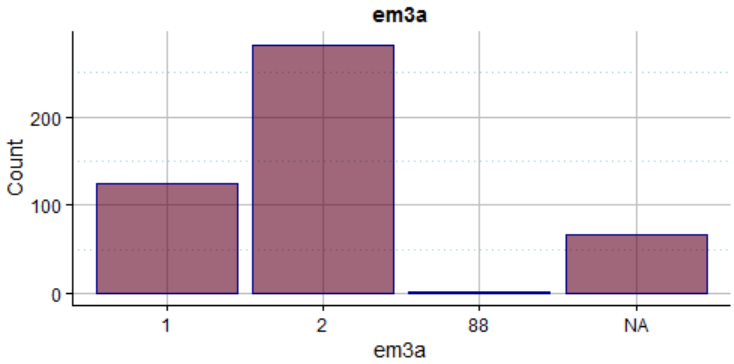
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.158: Description of *em3b*

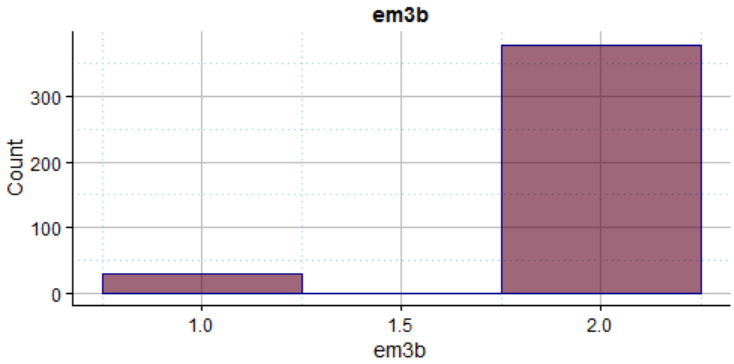
em3b				
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you need help when doing it?".			
Statistics	mode	levels	# missings	% missings
	2.00	2.00	66.00	13.92
Distribution				
Discretization & Semantic scales	1: yes 2: no			

Table A.159: Description of *em4*

em4

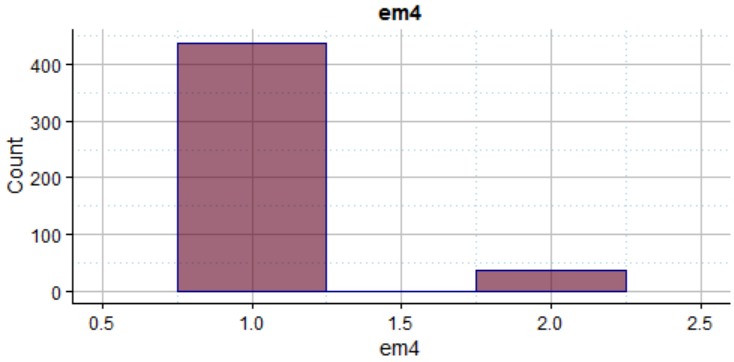
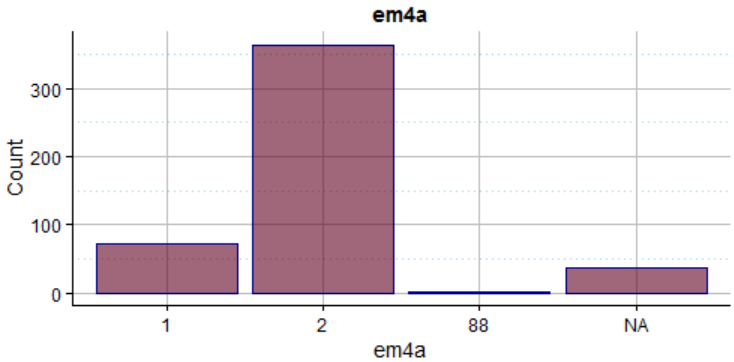
Meaning	This Mobility Score (MS) related features gives binary information about the question "Are you able to walk outside (nice weather)?".
Statistics	mode: 1.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.160: Description of *em4a*

em4a	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you get tired when doing it?".
Statistics	mode: 2.00; levels: 3.00; # missings: 36.00; % missings: 7.59
Distribution	

Discretization & Semantic scales	1: yes 2: no

Table A.161: Description of *em4b*

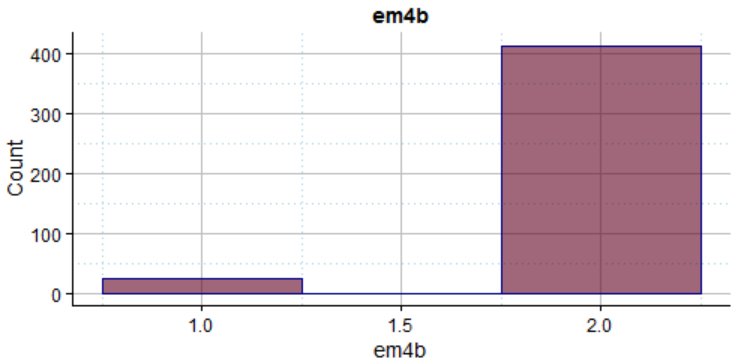
em4b	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you need help when doing it?".
Statistics	mode: 2.00; levels: 2.00; # missings: 36.00; % missings: 7.59
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.162: Description of *em5*

em5	
Meaning	This Mobility Score (MS) related features gives binary information about the question "Are you able to walk outside (bad weather)?".
Statistics	mode: 1.00; levels: 2.00; # missings: 1.00; % missings: 0.21

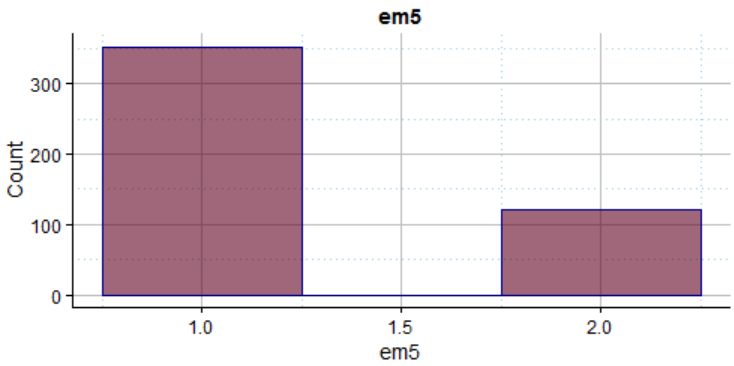
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.163: Description of *em5a*

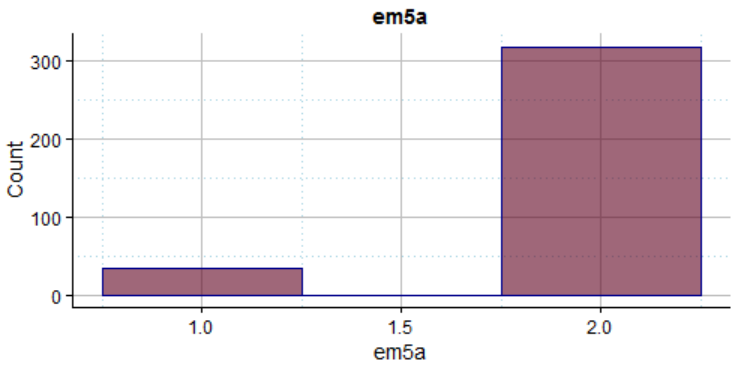
em5a	
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you get tired when doing it?".
Statistics	mode: 2.00; levels: 2.00; # missings: 123.00; % missings: 25.95
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.164: Description of *em5b*

em5b

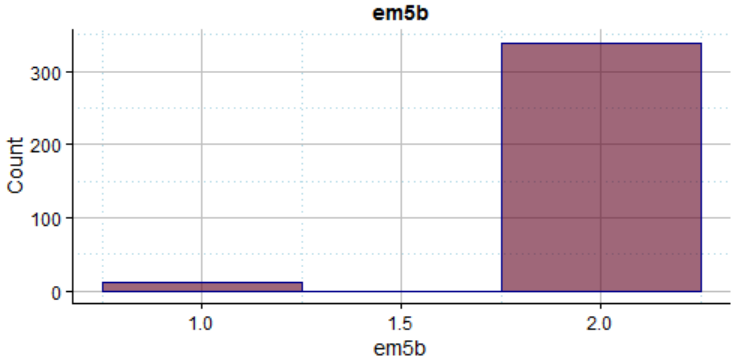
Meaning	This Mobility Score (MS) related features gives binary information about the question "If answered YES; Do you need help when doing it?".
Statistics	mode: 2.00; levels: 2.00; # missings: 124.00; % missings: 26.16
Distribution	
Discretization & Semantic scales	1: yes 2: no

Table A.165: Description of *tads*

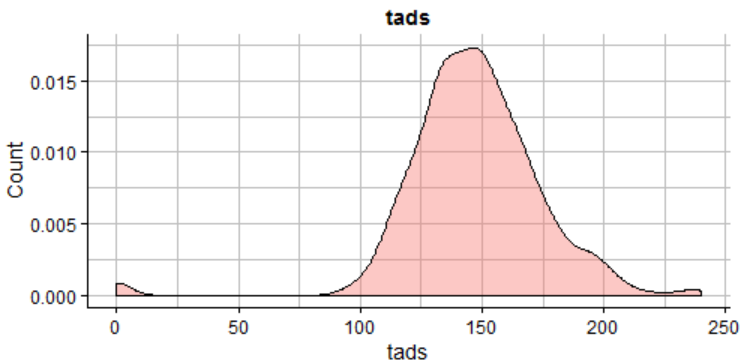
tads	
Meaning	This feature gives numeric information about the systolic blood pressure.
Statistics	min: 0.00; max: 240.00; average: 146.65; median: 146.00; σ^2: 816.69; σ: 28.58; # missings: 4.00; % missings: 0.84
Distribution	

Table A.166: Description of *tadd*

tadd

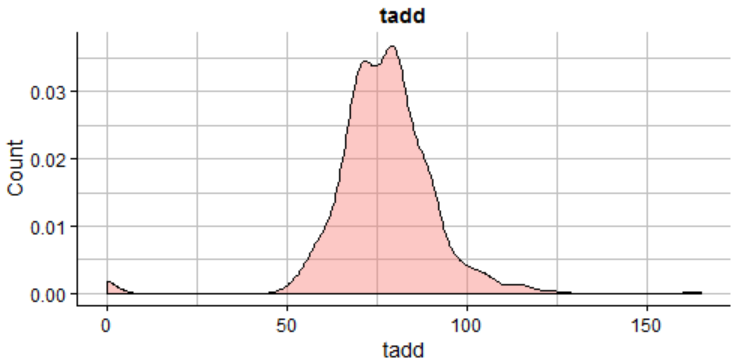
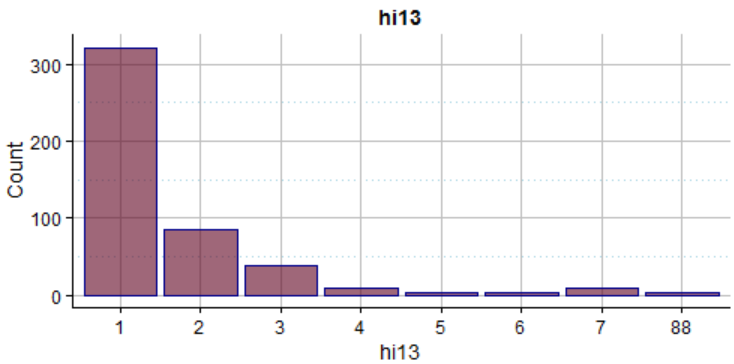
Meaning	This feature gives numeric information about the diastolic blood pressure.
Statistics	min: 0.00; max: 165.00; average: 77.14; median: 77.50; σ^2: 219.77; σ: 14.82; # missings: 4.00; % missings: 0.84
Distribution	

Table A.167: Description of *hi*13

*hi*13

Meaning	This feature gives categorical information about the education level.
Statistics	mode: 1.00; levels: 8.00; # missings: 0.00; % missings: 0.00
Distribution	

Discretization & Semantic scales	1: none 2: unfinished school 3: school 4: secondary school 5: professional school 6: university, technical grade (3 years) 7 university, grade (5 years) 8-10 nan or missing

Table A.168: Description of *enpot*1

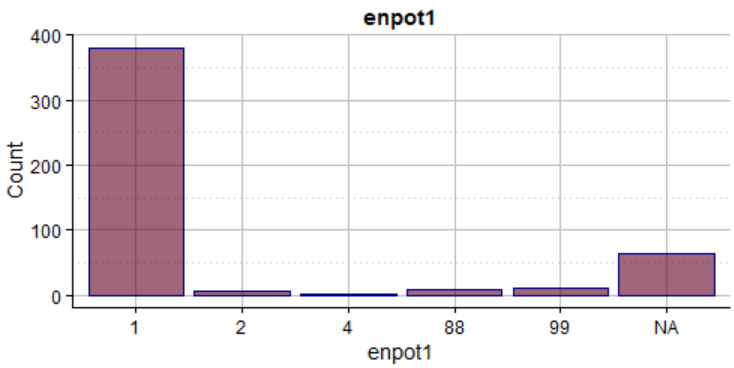
*enpot*1	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What day of the week is this?".
Statistics	mode: 1.00; levels: 5.00; # missings: 65.00; % missings: 13.71
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.169: Description of *enpot*2

*enpot*2

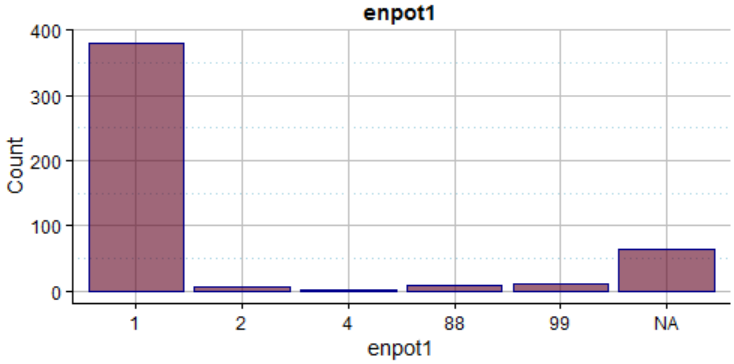
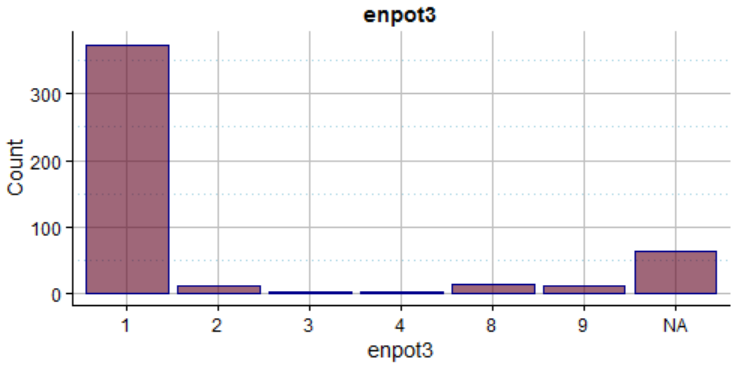
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What is today's date?".
Statistics	mode: 1.00; levels: 5.00; nMissings: 65.00; nMissingsPerc: 13.71
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.170: Description of *enpot3*

enpot3	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What month is this?".
Statistics	mode: 1.00; levels: 6.00; # missings: 62.00; % missings: 13.08
Distribution	

Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.171: Description of *enpot*4

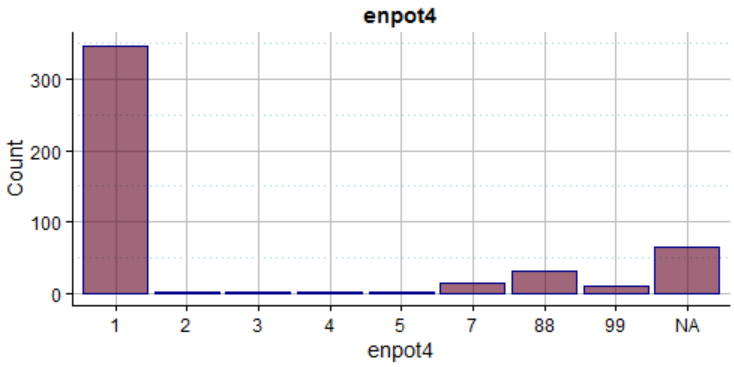
*enpot*4	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What year is this?".
Statistics	mode: 1.00; levels: 8.00; # missings: 65.00; % missings: 13.71
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.172: Description of *enpot*6

*enpot*6	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "Which season is this?".
Statistics	mode: 1.00; levels: 4.00; # missings: 61.00; % missings: 12.87

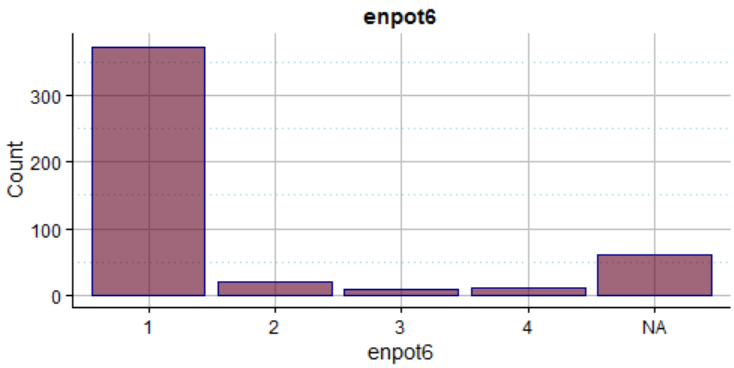
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.173: Description of *enpol*1

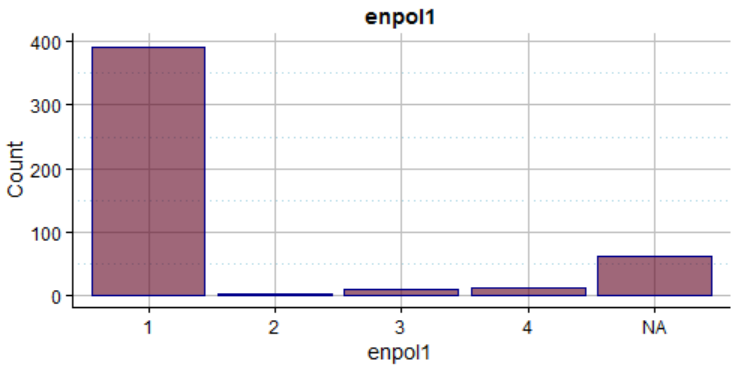
*enpol*1	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "IN HOME: What is the street address of this house? // IN FACILITY: What is the name of this building?".
Statistics	mode: 1.00; levels: 4.00; # missings: 62.00; % missings: 13.08
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.174: Description of *enpol*2

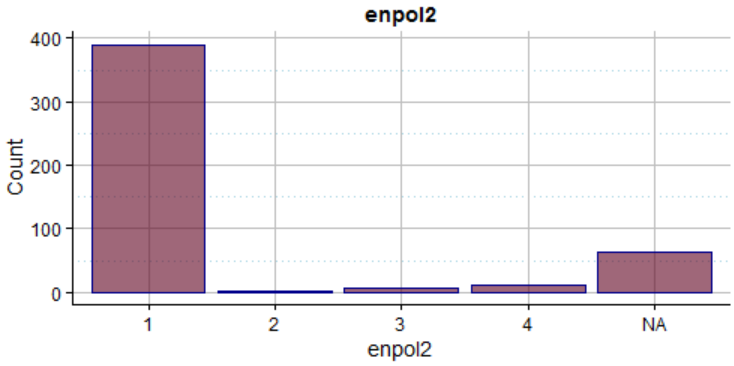
*enpol*2	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "IN HOME: What room are we in? // IN FACILITY: What floor are we on?".
Statistics	mode: 1.00; levels: 4.00; # missings: 63.00; % missings: 13.29
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.175: Description of *enpol*3

*enpol*3	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What city/town are we in?".
Statistics	mode: 1.00; levels: 4.00; # missings: 63.00; % missings: 13.29

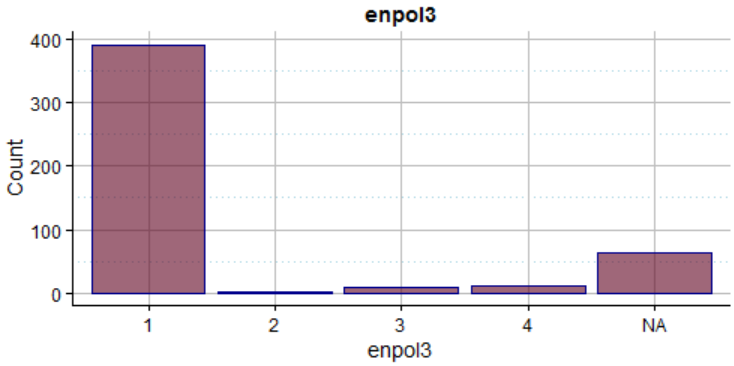
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.176: Description of *enpol*4

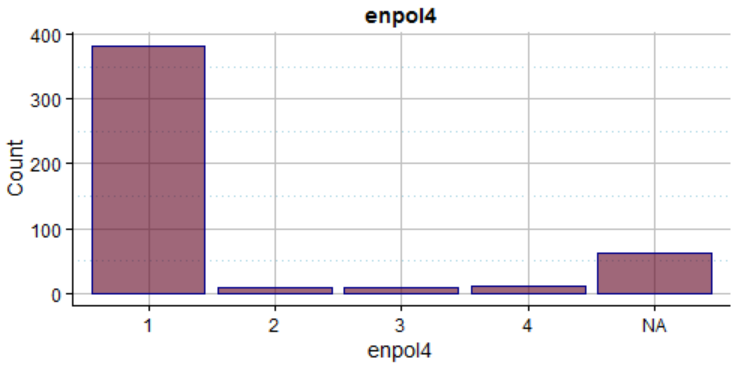
*enpol*4	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What province are we in?".
Statistics	mode: 1.00; levels: 4.00; # missings: 63.00; % missings: 13.29
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.177: Description of *enpol*5

enpol5

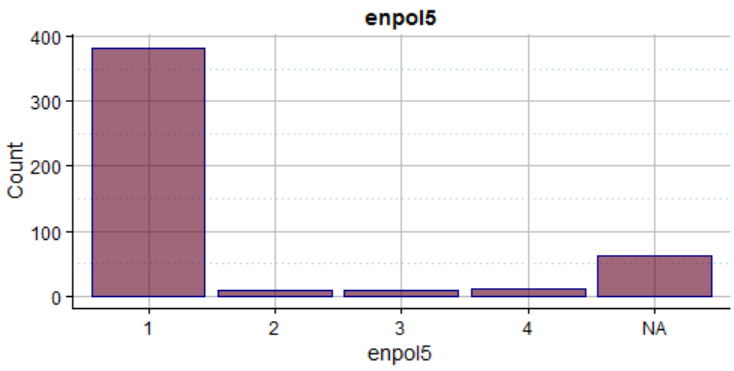
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "What county are we in?".
Statistics	mode: 1.00; levels: 4.00; # missings: 63.00; % missings: 13.29
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.178: Description of *enmem1a*

enmem1a	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do the following MMSE-task: "SAY: I am going to name three objects. When I am finished, I want you to repeat theM. Remember what they are because I am going to ask you to name them again in a few minutes. // Say the following words slowly at 1-second intervals - peseta (coin in spanish), caballo (horse in spanish), manzana (apple in spanish)".
Statistics	mode: 4.00; levels: 6.00; # missings: 62.00; % missings: 13.08

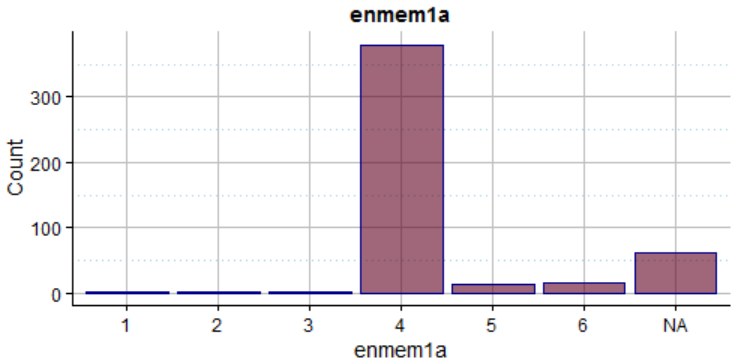
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.179: Description of *enpmem*2

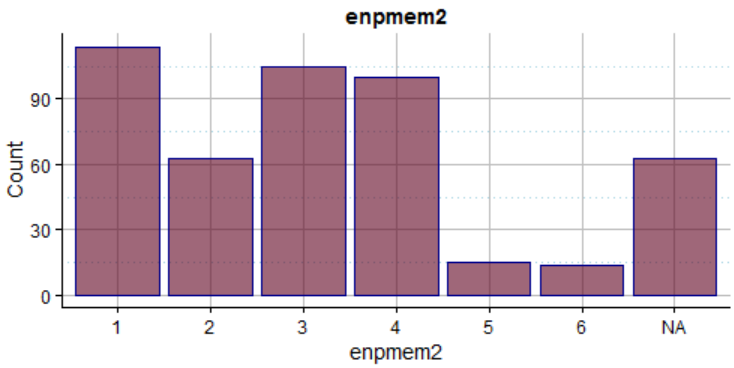
*enpmem*2	
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the following MMSE-question: "Now what were the three objects I asked you to remember?"
Statistics	mode: 1.00; levels: 6.00; # missings: 63.00; % missings: 13.29
Distribution	
Discretization & Semantic scales	1: correct answered 2: not correct answered other: missing

Table A.180: Description of *enpat*1

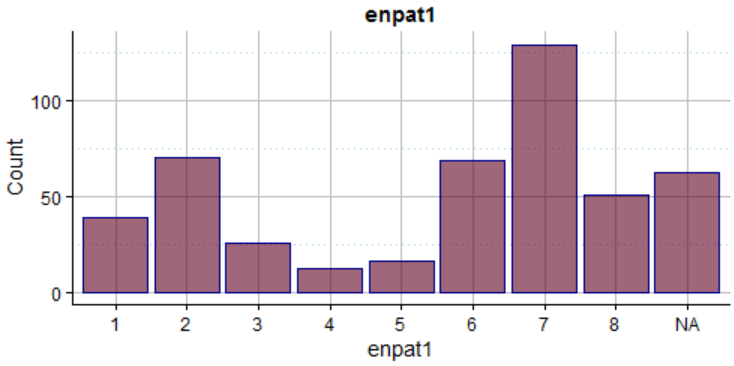
*enpat*1	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do the following MMSE-task: "Count backwards by 7 starting from 100".
Statistics	mode: 7.00; levels: 8.00; # missings: 62.00; % missings: 13.08
Distribution	
Discretization & Semantic scales	0: not correct 1: one letter correct 2: two letters correct 3: three letters correct 4: four letters correct 5: five letters correct 8: can't do it 9: won't do it

Table A.181: Description of *enpat*2

*enpat*2	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do the following MMSE-task: "Spell the word MUNDO (world in spanish). Now spell it backwards."

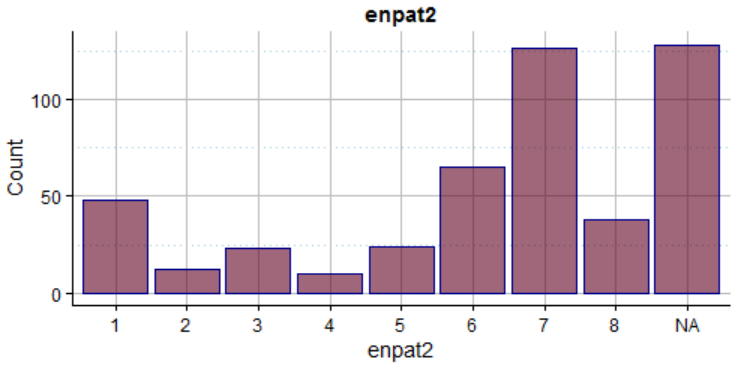
Statistics	mode	levels	# missings	% missings
	7.00	8.00	128.00	27.00
Distribution				
Discretization & Semantic scales	0: not correct 1: one letter correct 2: two letters correct 3: three letters correct 4: four letters correct 5: five letters correct 8: can't do it 9: won't do it			

Table A.182: Description of *enleng*1

*enleng*1				
Meaning	This MMSE related feature gives binary information about the ability of the patient to answer the MMSE-question "Show a wristchatch and a pencil. What are these called?".			
Statistics	mode	levels	# missings	% missings
	3.00	4.00	66.00	13.92

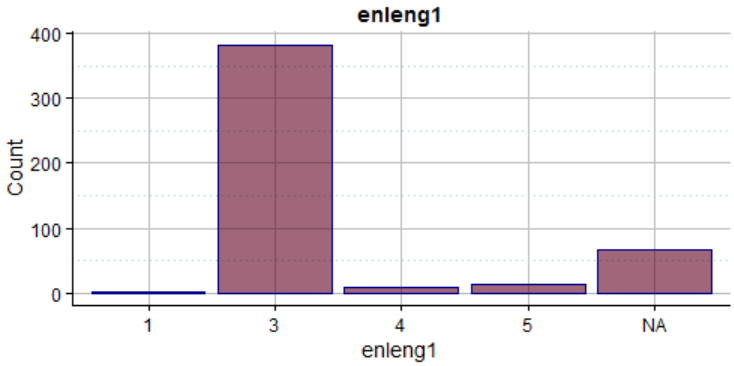
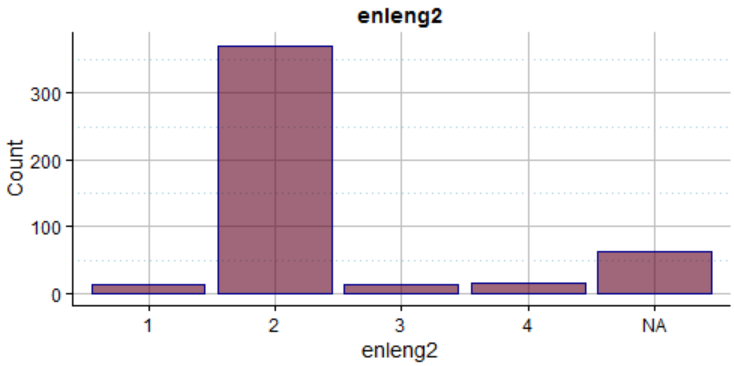
Distribution	
Discretization & Semantic scales	0: not correct 1: one correct 2: two correct 8: can't do it 9: won't do it

Table A.183: Description of *enleng*2

*enleng*2	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do following MMSE-task: "SAY: I would like you to repeat this phrase after me: Ni si, ni no, ni pero. (No ifs, ands or buts. In spanish) ".
Statistics	mode: 2.00; levels: 4.00; # missings: 62.00; % missings: 13.08
Distribution	

Discretization & Semantic scales	0: not correct 1: correct 8: can't do it 9: won't do it

Table A.184: Description of *enleng3*

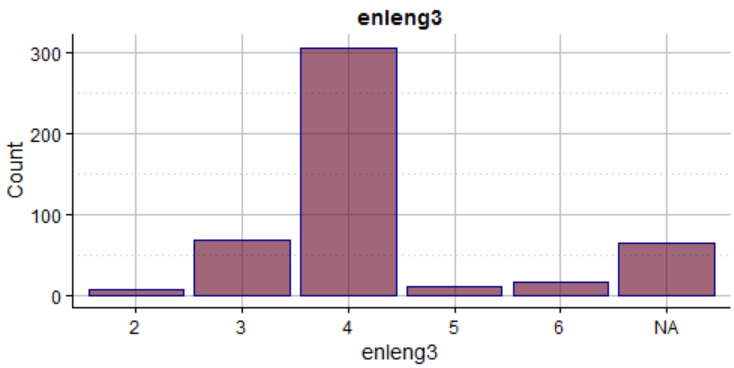
enleng3	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do following MMSE-task: "Ask the person if he is right or left handed. Take a piece of paper and hold it up in front of the person. SAY: Take this paper in your right/left hand (whichever is non-dominant), fold the paper in half once with both hands and put the paper down on the floor. Score 1 point for each instruction executed correctly.".
Statistics	mode: 4.00; levels: 5.00; # missings: 64.00; % missings: 13.50
Distribution	
Discretization & Semantic scales	0: not correct 1: one correct 2: two correct 3: three correct 8: can't do it 9: won't do it

Table A.185: Description of *enleng4*

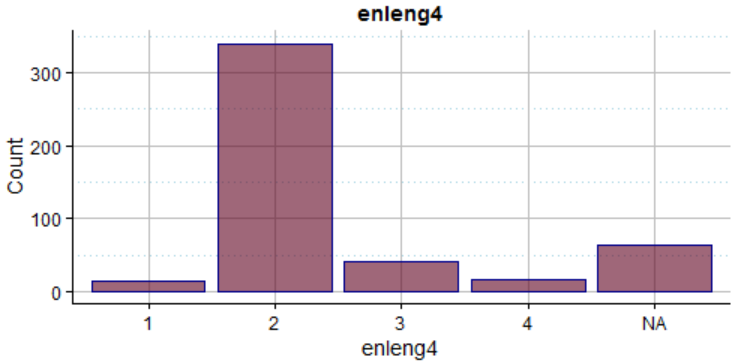
enleng4	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do following MMSE-task: "SAY: Read the words on the page and then do what it says. Then hand the person the sheet with "Cierre los ojos" (close your eyes in spanish) on it. If the subject read and does not close their eyes, repeat yp to three times. Score only if subject closes eyes.".
Statistics	mode: 2.00; levels: 4.00; # missings: 63.00; % missings: 13.29
Distribution	
Discretization & Semantic scales	0: not correct 1: correct 8: can't do it 9: won't do it

Table A.186: Description of *enpprx1*

enpprx1	
Meaning	This MMSE related feature gives binary information about the ability of the patient to do following MMSE-task: "Hand the person a pencil and paper. SAY: write any complete sentence on that piece of paper. (Note: The sentence must make sense. Ignore spelling errors)".

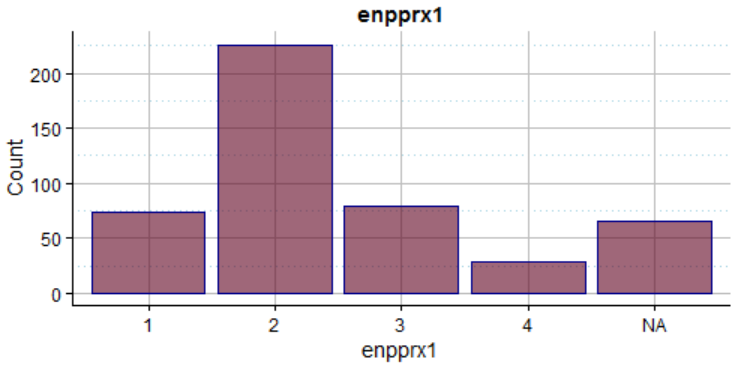
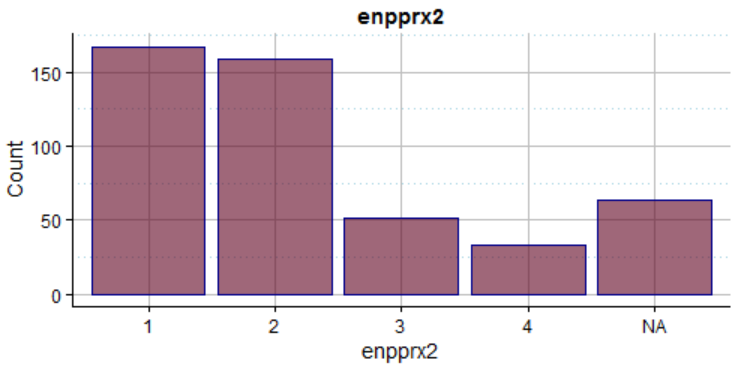
Statistics	mode	levels	# missings	% missings
	2.00	4.00	66.00	13.92
Distribution				
Discretization & Semantic scales	0: not correct 1: correct 8: can't do it 9: won't do it			

Table A.187: Description of *enpprx2*

enpprx2				
Meaning	This MMSE related feature gives binary information about the ability of the patient to do following MMSE-task: "Place design, eraser and pencil in front of the person. SAY: copy this design please. // Allow multiple tries. Wait until person is finished and hands it back. Score only for correctly copied diagram with a 4-sided figure between two 5-sided figures. ".			
Statistics	mode	levels	# missings	% missings
	1.00	4.00	64.00	13.50
Distribution				

Discretization & Semantic scales	0: not correct 1: correct 8: can't do it 9: won't do it

Table A.188: Description of $k1\_2013$

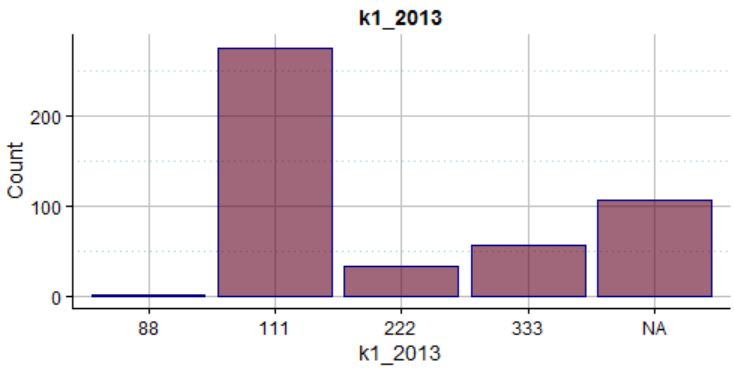
$k1\_2013$	
Meaning	This feature gives categorical information about the WHO activity 6: "Any difficulty washing face and arms?". This feature is associated with the ADL test, and represents question 1 (follow up, 2008).
Statistics	mode: 111.00; levels: 4.00; # missings: 107.00; % missings: 22.57
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (independent, score=1) 333: Unable to do it (dependent, score=0)

Table A.189: Description of $k2\_2013$

$k2\_2013$

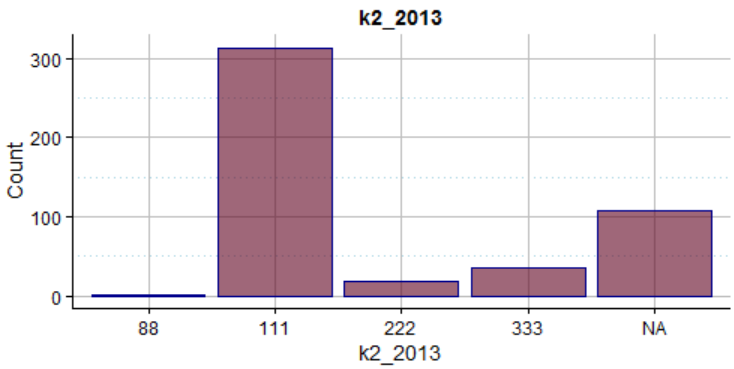
Meaning	This feature gives categorical information about the WHO activity 8: "Any difficulty dressing and undressing?". This feature is associated with the ADL test, and represents question 2 (follow up, 2013).
Statistics	mode: 111.00; levels: 4.00; # missings: 107.00; % missings: 22.57
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (independent, score=1) 333: Unable to do it (dependent, score=0)

Table A.190: Description of $k3\_2013$

$k3\_2013$	
Meaning	This feature gives categorical information about the WHO activity 11: "Any difficulty using the toilet?". This feature is associated with the ADL test, and represents question 3.
Statistics	mode: 111.00; levels: 4.00; # missings: 107.00; % missings: 22.57

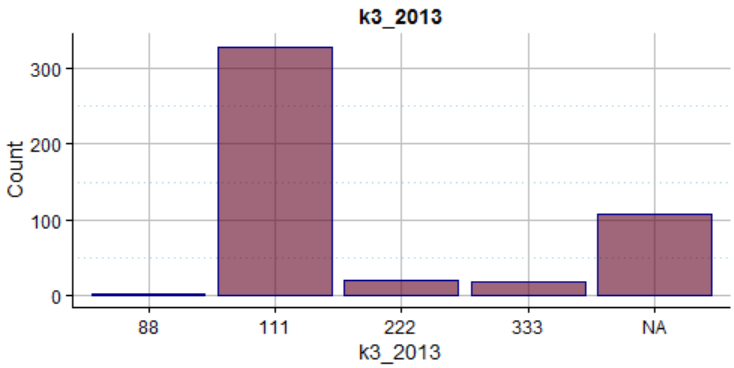
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (dependent, score=0) 333: Unable to do it (dependent, score=0)

Table A.191: Description of $k4\_2013$

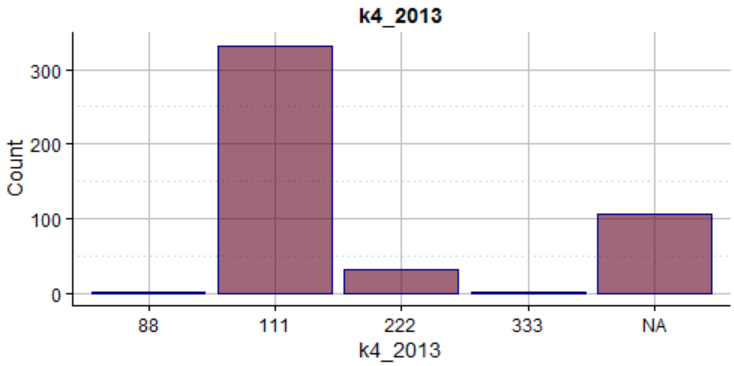
$k4\_2013$	
Meaning	This feature gives categorical information about the WHO activity 12: "Any difficulty getting in and out of bed?". This feature is associated with the ADL test, and represents question 4 (follow up, 2013).
Statistics	mode: 111.00; levels: 4.00; # missings: 107.00; % missings: 22.57
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (dependent, score=0) 333: Unable to do it (dependent, score=0)

Table A.192: Description of $k5\_2013$

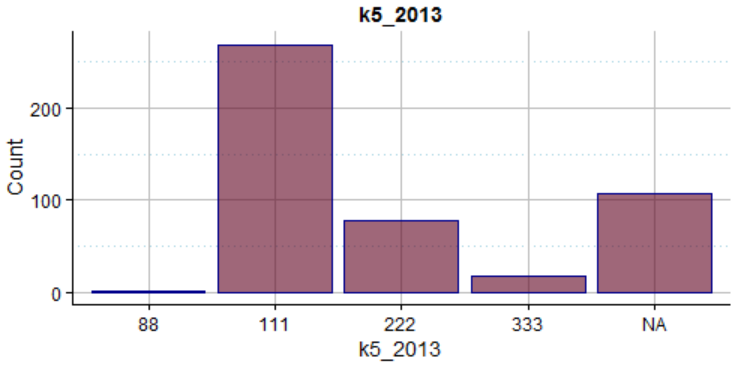
$k5\_2013$	
Meaning	This feature gives categorical information about the WHO activity 19: "Any difficulty controlling urination and bowel movements?".This feature is associated with the ADL test, and represents question 5 (follow up, 2013).
Statistics	mode: 111.00; levels: 4.00; # missings: 108.00; % missings: 22.78
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (dependent, score=0) 333: Unable to do it (dependent, score=0)

Table A.193: Description of $k6\_2013$

$k6\_2013$	
Meaning	This feature gives categorical information about the WHO activity 9: "Any difficulty eating (e.g.,holding a fork, cutting food, drinking from a glass)?". This feature is associated with the ADL test, and represents question 6 (follow up, 2013).
Statistics	mode: 111.00; levels: 4.00; # missings: 110.00; % missings: 23.21

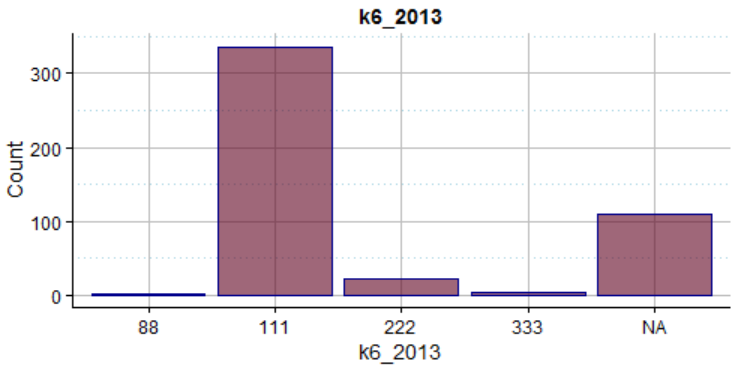
Distribution	
Discretization & Semantic scales	111: Without help (independent, score=1) 222: With some help from another person (independent, score=1) 333: Unable to do it (dependent, score=0)

Table A.194: Description of $lw1\_2013$

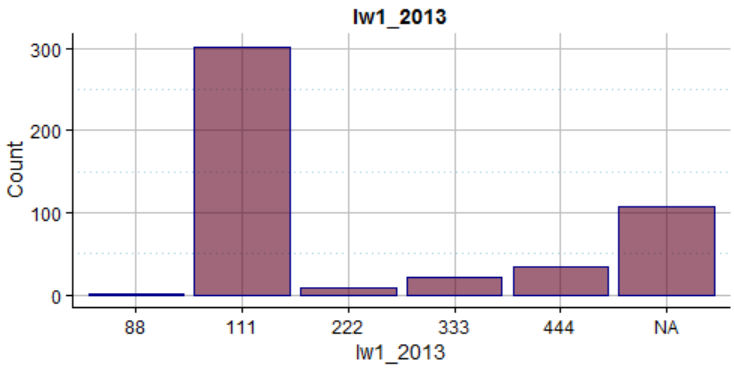
$lw1\_2013$	
Meaning	This feature is associated with the IADL test, and represents question 1.
Statistics	mode: 111.00; levels: 5.00; # missings: 107.00; % missings: 22.57
Distribution	
Discretization & Semantic scales	111: Operates telephone on own initiative(independent, score=1); looks up and dials numbers, etc. 222: Dials a few well-known numbers (independent, score=1) 333: Answers telephone but does not dial (independent, score=1) 444: Does not use telephone at all (dependent, score=0)

Table A.195: Description of *lw*2_2013

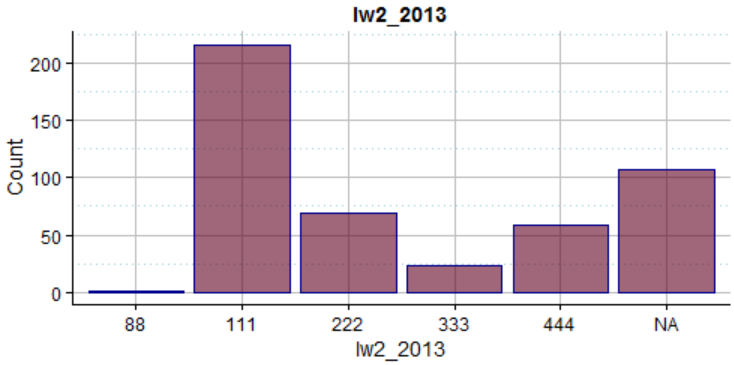
*lw*2_2013	
Meaning	This feature is associated with the IADL test, and represents question 2.
Statistics	mode: 111.00; levels: 5.00; # missings: 107.00; % missings: 22.57
Distribution	
Discretization & Semantic scales	111: Takes care of all shopping needs independently (independent, score=1) 222: Shops independently for small purchases (dependent, score=0) 333: Needs to be accompanied on any shopping trip (dependent, score=0) 444: Completely unable to shop (dependent, score=0)

Table A.196: Description of *lw*3_2013

*lw*3_2013	
Meaning	This feature is associated with the IADL test, and represents question 3.
Statistics	mode: 111.00; levels: 5.00; # missings: 107.00; % missings: 22.57

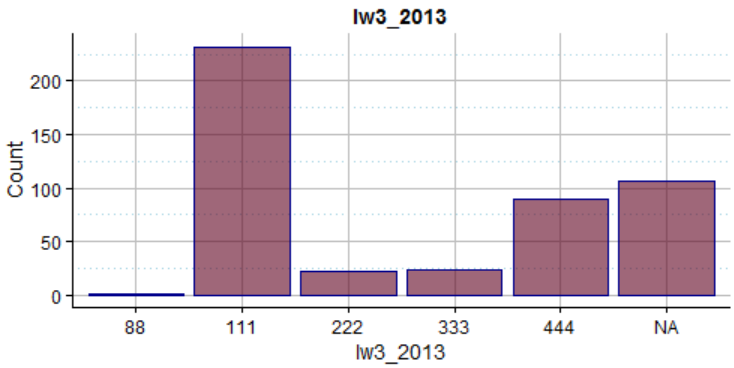
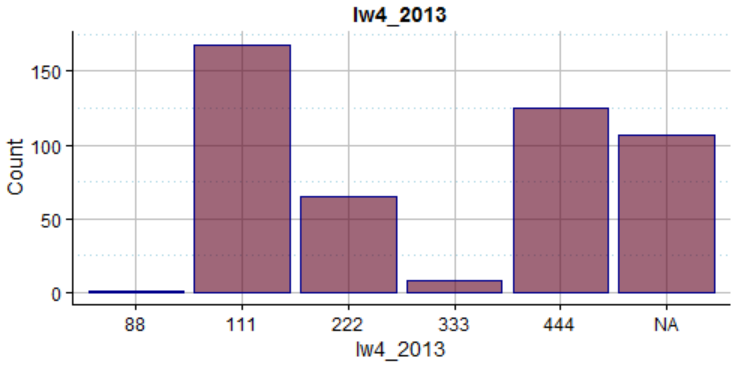
Distribution	
Discretization & Semantic scales	111: Plans, prepares, and serves adequate meals independently(independent, score=1) 222: Prepares adequate meals if supplied with ingredients (dependent, score=0) 333: Heats and serves prepared meals, or prepares meals but does not maintain adequate diet (dependent, score=0) 444: Needs to have meals prepared and served (dependent, score=0)

Table A.197: Description of *lw*4_2013

*lw*4_2013				
Meaning	This feature is associated with the IADL test, and represents question 4.			
Statistics	mode	levels	# missings	% missings
	111.00	5.00	107.00	22.57
Distribution				

Discretization & Semantic scales	111: Maintains house alone or with occasional assistance (e.g., "heavy work domestic help")(independent, score=1) 222: Performs light daily tasks such as dishwashing, bed making(independent, score=1) 333: Performs light daily tasks but cannot maintain acceptable level of cleanliness (dependent, score=0) 444: Needs help with all home maintenance tasks (dependent, score=0) 555: Does not participate in any housekeeping tasks (dependent, score=0)

Table A.198: Description of *lw*5_2013

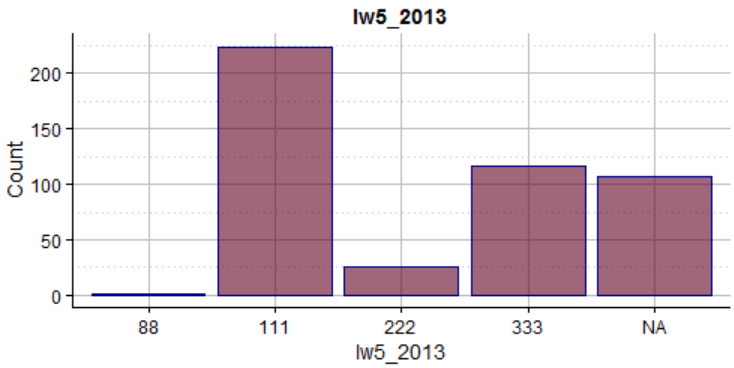
*lw*5_2013	
Meaning	This feature is associated with the IADL test, and represents question 5.
Statistics	mode: 111.00; levels: 4.00; # missings: 107.00; % missings: 22.57
Distribution	
Discretization & Semantic scales	111: Does personal laundry completely (independent, score=1) 222: Launders small items; rinses stockings, etc. (dependent, score=0)) 333: All laundry must be done by others (dependent, score=0)

Table A.199: Description of *lw6_2013*

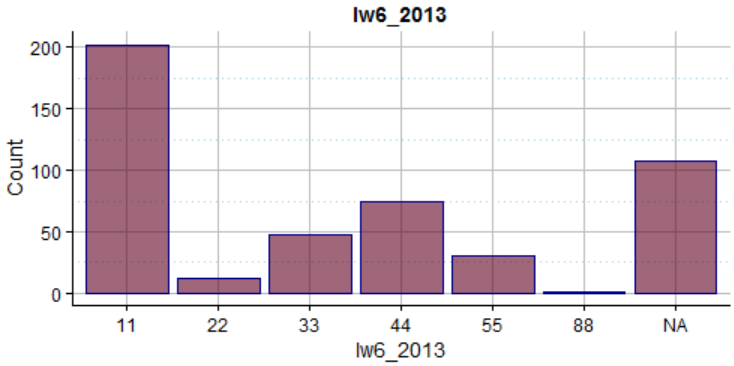
lw6_2013	
Meaning	This feature is associated with the IADL test, and represents question 6.
Statistics	mode: 11.00; levels: 6.00; # missings: 108.00; % missings: 22.78
Distribution	
Discretization & Semantic scales	111: Travels independently on public transportation or drives own car(independent, score=1) 222: Arranges own travel via taxi, but does not otherwise use public transportation (independent, score=1) 333: Travels on public transportation when assisted or accompanied by another (dependent, score=0) 444: Travel limited to taxi or automobile with assistance of another (dependent, score=0) 555: Does not travel at all (dependent, score=0)
Note	The values should be "111,222,333,444,555" instead of "11,22,33,44,55".

Table A.200: Description of *lw7_2013*

lw7_2013	
Meaning	This feature is associated with the IADL test, and represents question 7.

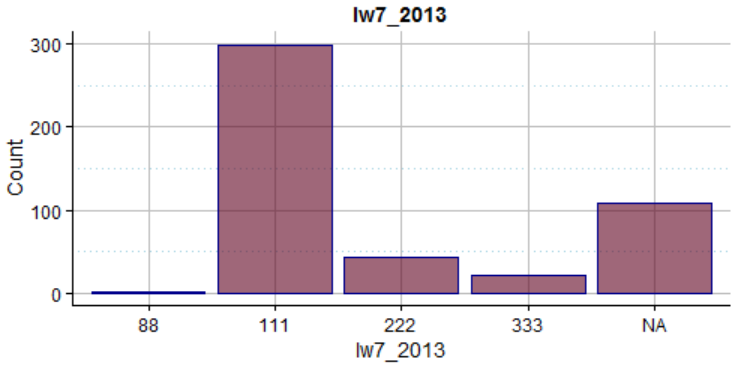
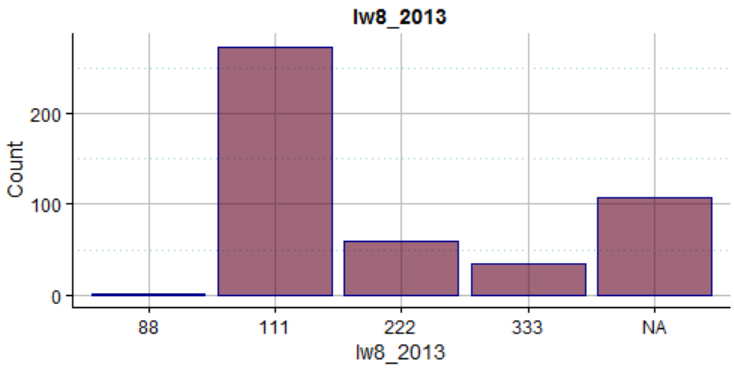
Statistics	mode	levels	# missings	% missings
	111.00	4.00	109.00	23.00
Distribution				
Discretization & Semantic scales	111: Is responsible for taking medication in correct dosages at correct time(independent, score=1) 222: Takes responsibility if medication is prepared in advance in separate dosages (dependent, score=0) 333: Is not capable of dispensing own medication (dependent, score=0)			

Table A.201: Description of *lw*8_2013

*lw*8_2013				
Meaning	This feature is associated with the IADL test, and represents question 8.			
Statistics	mode	levels	# missings	% missings
	111.00	4.00	107.00	22.57
Distribution				

Discretization & Semantic scales	111: Manages financial matters independently (budgets, writes checks, pays rent and bills, goes to bank), collects and keeps track of income (independent, score=1) 222: Manages day-to-day purchases, but needs help with banking, major purchases, etc: (independent, score=1) 333: Incapable of handling money (dependent, score=0)

Table A.202: Description of *cq*8

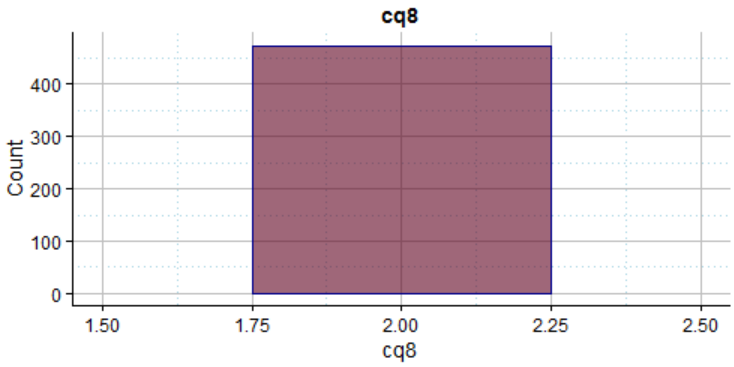
*cq*8	
Meaning	This feature gives binary information about the presence of Leukemia or Polycytemia.
Statistics	mode: 2.00; levels: 1.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	1: present 2: not present 88: not available

Table A.203: Description of *cq*9

*cq*9	
Meaning	This feature gives binary information about the presence of Lymphoma.

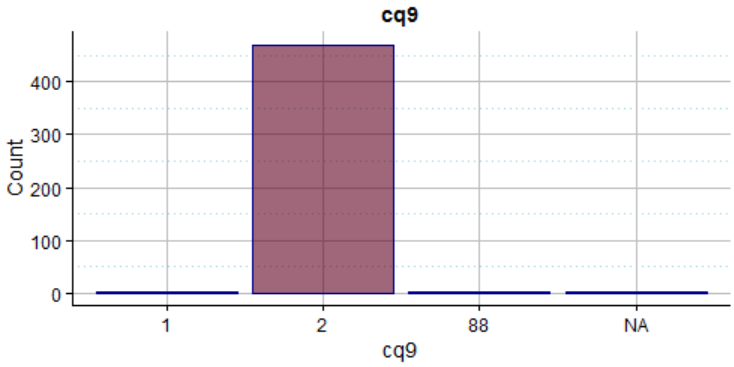
Statistics	mode	levels	# missings	% missings
	2.00	3.00	2.00	0.42
Distribution				
Discretization & Semantic scales	1: present 2: not present 88: not available			

Table A.204: Description of *cq*10

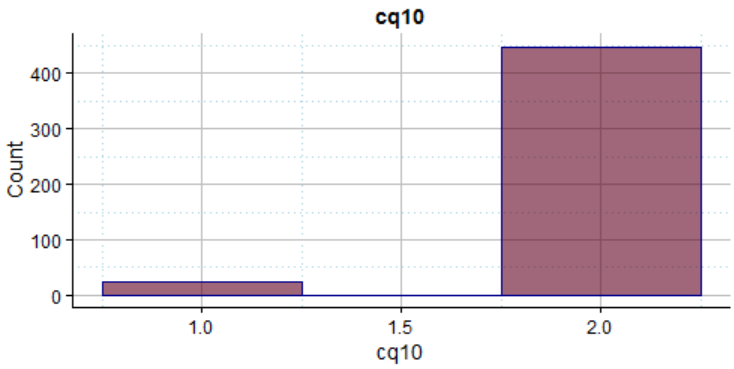
*cq*10				
Meaning	This feature gives binary information about the presence of cancer (except leukemia, polycythemia). and lymphoma)			
Statistics	mode	levels	# missings	% missings
	2.00	2.00	1.00	0.21
Distribution				
Discretization & Semantic scales	1: present 2: not present 88: not available			

Table A.205: Description of *ppeso_2013*

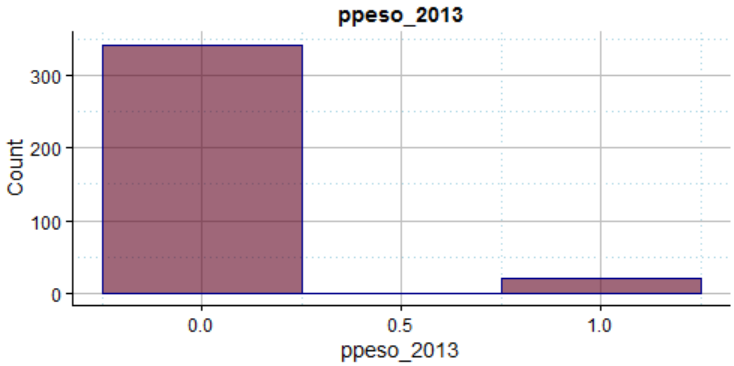
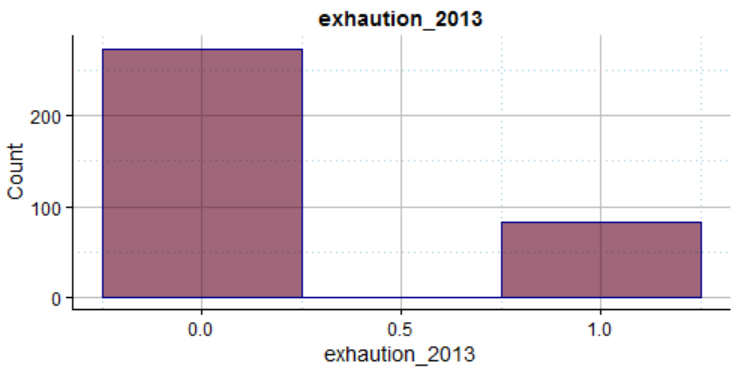
ppeso_2013	
Meaning	This feature gives binary information about the Fried criterion: "weight loss >10 lbs. in past year".
Statistics	mode: 0.00; levels: 2.00; # missings: 112.00; % missings: 23.63
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.206: Description of *exhaution_2013*

exhaution_2013	
Meaning	This feature gives binary information about the Fried criterion: "exhaustion >=3days in past week".
Statistics	mode: 0.00; levels: 2.00; # missings: 118.00; % missings: 24.89
Distribution	

Discretization & Semantic scales	0: not true 1: true

Table A.207: Description of $pasefrag\_2013$

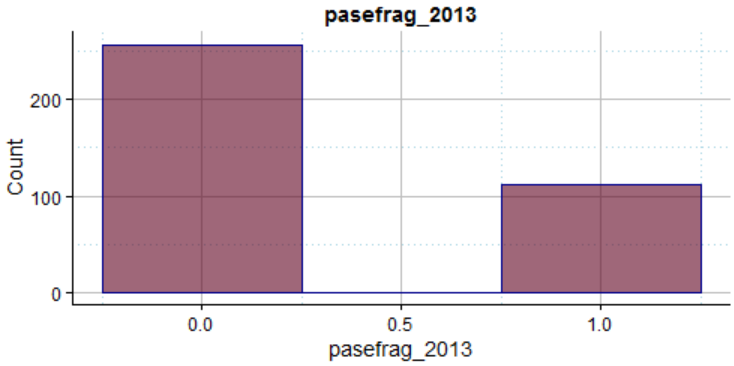
$pasefrag\_2013$	
Meaning	This feature gives binary information about the Fried criterion: "pase score $\leqslant 20^{\text{th}}$ percentile"
Statistics	mode: 0.00; levels: 2.00; # missings: 105.00; % missings: 22.15
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.208: Description of $fuerzafragil\_2013$

$fuerzafragil\_2013$	
Meaning	This feature gives binary information about the Fried criterion: "grip strength $\leqslant 20^{\text{th}}$ percentile".
Statistics	mode: 0.00; levels: 2.00; # missings: 120.00; % missings: 25.32

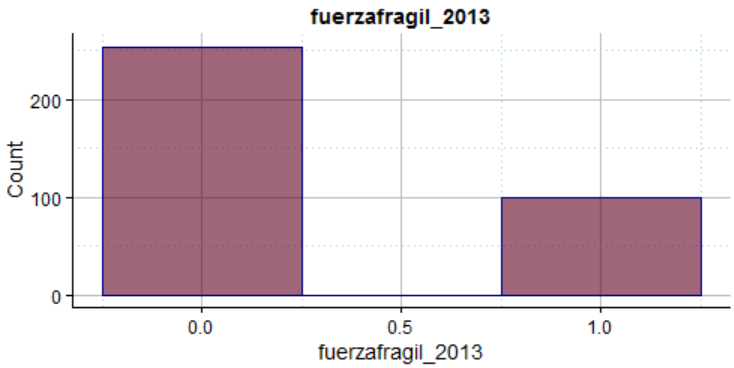
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.209: Description of $marchafragil\_2013$

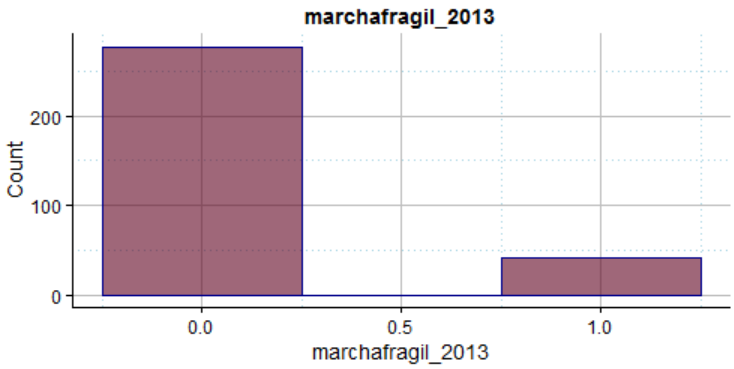
$marchafragil\_2013$	
Meaning	This feature gives binary information about the Fried criterion: "time to walk $\geqslant 80^{\text{th}}$ percentile".
Statistics	mode: 0.00; levels: 2.00; # missings: 156.00; % missings: 32.91
Distribution	
Discretization & Semantic scales	0: not true 1: true

Table A.210: Description of num_drug

num_drug	
Meaning	This feature gives numeric information about the number of drugs the patient takes.

Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	0.00	11.00	4.66	4.00	8.65	2.94	0.00	0.00

Distribution

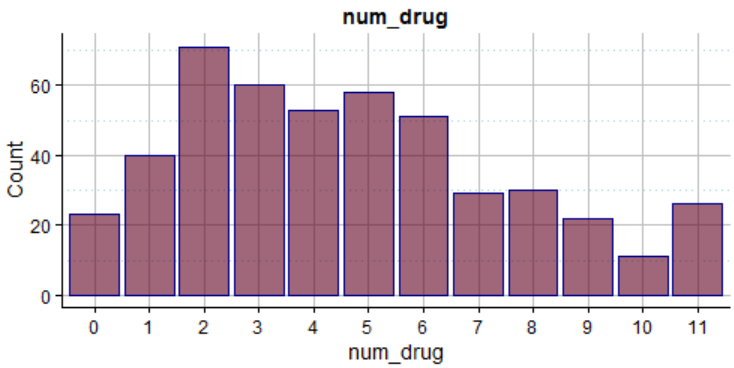


Table A.211: Description of *polypharmacy*

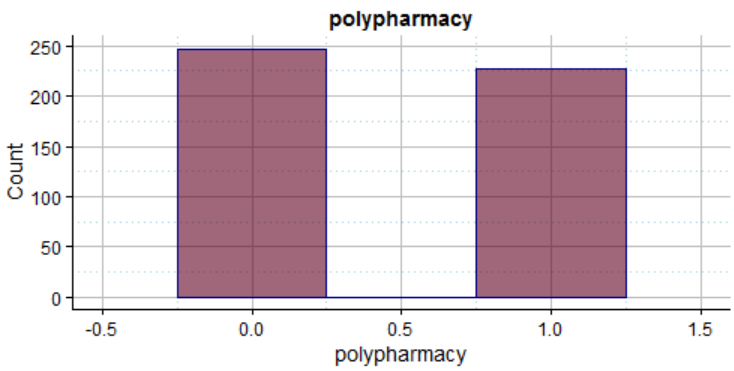
polypharmacy	
Meaning	This feature gives binary information about the presence of polypharmacy (when the number of drugs is equal or higher 5).
Statistics	mode: 0.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	0: no 1: yes

Table A.212: Description of $cognitive_impairment_MMSE_educative_level$

$cognitive_impairment_MMSE_educative_level$

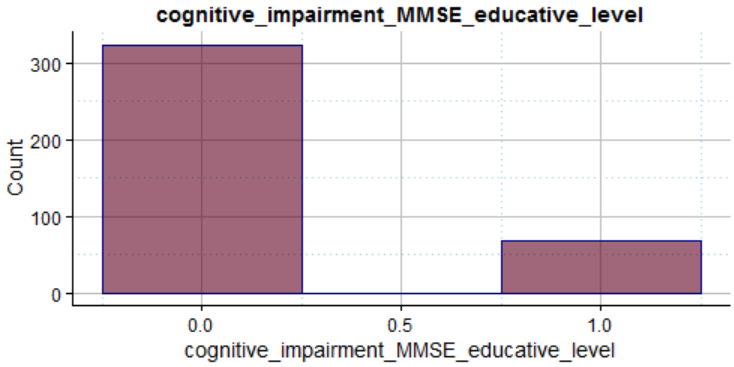
Meaning	This feature gives binary information about the question "Has the patient a cognitive impairment?".
Statistics	mode: 0.00; levels: 2.00; # missings: 81.00; % missings: 17.09
Distribution	
Discretization & Semantic scales	0: no 1: yes

Table A.213: Description of $X.1\_year\_smoker$

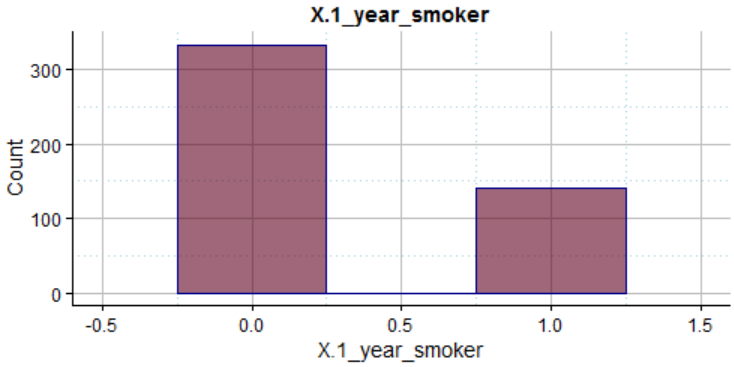
$X.1\_year\_smoker$	
Meaning	This feature gives binary information about if the patient was a smoker for at least one year.
Statistics	mode: 0.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	0: no 1: yes

Table A.214: Description of $current_smoker$

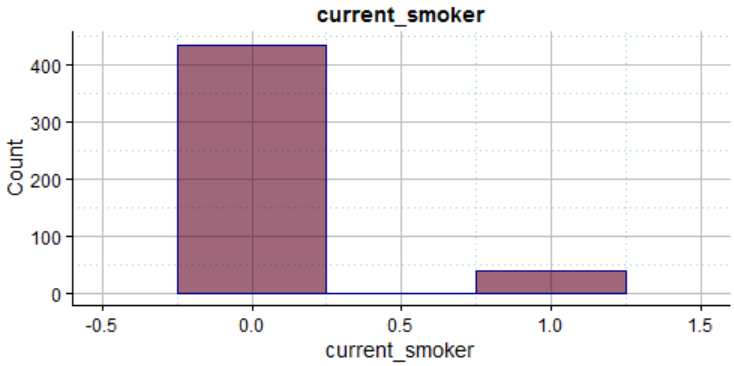
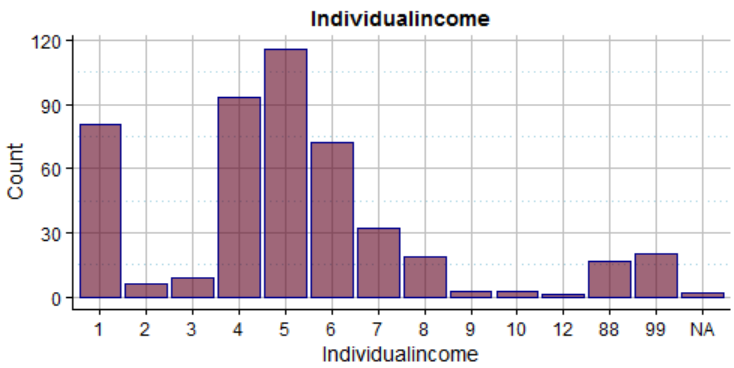
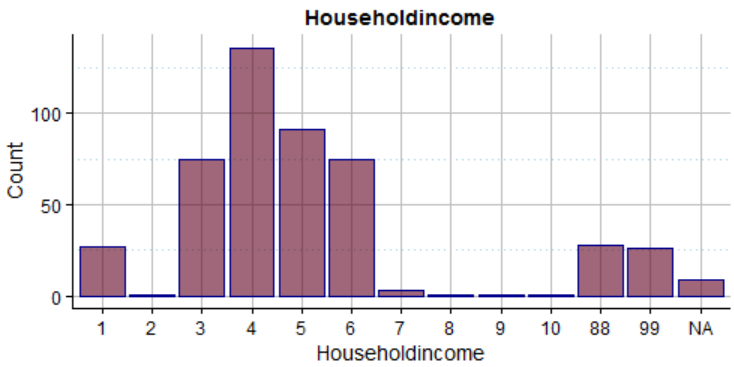
$current_smoker$	
Meaning	This feature gives binary information about if the patient is currently a smoker.
Statistics	mode: 0.00; levels: 2.00; # missings: 0.00; % missings: 0.00
Distribution	
Discretization & Semantic scales	0: no 1: yes

Table A.215: Description of $Individualincome$

$Individualincome$	
Meaning	This feature gives categorical information about income of the individual.
Statistics	mode: 5.00; levels: 13.00; nMissings: 2.00; nMissingsPerc: 0.42
Distribution	

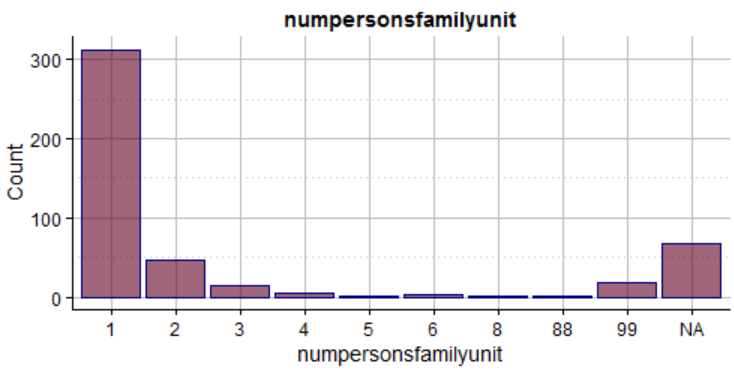
Discretization & Semantic scales	1:None 2:<300 euros 3: 301-500 euros 4:501-700 euros 5:701-900 euros 6:901-1:500 euros 7:1:501- 2:000 euros 8:2001- 3:000 euros 9:3:001-4000 euros 10:más de 4001 euros: 11:NS 12:NC

Table A.216: Description of *Householdincome*

Householdincome	
Meaning	This feature gives categorical information about the income of the household in which the individual lives.
Statistics	mode: 4.00; levels: 12.00; nMissings: 9.00; nMissingsPerc: 1.90
Distribution	

Discretization & Semantic scales	1:None 2:< 200 euros 3:201-300 euros 4:301-500 euros 5:501-700 euros 6:701-900 euros 7:901-1:100 euros 8:1:101-1:300 euros 9:1301-1:500 euros 10:1501-2000 euros 11:2001-3000 euros 12:3001-4000 euros 13:more than 4001 euros: 14:NS 15:NC

Table A.217: Description of *numpersonsfamilyunit*

numpersonsfamilyunit	
Meaning	This feature gives binary information about ...
Statistics	mode: 1.00; levels: 9.00; nMissings: 68.00; nMissingsPerc: 14.35
Distribution	

Discretization & Semantic scales	1: 1 person 2: 2 persons 3: 3 persons 4: 4 persons 5: 5 persons 6: 6persons 7: 7persons 8: 8 or more persons 9: don't know 10: No answer

Table A.218: Description of *Charlsonindex*

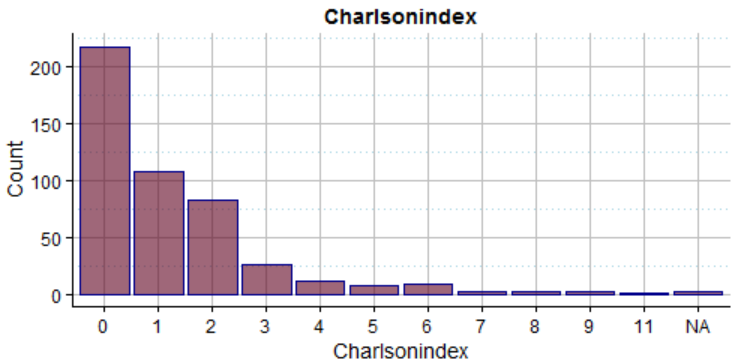
Charlsonindex	
Meaning	This feature gives numeric information about the Charlson co-morbidity index score.
Statistics	min: 0.00; max: 11.00; average: 1.19; median: 1.00; σ^2: 2.79; σ: 1.67; # missings: 3.00; % missings: 0.63
Distribution	

Table A.219: Description of *cv1cv4*

cv1cv4	
Meaning	This feature gives binary information about the presence of myocardial infarction or Heart attack (self reported) or angina pectoris.

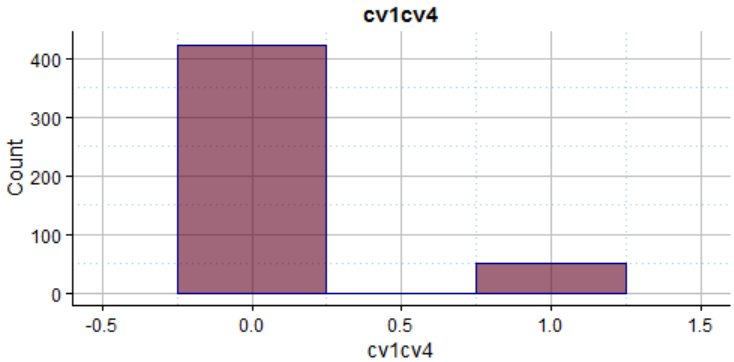
Statistics	mode	levels	# missings	% missings
	0.00	2.00	0.00	0.00
Distribution				
Discretization & Semantic scales	0: no 1: yes			

Table A.220: Description of $IGF1$

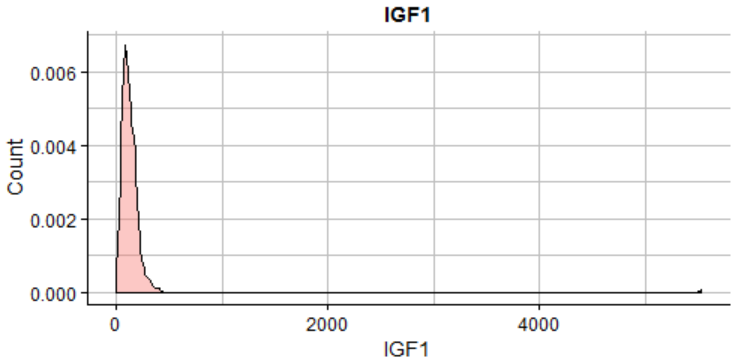
$IGF1$								
Meaning	This feature gives numeric information about the insulin like growth factor 1 (IGF1) [ng/mL].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	0.00	5530.00	135.53	105.00	88748.49	297.91	127.00	26.79
Distribution								

Table A.221: Description of $E2$

$E2$	
Meaning	This feature gives numeric information about 17β-estradiol (E2) [pmol/L].

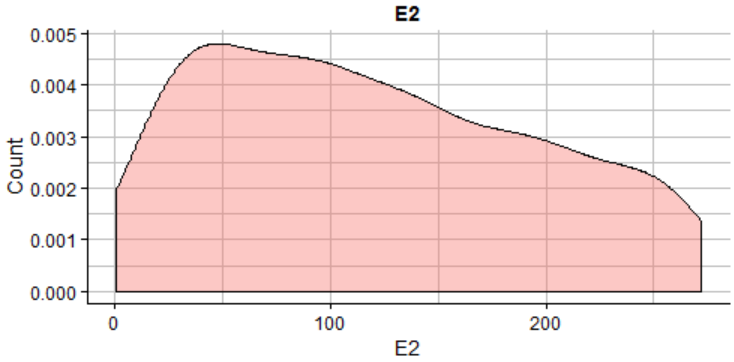
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	1.00	272.00	120.60	110.50	5622.66	74.98	16.00	3.38
Distribution								

Table A.222: Description of *Dheas*

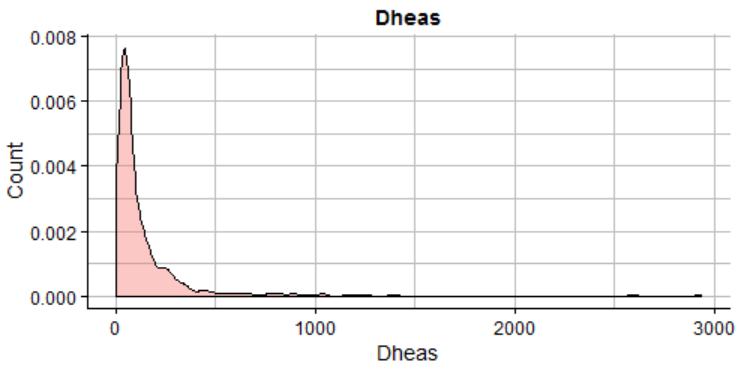
Dheas								
Meaning	This feature gives numeric information about dehydroepiandrosterone sulfate (DHEA-S) [µg/dL].							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	4.00	2936.00	143.85	67.00	66122.32	257.14	16.00	3.38
Distribution								

Table A.223: Description of *Dhea*

Dhea								
Meaning	This feature gives numeric information about dehydroepiandrosterone (DHEA) [ng/mL]							
Statistics	min	max	average	median	σ^2	σ	# missings	% missings
	1.00	130.00	7.67	5.00	86.84	9.32	16.00	3.38

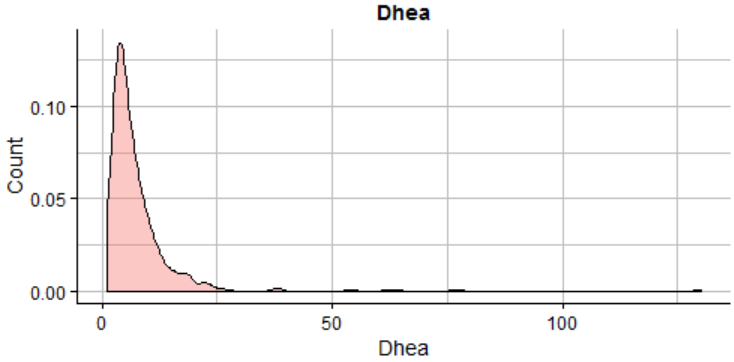
Distribution	

Table A.224: Description of *epoc*1

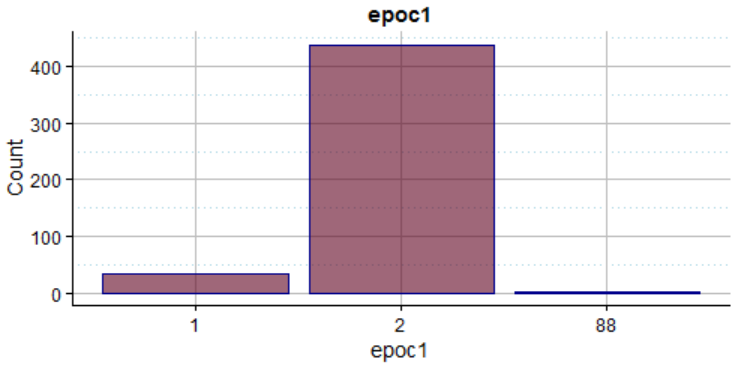
*epoc*1	
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor tell you that you had a chronic obstructive pulmonary disease: emphysema or chronic bronchitis?"
Statistics	mode: 2.00; levels: 2.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.225: Description of *epoc*2

epoc2

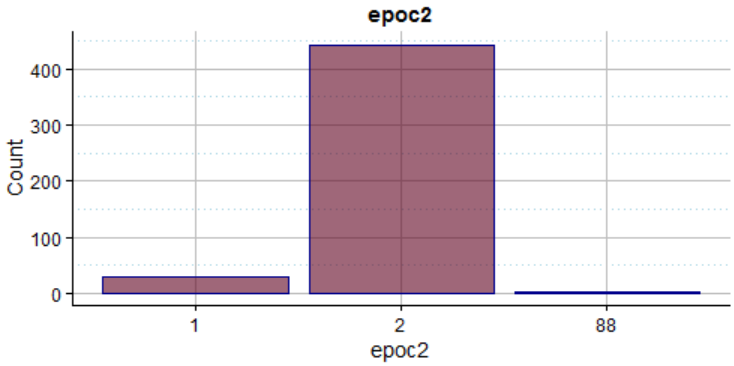
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor say tell that you had asthma?"
Statistics	mode: 2.00; levels: 2.00; # missings: 2.00; % missings: 0.42
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.226: Description of *epoc3*

epoc3	
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor tell you that you had any lung disease?"
Statistics	mode: 2.00; levels: 2.00; # missings: 2.00; % missings: 0.42

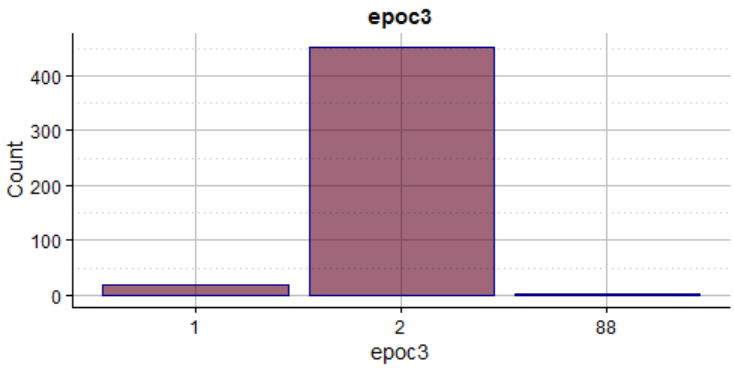
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.227: Description of *epoc4*

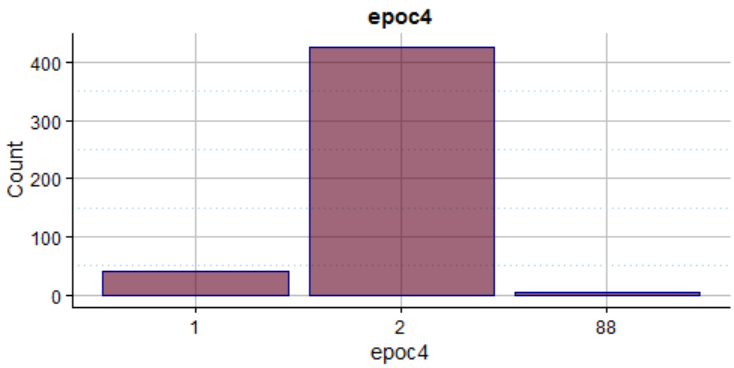
epoc4	
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor tell you that you had had a pneumonía or bronchopneumonía?"
Statistics	mode: 2.00; levels: 2.00; # missings: 5.00; % missings: 1.05
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.228: Description of *epoc*5

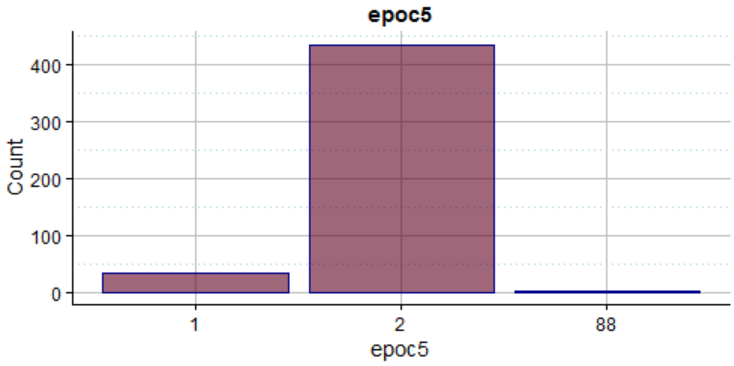
*epoc*5	
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor tell you that you had had an acute bronchitis?"
Statistics	mode: 2.00; levels: 2.00; # missings: 3.00; % missings: 0.63
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.229: Description of *epoc*6

*epoc*6	
Meaning	This feature gives categorical information about the answer to the question: "Have you ever been operated of your lung?"
Statistics	mode: 2.00; levels: 2.00; # missings: 1.00; % missings: 0.21

Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.230: Description of *epoc7*

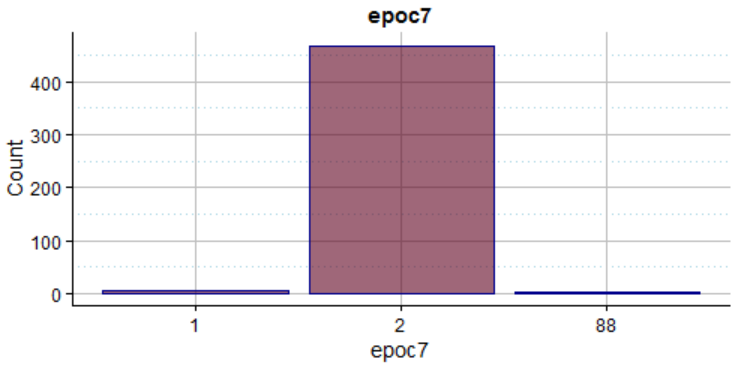
epoc7	
Meaning	This feature gives categorical information about the answer to the question: "Do you have any other lung disease?"
Statistics	mode: 2.00; levels: 2.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.231: Description of *cq6*

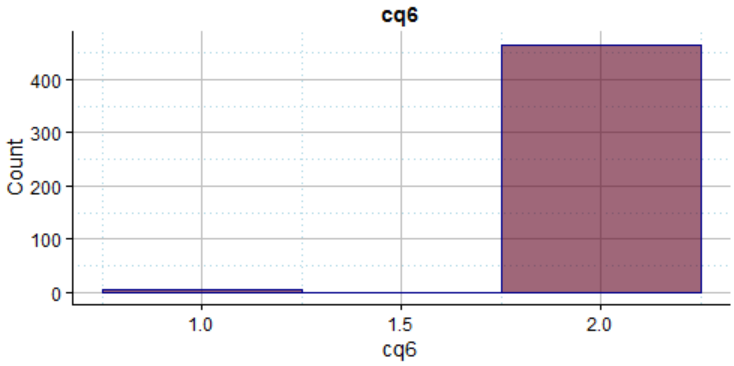
cq6	
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor tell you that you had Alzheimer's disease, senile dementia or another dementia?"
Statistics	mode: 2.00; levels: 2.00; # missings: 3.00; % missings: 0.63
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.232: Description of *cq6a*

cq6a	
Meaning	This feature gives categorical information about the answer to the question: "What kind of dementia did your say doctor that you had?"
Statistics	mode: 1.00; levels: 1.00; # missings: 470.00; % missings: 99.16

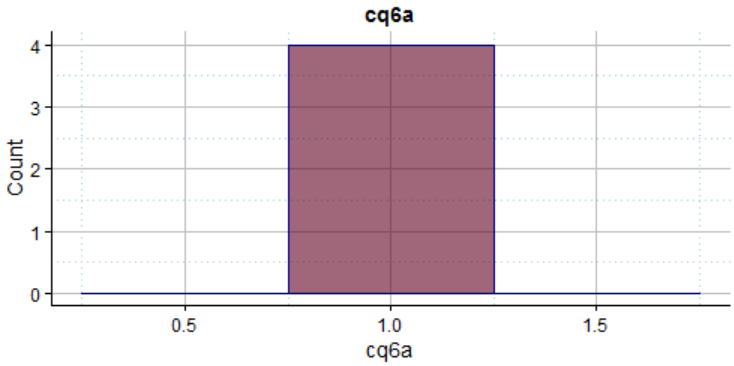
Distribution	
Discretization & Semantic scales	1: Alzheimer's disease 2: Vascular dementia 3: Mixed dementia 4: Dementia with Lewy bodies 5: Frontotemporal dementia 6: Dementia associated to Parkinson's disease 7: Senile dementia 8: Other dementia 9: Don't Know 10: No answer

Table A.233: Description of $reum1$

$reum1$	
Meaning	This feature gives categorical information about the answer to the question: "Have you ever had any joint inflammated for more than 4 weeks in a row?"
Statistics	mode: 2.00; levels: 3.00; # missings: 1.00; % missings: 0.21

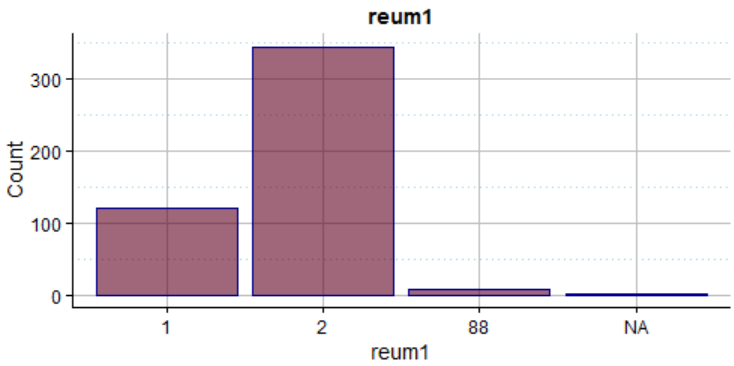
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.234: Description of *reum2*

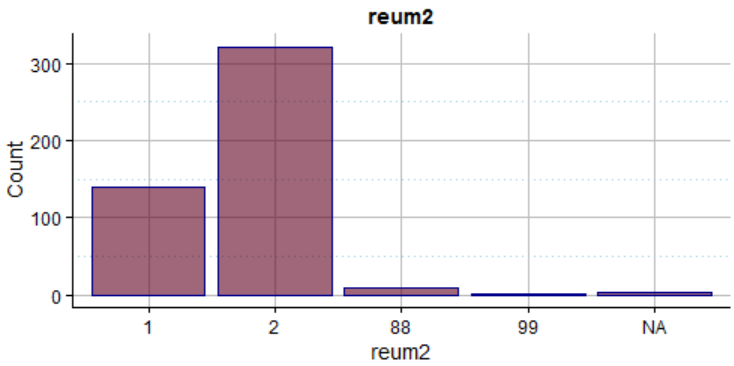
reum2	
Meaning	This feature gives categorical information about the answer to the question: "Have you ever felt pain in any joint for more than 4 weeks in a row?"
Statistics	mode: 2.00; levels: 4.00; # missings: 3.00; % missings: 0.63
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.235: Description of *reum3*

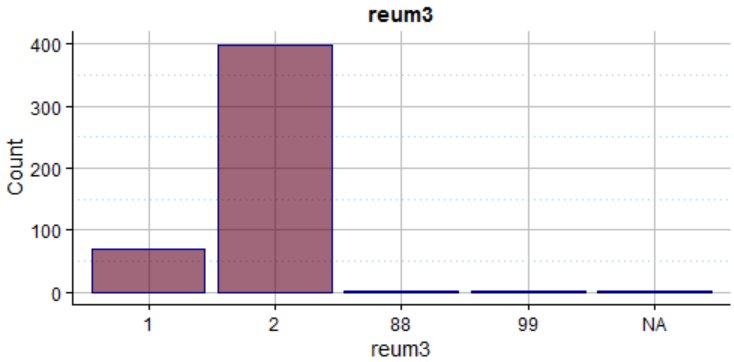
reum3	
Meaning	This feature gives categorical information about the answer to the question: "Do you ever feel that you can't move or feel rigid for over half an hour during the morning?"
Statistics	mode: 2.00; levels: 4.00; # missings: 1.00; % missings: 0.21
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.236: Description of *reum4*

reum4	
Meaning	This feature gives categorical information about the answer to the question: "Have you ever been told you have arthritis?"
Statistics	mode: 2.00; levels: 3.00; # missings: 1.00; % missings: 0.21

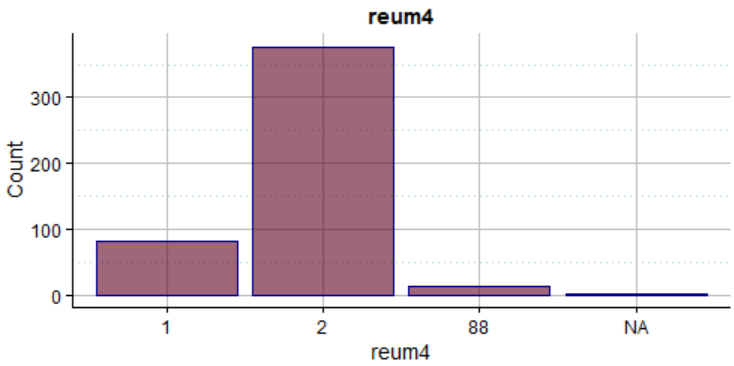
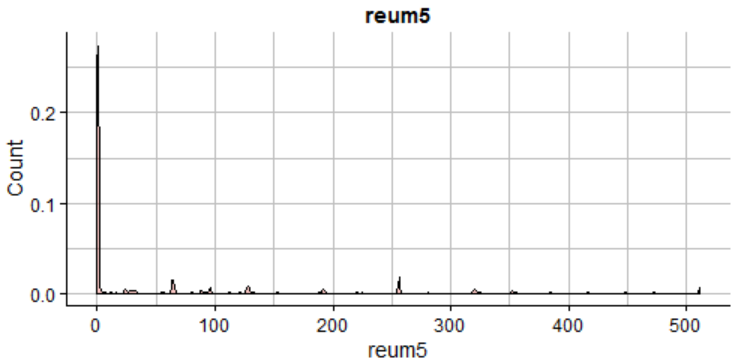
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.237: Description of *reum5*

reum5	
Meaning	This feature gives categorical information about the answer to the question: "Please select in the mannequin the joints in which you have had or have now inflammation for more than 4 weeks in a row (note the location of the affected joints). SHOW CARD 2."
Statistics	min: 0.00; max: 511.00; average: 43.62; median: 0.00; σ^2: 10711.39; σ: 103.50; # missings: 0.00; % missings: 0.00
Distribution	

Discretization & Semantic scales	1: Shoulders 2: Elbows 3: Wrists 4: Metacarpophalangeal 5: Proximal interphalangeal 6: Hips 7: Knees 8: Ankles 9: Others

Table A.238: Description of *reum6*

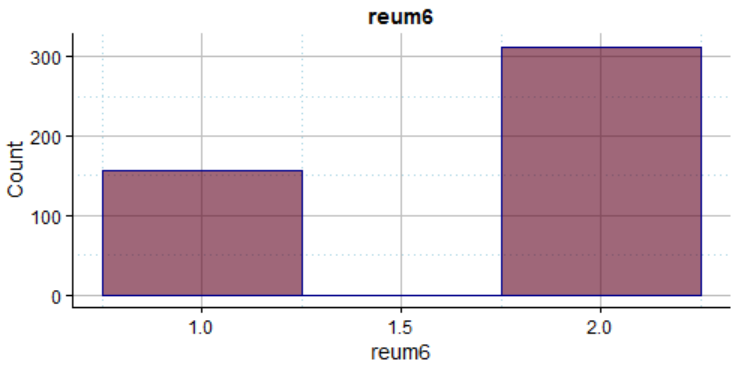
reum6	
Meaning	This feature gives categorical information about the answer to the question: "Do you feel pain or have inflammation in any joint?"
Statistics	mode: 2.00; levels: 2.00; # missings: 5.00; % missings: 1.05
Distribution	
Discretization & Semantic scales	1: Yes 2: No 3: Don't know 4: No answer

Table A.239: Description of *reum7*

reum7

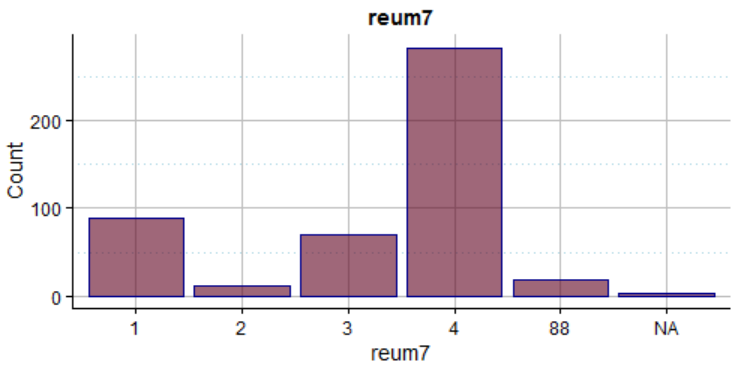
Meaning	This feature gives categorical information about the answer to the question: "Did any doctor tell you that you had arthritis or arthrosis in your..?"
Statistics	mode: 4.00; levels: 5.00; # missings: 3.00; % missings: 0.63
Distribution	
Discretization & Semantic scales	1: Knees 2: Hips 3: Knees and hips 4: Others 5: Don't Know 6: Don't answer

Table A.240: Description of *drug_1a*

drug_1a	
Meaning	This feature gives categorical information about the drug related question : "How do you take it?"
Statistics	mode: 1.00; levels: 3.00; # missings: 31.00; % missings: 6.54

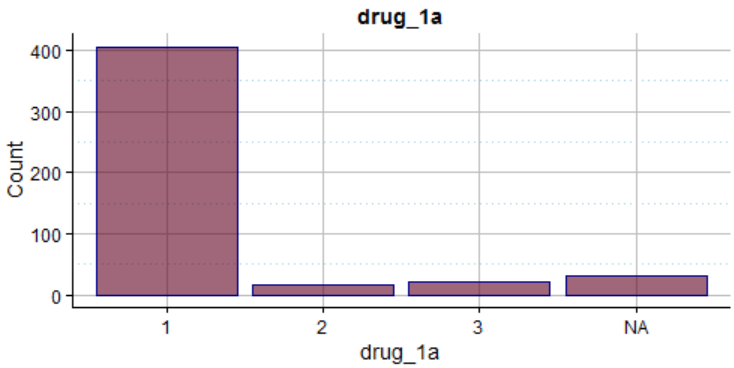
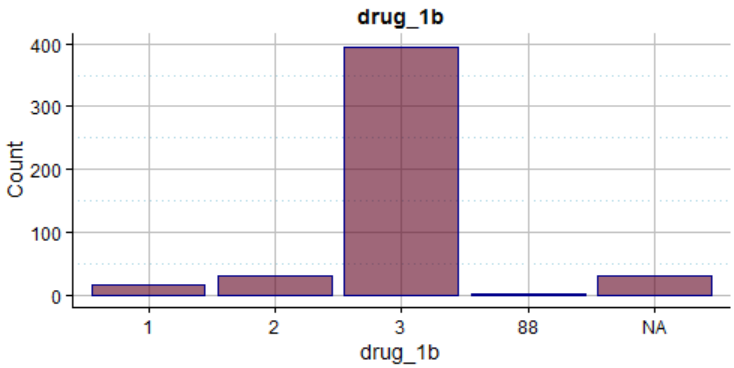
Distribution	
Discretization & Semantic scales	1: continuous 2: intermittent 3: Sporadic other: not available

Table A.241: Description of *drug_1b*

drug_1b	
Meaning	This feature gives categorical information about the drug related question : "When did you start to take it?"
Statistics	mode: 3.00; levels: 4.00; # missings: 30.00; % missings: 6.33
Distribution	
Discretization & Semantic scales	1: less than 1 month 2: from 1 month to 1 year 3: more than 1 year other: NAN

A.2 Codebook

	VARIABLE	LABEL	TYPE	DESCRIPTION	LEVELS OR NOTES
1	ETES_CODE	hi1	Numeric	ETES ID	
2	FRAILOMIC_CODE	FRAILOMIC_CODE	Numeric	FRAILOMIC ID	
3	AGE	hi8	Numeric	Age (in years)	
4	GENDER	hi11	Cathegorical		1.Male; 2.Female
5	FRAIL_1	ppeso	Cathegorical	Frailty: weight loss >10 lbs. in past yr	0.No; 1.Yes; missing.Undetermined
6	FRAIL_2	exhaustion	Cathegorical	Frailty: exhaustion >=3days in past week	0.No; 1.Yes;missing.Undetermined
7	FRAIL_3	pasefrag	Cathegorical	Frailty: PASE <=20 percentile	0.No; 1.Yes; missing.Undetermined
8	FRAIL_4	marchafragil	Cathegorical	Frailty: time to walk >=80th percentile	0.No; 1.Yes;;missing.Undetermined
9	FRAIL_5	fuerzafragil	Cathegorical	Frailty: grip strength <=20th percentile	0.No; 1.Yes;missing.Undetermined
10	FRAILTY STATUS	Fragil	Cathegorical	Frail status according to Fried scale	0.Health; 1.Prefrail; 2. Frail; missing.Undetermined
11	MMSE	MMSE2009	Numeric	MMSE raw score (0-30)	Score 0-30
12	GDS_1	YS1	Cathegorical	GDS1:Are you basically satisfied with your life?	1.YES (score 0); 2.NO (score 1)
13	GDS_2	YS2	Cathegorical	GDS2:Have you dropped many of your activities and interests?	1.YES (score 1); 2.NO (score 0)
14	GDS_3	YS3	Cathegorical	GDS3:Do you feel that your life is empty?	1.YES (score 1); 2.NO (score 0)
15	GDS_4	YS4	Cathegorical	GDS4:Do you often get bored?	1.YES (score 1); 2.NO (score 0)
16	GDS_5	YS5	Cathegorical	GDS5:Are you in good spirits most of the time?	1.YES (score 0); 2.NO (score 1)
17	GDS_6	YS6	Cathegorical	GDS6:Are you afraid that something bad is going to happen to you?	1.YES (score 1); 2.NO (score 0)
18	GDS_7	YS7	Cathegorical	GDS7:Do you feel happy most of the time?	1.YES (score 0); 2.NO (score 1)
19	GDS_8	YS8	Cathegorical	GDS8:Do you often feel helpless?	1.YES (score 1); 2.NO (score 0)
20	GDS_9	YS9	Cathegorical	GDS9:Do you prefer to stay at home, rather than going out and doing new things?	1.YES (score 1); 2.NO (score 0)
21	GDS_10	YS10	Cathegorical	GDS10:Do you feel you have more problems with memory than most?	1.YES (score 1); 2.NO (score 0)
22	GDS_11	YS11	Cathegorical	GDS11:Do you think it is wonderful to be alive now?	1.YES (score 0); 2.NO (score 1)
23	GDS_12	YS12	Cathegorical	GDS12:Do you feel pretty worthless the way you are now?	1.YES (score 1); 2.NO (score 0)
24	GDS_13	YS13	Cathegorical	GDS13:Do you feel full of energy?	1.YES (score 0); 2.NO (score 1)
25	GDS_14	YS14	Cathegorical	GDS14:Do you feel that your situation is hopeless?	1.YES (score 1); 2.NO (score 0)
26	GDS_15	YS15	Cathegorical	GDS15:Do you think that most people are better off than you are?	1.YES (score 1); 2.NO (score 0)
27	GDS	gdstotal	Numeric	GDS: Total Score	Score 0-15
28	ADL_1	K1	Cathegorical	WHO activity 6: Any difficulty washing face and arms?	111.Without help (independent, score=1); 222.With some help from another person (independent, score=1); 333.Unable to do it (dependent, score=0)
29	ADL_2	K2	Cathegorical	WHO activity 8: Any difficulty dressing and undressing?	111.Without help (independent, score=1); 222.With some help from another person (independent, score=1); 333.Unable to do it (dependent, score=0)
30	ADL_3	K3	Cathegorical	WHO activity 11: Any difficulty using the toilet?	111.Without help (independent, score=1); 222.With some help from another person (dependent, score=0); 333.Unable to do it (dependent, score=0)
31	ADL_4	K4	Cathegorical	WHO activity 12: Any difficulty getting in and out of bed?	111.Without help (independent, score=1); 222.With some help from another person (dependent, score=0); 333.Unable to do it (dependent, score=0)
32	ADL_5	K5	Cathegorical	WHO activity 19: Any difficulty controlling urination and bowel movements?	111.Without help (independent, score=1); 222.With some help from another person (dependent, score=0); 333.Unable to do it (dependent, score=0)
33	ADL_6	K6	Cathegorical	WHO activity 9: Any difficulty eating (e.g.,holding a fork, cutting food, drinking from a glass)?	111.Without help (independent, score=1); 222.With some help from another person (independent, score=1); 333.Unable to do it (dependent, score=0)
34	ADL	KATZ2008	Numeric	Number of ADL abilities (0-6)	Score (0-6)
35	IADL_1	LW1	Cathegorical	WHO activity 20: Any difficulty using the telephone?	111. Operates telephone on own initiative(independent, score=1); looks up and dials numbers, etc. 222. Dials a few well-known numbers (independent, score=1) 333. Answers telephone but does not dial (independent, score=1) 444. Does not use telephone at all (dependent, score=0)
36	IADL_2	LW2	Cathegorical	WHO activity 5: Any difficulty shopping daily for basic necessities?	111. Takes care of all shopping needs independently (independent, score=1); 222. Shops independently for small purchases (dependent, score=0) 333. Needs to be accompanied on any shopping trip (dependent, score=0) 444. Completely unable to shop (dependent, score=0)
37	IADL_3	LW3	Cathegorical	WHO activity 10: Any difficulty cooking a simple meal?	111. Plans, prepares, and serves adequate meals independently(independent, score=1); 222. Prepares adequate meals if supplied with ingredients (dependent, score=0) 333. Heats and serves prepared meals, or prepares meals but does not maintain adequate diet (dependent, score=0) 444. Needs to have meals prepared and served (dependent, score=0)
38	IADL_4	LW4	Cathegorical	WHO activity 13: Any difficulty doing light housework (e.g., doing dishes, light cleaning)?	111. Maintains house alone or with occasional assistance (e.g., "heavy work domestic help")(independent, score=1); 222. Performs light daily tasks such as dishwashing, bed making(independent, score=1); 333. Performs light daily tasks but cannot maintain acceptable level of cleanliness (dependent, score=0) 444. Needs help with all home maintenance tasks (dependent, score=0) 555. Does not participate in any housekeeping tasks (dependent, score=0)
39	IADL_5	LW5	Cathegorical	WHO activity 14: Any difficulty doing heavy housework (e.g., washing windows, floor)?	111. Does personal laundry completely (independent, score=1); 222. Launders small items; rinses stockings, etc. (dependent, score=0) 333. All laundry must be done by others (dependent, score=0)
40	IADL_6	LW6	Cathegorical	WHO activity 22: Any difficulty using public transportation?	111. Travels independently on public transportation or drives own car(independent, score=1); 222. Arranges own travel via taxi, but does not otherwise use public transportation (independent, score=1); 333. Travels on public transportation when assisted or accompanied by another (dependent, score=0) 444. Travel limited to taxi or automobile with assistance of another (dependent, score=0) 555. Does not travel at all (dependent, score=0)
41	IADL_7	LW7	Cathegorical	WHO activity 23: Any difficulty taking medications correctly?	111. Is responsible for taking medication in correct dosages at correct time(independent, score=1); 222. Takes responsibility if medication is prepared in advance in separate dosages (dependent, score=0) 333. Is not capable of dispensing own medication (dependent, score=0)
42	IADL_8	LW8	Cathegorical	WHO activity 24: Any difficulty managing home finances?	111. Manages financial matters independently (budgets, writes checks, pays rent and bills, goes to bank), collects and keeps track of income (independent, score=1); 222. Manages day-to-day purchases, but needs help with banking, major purchases, etc. (independent, score=1); 333. Incapable of handling money (dependent, score=0);
43	IADL	lawton2008	Numeric	Number of IADL abilities (0-8)	Score (0-8)
44	ALC_CONSUM	Alch1	Cathegorical	How many drinks do you have?	0. Never (in the last year); 1. One or less per month; 2. from 2 to 4 per month; 3. Twice per week; 4. 3 Times per week; 5. 4 Times per week; 6. 5 Times per week; 7. 6 Times per week; 8 Daily
45	WINE	Alch1a1	Numeric	how many glasses of wine do you drink daily?	units of alcohol/day
46	BEER	Alch1a2	Numeric	how many glasses of beer do you drink daily?	units of alcohol/day
47	SPIRITS	Alch1a3	Numeric	how many glasses of spirits do you drink daily?	units of alcohol/day
48	Alcohol consumption. Current period	Alch1b	Cathegorical	For how many years?	given the answer to the question Alch1
49	Previous_alcohol_consumption_1	Alch2	Cathegorical	did you drink previously?	1. YES; 2. NO
50	Previous_alcohol_consumption_2	Alch2a	Cathegorical	Kind of drinker	1. M>12,W>8; 2. M=9-12, W=7-8; 3. M=7-8, W=5-6; 4. M=3-6, W=3-4; 5. M=1-2, W=1-2 (units of alcohol/day)
51	Previous_alcohol_consumption_3	Alch2b	Cathegorical	Starting age	1. <15; 2. 15-20; 3. 21-30; 4. 31-40; 5. 41-50; 6. 51-60; 7. 61-70; 8. 71-80; 8 >80
52	Previous_alcohol_consumption_4	Alch2c	Cathegorical	Ending age	1. <15; 2. 15-20; 3. 21-30; 4. 31-40; 5. 41-50; 6. 51-60; 7. 61-70; 8. 71-80; 8 >80
53	tabacco_consumption_1	tab1	Cathegorical	Have you smoked at least 100 cigarettes in your entire life?	1. YES; 2. NO;3. Unknown; 4 NA
54	tabacco_consumption_2	tab1a	Cathegorical	If yes, Did you smoke cigarettes daily, ocassionally, or not at all?	1. Daily; 2. Ocassionally. 3. Undecided
55	tabacco_consumption_3	tab1a1	Cathegorical	Do you smoke actually?	1. Yes, daily; 2. Yes, ocassionally; 3. No
56	tabacco_consumption_4	tab1a1a	Cathegorical	If not, How many time have you stopped smoking?	1. Yesterday; 2. 2-6 days ago; 3. 7-30 days ago; 4. 1-12 months ago; 5. 1-5 years ago; 6. 6-10 years ago; 7. 11-20 years ago; 8. more than 20 years ago
57	tabacco_consumption_5	tab1a3	Numeric	For how many years did you smoke?	
58	GRIP STRENGHT	fuerza1a	Numeric	Muscle strength (upper) with dynamometer: hand grip dominant limb (kg)	
59	WEIGHT	peso1	Numeric	Weight (kg)	
60	HEIGHT	altura1	Numeric	Height (cm)	
61	SELF_REP_CLINICAL_CONDITION_1	PS1	Cathegorical	How would you evaluate your current health? How do you feel now?	1.Very good; 2.Good;3. Fair (so-so);4.Poor;5.Very poor;6. Undetermined

62	SELF_REP_CLINICAL_CONDITION_2	PS2	Cathegorical	How is your health compared to 1 yr ago?	1.Much better;2.Better;3.The same;4.Slightly worse;5.Much worse;6.Undetermined
63	SELF_REP_CLINICAL_CONDITION_3	PS3	Cathegorical	How would you judge your health compared to other people of your same age?	1.Much worse; 2.Sligthly worse;3.The same; 4.Better; 5. Much better; 6.Undetermined
64	COMORBIDITY_1	ccv1	Cathegorical	Myocardial infarction / Heart attack (self reported)+E148	1.YES; 2. NO
65	COMORBIDITY_2	ccv2	Cathegorical	Congestive heart failure (self reported)	1.YES; 2. NO
66	COMORBIDITY_3	ccv4	Cathegorical	Angina pectoris (self reported)	1.YES; 2. NO
67	COMORBIDITY_4	ccv6	Cathegorical	Hypertension (self-report,drugs,BP tests)	1.YES; 2. NO
68	COMORBIDITY_5	ccv8	Cathegorical	Diabetes mellitus (self reported, drugs)	1.YES; 2. NO
69	EKG	EKG1	Numeric	EKG: Heart rate (beats/minute)	
70	WAIST_PERIMETER	ppci	Numeric	Anthropometry: waist perimeter (cm)	
71	HIP_PERIMETER	ppca	Numeric	Anthropometry: hip perimeter (cm)	
72	PASE_SCORE	pasetot	Numeric	Physical activity scale for elderly score	
73	DEATH	codigo01	Cathegorical	Status at Follow Up 1	0.Alive;1.Dead
74	FRAILTY STATUS 2013	Fragil2013	Cathegorical	Frail status according to Fried scale	0.Health; 1.Prefrail; 2. Frail; missing.Undetermined
75	ADL	Katz2013	Numeric	Number of ADL abilities (0-6)	Score (0-6)
76	IADL	Lawton2013	Numeric	Number of IADL abilities (0-8)	Score (0-8)
77	Tadd	tadd	Numeric	Pressure arterial. Diastolic	
78	Tads	tads	Numeric	Pressure arterial. Systolic	
79	Movility scale 1	em1	Cathegorical	Are you able to walk at home?	1.YES; 2. NO
80	Movility scale 2	em1a	Cathegorical	If answered YES; Do you get tired when doing it?	1.YES; 2. NO
81	Movility scale 3	em1b	Cathegorical	If answered YES; Do you need help when doing it?	1.YES; 2. NO
82	Movility scale 4	em2	Cathegorical	Are you able to go out from home?	1.YES; 2. NO
83	Movility scale 5	em2a	Cathegorical	If answered YES; Do you get tired when doing it?	1.YES; 2. NO
84	Movility scale 6	em2b	Cathegorical	If answered YES; Do you need help when doing it?	1.YES; 2. NO
85	Movility scale 7	em3	Cathegorical	Are you able to climb stairs?	1.YES; 2. NO
86	Movility scale 8	em3a	Cathegorical	If answered YES; Do you get tired when doing it?	1.YES; 2. NO
87	Movility scale 9	em3b	Cathegorical	If answered YES; Do you need help when doing it?	1.YES; 2. NO
88	Movility scale 10	em4	Cathegorical	Are you able to walk outside (nice weather)?	1.YES; 2. NO
89	Movility scale 11	em4a	Cathegorical	If answered YES; Do you get tired when doing it?	1.YES; 2. NO
90	Movility scale 12	em4b	Cathegorical	If answered YES; Do you need help when doing it?	1.YES; 2. NO
91	Movility scale 13	em5	Cathegorical	Are you able to walk outside (bad weather)?	1.YES; 2. NO
92	Movility scale 14	em5a	Cathegorical	If answered YES; Do you get tired when doing it?	1.YES; 2. NO
93	Movility scale 15	em5b	Cathegorical	If answered YES; Do you need help when doing it?	1.YES; 2. NO
94	Educative level	Hi13	Cathegorical	Educative level	1. NONE; 2. UNFINISHED SCHOOL; 3 SCHOOL; 4 SECONDARY SCHOOL; 5 PROFESSIONAL SCHOOL; 6. UNIVERSITY. TECHNICAL GRADE (3 YEARS); 7 UNIVERSITY. GRADE (5 YEARS); 8-10 NAN OR MISSING
95	MMSE temporal domain 1	Enpot1	Cathegorical	What day of the week is this?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
96	MMSE temporal domain 2	Enpot2	Cathegorical	What is today's date?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
97	MMSE temporal domain 3	Enpot3	Cathegorical	What month is this?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
98	MMSE temporal domain 4	Enpot4	Cathegorical	What year is this?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
99	MMSE temporal domain 5	Enpot6	Cathegorical	Which season is this?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
100	MMSE spatial domain 1	Enpol1	Cathegorical	IN HOME: What is the street address of this house? // IN FACILITY: What is the name of this building?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
101	MMSE spatial domain 2	Enpol2	Cathegorical	IN HOME: What room are we in? // IN FACILITY: What floor are we on?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
102	MMSE spatial domain 3	Enpol3	Cathegorical	What city/town are we in?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
103	MMSE spatial domain 4	Enpol4	Cathegorical	What province are we in?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
104	MMSE spatial domain 5	Enpol5	Cathegorical	What county are we in?	0. 0 POINTS; 1. 1 POINT; OTHER. NAN
105	MMSE Three objects. Repetition	enpmem1a	Cathegorical	SAY: I am going to name three objects. When I am finished, I want you to repeat theM. Remember what they are because I am going to ask you to name them again in a few minutes. // Say the following words slowly at 1-second intervals - peseta (coin in spanish), caballo (horse in spanish), manzana (apple in spanish)	1. 0 POINTS; 2. 1 POINT; 3. 2 POINTS; 4. 3 POINTS; OTHER. NAN
106	MMSE spell the word	enpat2	Cathegorical	Spell the word MUNDO (world in spanish). Now spell it backwards.	1. 0 POINTS; 2. 1 POINT; 3. 2 POINTS; 4. 3 POINTS; 5. 4 POINTS; 6. 5 POINTS; OTHER. NAN
107	MMSE backward counting	enpat1	Cathegorical	Count backwards by 7 starting from 100	1. 0 POINTS; 2. 1 POINT; 3. 2 POINTS; 4. 3 POINTS; 5. 4 POINTS; 6. 5 POINTS; OTHER. NAN
	In the total score we used the best result of spelling and backward counting				
108	MMSE Three objects. Short term memory	enpmem2	Cathegorical	Now what were the three objects I asked you to remember?	1. 0 POINTS; 2. 1 POINT; 3. 2 POINTS; 4. 3 POINTS; OTHER. NAN
109	MMSE Wristcatch and pencil	enpleng1	Cathegorical	Show a wristchatch and a pencil. What are these called?	1. 0 POINTS; 2. 1 POINT; 3. 2 POINTS; OTHER. NAN
110	MMSE Phrase. Repetition	enpleng2	Cathegorical	SAY: I would like you to repeat this phrase after me: Ni si, ni no, ni pero. (No ifs, ands or buts. In spanish)	1. 0 POINTS; 2. 1 POINT; OTHER. NAN
111	MMSE Read and comprehension	enpleng4	Cathegorical	SAY: Read the words on the page and then do what it says. Then hand the person the sheet with "Cierre los ojos" (close your eyes in spanish) on it. If the subject read and does not close their eyes, repeat yp to three times. Score only if subject closes eyes.	1. 0 POINTS; 2. 1 POINT; OTHER. NAN
112	MMSE Writing	enpprx1	Cathegorical	Hand the person a pencil and paper. SAY: write any complete sentence on that piece of paper. (Note: The sentence must make sense. Ignore spelling errors)	1. 0 POINTS; 2. 1 POINT; OTHER. NAN
113	MMSE Drawing	enpprx2	Cathegorical	Place design, eraser and pencil in front of the person. SAY: copy this design please. // Allow multiple tries. Wait until person is finished and hands it back. Score only for correctly copied diagram with a 4-sided figure between two 5-sided figures.	1. 0 POINTS; 2. 1 POINT; OTHER. NAN
114	MMSE Listening comprehension	enpleng3	Cathegorical	Ask the person if he is right or left handed. Take a piece of paper and hold it up in front of the person. SAY: Take this paper in your right/left hand (whichever is non-dominant), fold the paper in half once with both hands and put the paper down on the floor. Score 1 point for each instruction executed correctly.	1. 0 POINTS; 2. 1 POINT; 3. 2 POINTS; 4. 3 POINTS; OTHER. NAN
115	ADL_1_2013	K1_2013	Cathegorical	WHO activity 6: Any difficulty washing face and arms?	111.Without help (independent, score=1); 222.With some help from another person (independent, score=1); 333.Unable to do it (dependent, score=0)
116	ADL_2_2013	K2_2013	Cathegorical	WHO activity 8: Any difficulty dressing and undressing?	111.Without help (independent, score=1); 222.With some help from another person (independent, score=1); 333.Unable to do it (dependent, score=0)
117	ADL_3_2013	K3_2013	Cathegorical	WHO activity 11: Any difficulty using the toilet?	111.Without help (independent, score=1); 222.With some help from another person (dependent, score=0); 333.Unable to do it (dependent, score=0)
118	ADL_4_2013	K4_2013	Cathegorical	WHO activity 12: Any difficulty getting in and out of bed?	111.Without help (independent, score=1); 222.With some help from another person (dependent, score=0); 333.Unable to do it (dependent, score=0)
119	ADL_5_2013	K5_2013	Cathegorical	WHO activity 19: Any difficulty controlling urination and bowel movements?	111.Without help (independent, score=1); 222.With some help from another person (dependent, score=0); 333.Unable to do it (dependent, score=0)
120	ADL_6_2013	K6_2013	Cathegorical	WHO activity 9: Any difficulty eating (e.g.,holding a fork, cutting food, drinking from a glass)?	111.Without help (independent, score=1); 222.With some help from another person (independent, score=1); 333.Unable to do it (dependent, score=0)
121	IADL_1_2013	LW1_2013	Cathegorical	WHO activity 20: Any difficulty using the telephone?	111. Operates telephone on own initiative(independent, score=1); looks up and dials numbers, etc. 222. Dials a few well-known numbers (independent, score=1) 333. Answers telephone but does not dial (independent, score=1) 444. Does not use telephone at all (dependent, score=0)
122	IADL_2_2013	LW2_2013	Cathegorical	WHO activity 5: Any difficulty shopping daily for basic necessities?	111. Takes care of all shopping needs independently (independent, score=1); 222. Shops independently for small purchases (dependent, score=0) 333. Needs to be accompanied on any shopping trip (dependent, score=0) 444. Completely unable to shop (dependent, score=0)
123	IADL_3_2013	LW3_2013	Cathegorical	WHO activity 10: Any difficulty cooking a simple meal?	111. Plans, prepares, and serves adequate meals independently(independent, score=1); 222. Prepares adequate meals if supplied with ingredients (dependent, score=0) 333. Heats and serves prepared meals, or prepares meals but does not maintain adequate diet (dependent, score=0) 444. Needs to have meals prepared and served (dependent, score=0)

124	IADL_4_2013	LW4_2013	Cathegorical	WHO activity 13: Any difficulty doing light housework (e.g., doing dishes, light cleaning)?
125	IADL_5_2013	LW5_2013	Cathegorical	WHO activity 14: Any difficulty doing heavy housework (e.g., washing windows, floor)?
126	IADL_6_2013	LW6_2013	Cathegorical	WHO activity 22: Any difficulty using public transportation?
127	IADL_7_2013	LW7_2013	Cathegorical	WHO activity 23: Any difficulty taking medications correctly?
128	IADL_8_2013	LW8_2013	Cathegorical	WHO activity 24: Any difficulty managing home finances?
129	FRAIL_1_2013	ppeso_2013	Cathegorical	Frailty: weight loss >10 lbs. in past yr
130	FRAIL_2_2013	exhaustion_2013	Cathegorical	Frailty: exhaustion >=3days in past week
131	FRAIL_3_2013	pasefrag_2013	Cathegorical	Frailty: PASE <=20 percentile
132	FRAIL_4_2013	marchafragil_2013	Cathegorical	Frailty: time to walk >=80th percentile
133	FRAIL_5_2013	fuerzafragil_2013	Cathegorical	Frailty: grip strength <=20th percentile
134	Leukemia or Polycytemia	CQ8	Cathegorical	
135	Lymphoma	CQ9	Cathegorical	
136	Cancer (except Leukemia, polycythemia an	CQ10	Cathegorical	
137	DEPRESSION	Depression	Cathegorical	gdstotal>=5
138	Dementia	**CQ6**	Cathegorical	Did any doctor tell you that you had Alzheimer's disease, senile dementia or another dementia?
139	Dementia (kind)	CQ6a	Cathegorical	What kind of dementia did your say doctor that you had?
140	Arthosis/arthritis	reum1	Cathegorical	Have you ever had any joint inflammated for more than 4 weeks in a row?
141	Arthosis/arthritis	reum2	Cathegorical	Have you ever felt pain in any joint for more than 4 weeks in a row?
142	Arthosis/arthritis	reum3	Cathegorical	Do you ever feel that you can't move or feel rigid for over half an hour during the morning?
143	Arthosis/arthritis	reum4	Cathegorical	Have you ever been told you have arthritis?
144	Arthosis/arthritis	reum5	Cathegorical	Please select in the mannequin the joints in which you have had or have now inflammation for more than 4 weeks in a row (note the location of the affected joints). SHOW CARD 2.
145	Arthosis/arthritis	reum6	Cathegorical	Do you feel pain or have inflammation in any joint?
146	Arthosis/arthritis	reum6a	Cathegorical	If yes,Please, show which joints. SHOW CARD 2:
147	Arthosis/arthritis	reum7	Cathegorical	Did any doctor tell you that you had arthritis or arthrosis in your..?
148	Arthosis/arthritis	reum7a	Cathegorical	if yes (1, 2 or 3)The doctor said that you had it after a hip or knee radiography, or both?
149	EPOC	EPOC1	Cathegorical	Did any doctor tell you that you had a chronic obstructive pulmonary disease: emphysema or chronic bronchitis?
150	EPOC	EPOC2	Cathegorical	Did any doctor say tell that you had asthma?
151	EPOC	EPOC3	Cathegorical	Did any doctor tell you that you had any lung disease?
152	EPOC	EPOC4	Cathegorical	Did any doctor tell you that you had had a pneumonía or bronchopneumonía?
153	EPOC	EPOC5	Cathegorical	Did any doctor tell you that you had had an acute bronchitis?
154	EPOC	EPOC6	Cathegorical	Have you ever been operated of your lung?
155	EPOC	EPOC7	Cathegorical	Do you have any other lung disease?
156	Income	Individualincome	Cathegorical	
157	Income	Householdincome	Cathegorical	
158	Income	numpersonsfamilyunit	Cathegorical	
159	comorbidities	Charlsonindex	Cathegorical	
160	MYO_HA_AP	cv1cv4	Cathegorical	Myocardial infarction / Heart attack (self reported)/angina pectoris
161	number of drugs	num_drug	Cathegorical	number of drugs
162	polypharmacy	polypharmacy	Cathegorical	number of drugs >=5
163	cognitive impairment	cognitive_impairment_MMSE_educative_level	Cathegorical	
164	Drug 1 Comercial name	drug_1_comercial_name	Cathegorical	Drug 1 comercial name
165	Drug 1 Active drug	drug_1_PA	Cathegorical	Drug 1 Active drug
166	Drug 1 ATC code	drug_1_ATC	Cathegorical	Drug 1 ATC code
167	Drug 1 therapy type	drug_1a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN
168	Drug 1 start date	drug_1b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN
169	Drug 2 Comercial name	drug_2_comercial_name	Cathegorical	Drug 2 comercial name
170	Drug 2 Active drug	drug_2_PA	Cathegorical	Drug 2 Active drug
171	Drug 2 ATC code	drug_2_ATC	Cathegorical	Drug 2 ATC code
172	Drug 2 therapy type	drug_2a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN
173	Drug 2 start date	drug_2b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN
174	Drug 3 Comercial name	drug_3_comercial_name	Cathegorical	Drug 3 comercial name
175	Drug 3 Active drug	drug_3_PA	Cathegorical	Drug 3 Active drug
176	Drug 3 ATC code	drug_3_ATC	Cathegorical	Drug 3 ATC code
177	Drug 3 therapy type	drug_3a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN
178	Drug 3 start date	drug_3b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN
179	Drug 4 Comercial name	drug_4_comercial_name	Cathegorical	Drug 4 comercial name
180	Drug 4 Active drug	drug_4_PA	Cathegorical	Drug 4 Active drug
181	Drug 4 ATC code	drug_4_ATC	Cathegorical	Drug 4 ATC code

111. Maintains house alone or with occasional assistance (e.g., "heavy work domestic help")(independent, score=1);
222. Performs light daily tasks such as dishwashing, bed making(independent, score=1);
333. Performs light daily tasks but cannot maintain acceptable level of cleanliness (dependent, score=0)
444. Needs help with all home maintenance tasks (dependent, score=0)
555. Does not participate in any housekeeping tasks (dependent, score=0)
111. Does personal laundry completely (independent, score=1); 222. Launders small items; rinses stockings, etc. (dependent, score=0)
333. All laundry must be done by others (dependent, score=0)
111. Travels independently on public transportation or drives own car(independent, score=1);
222. Arranges own travel via taxi, but does not otherwise use public transportation (independent, score=1);
333. Travels on public transportation when assisted or accompanied by another (dependent, score=0)
444. Travel limited to taxi or automobile with assistance of another (dependent, score=0)
555. Does not travel at all (dependent, score=0)
111. Is responsible for taking medication in correct dosages at correct time(independent, score=1);
222. Takes responsibility if medication is prepared in advance in separate dosages (dependent, score=0)
333. Is not capable of dispensing own medication (dependent, score=0)
111. Manages financial matters independently (budgets, writes checks, pays rent and bills, goes to bank),
collects and keeps track of income (independent, score=1);
222. Manages day-to-day purchases, but needs help with banking, major purchases, etc. (independent, score=1);
333. Incapable of handling money (dependent, score=0);

1=YES, 2=NO, 88=MISSING
1=YES, 2=NO, 88=MISSING
1=YES, 2=NO, 88=MISSING
1=YES, 2=NO
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Alzheimer's disease.; 2. Vascular dementia.; 3. Mixed dementia.; 4. Dementia with Lewy bodies.;
5. Frontotemporal dementia.; 6. Dementia associated to Parkinson's disease.;
7. Senile dementia.; 8. Other dementia.; 9. Don't Know.; 10. No answer.
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Shoulders ;2 Elbows ; 3 Wrists ; 4 Metacarpophalangeal; 5 Proximal interphalangeal; 6 Hips; 7 Knees; 8 Ankles; 9 Others
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Shoulders ;2 Elbows ; 3 Wrists ; 4 Metacarpophalangeal; 5 Proximal interphalangeal; 6 Hips; 7 Knees; 8 Ankles; 9 Others
1. Knees; 2 Hips; 3 Knees and hips; 4 Others; 5 Don't Know; 6 Don't answer.
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1. Yes; 2. No; 3. Don't know; 4. No answer
1.None; 2.<300 euros; 3. 301-500 euros; 4.501-700 euros; 5.701-900 euros; 6.901-1.500 euros;
7.1.501- 2.000 euros; 8.2001- 3.000 euros; 9.3.001-4000 euros;
10.más de 4001 euros.; 11.NS; 12.NC
1.None; 2.< 200 euros; 3.201-300 euros; 4.301-500 euros; 5.501-700 euros; 6.701-900 euros;
7.901-1.100 euros; 8.1.101-1.300 euros; 9.1301-1.500 euros;
10.1501-2000 euros; 11.2001-3000 euros; 12.3001-4000 euros; 13.more than 4001 euros.;
14.NS; 15.NC
1. 1 person; 2. 2 persons; 3. 3 persons; 4. 4 persons; 5. 5 persons; 6. 6persons;
7. 7persons; 8. 8 or more persons; 9. don't know; 10. No answer
REF.:Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic
comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987; 40: 373-83

182	Drug 4 therapy type	drug_4a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
183	Drug 4 start date	drug_4b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
184	Drug 5 Comercial name	drug_5_comercial_name	Cathegorical	Drug 5 comercial name	
185	Drug 5 Active drug	drug_5_PA	Cathegorical	Drug 5 Active drug	
186	Drug 5 ATC code	drug_5_ATC	Cathegorical	Drug 5 ATC code	
187	Drug 5 therapy type	drug_5a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
188	Drug 5 start date	drug_5b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
189	Drug 6 Comercial name	drug_6_comercial_name	Cathegorical	Drug 6 comercial name	
190	Drug 6 Active drug	drug_6_PA	Cathegorical	Drug 6 Active drug	
191	Drug 6 ATC code	drug_6_ATC	Cathegorical	Drug 6 ATC code	
192	Drug 6 therapy type	drug_6a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
193	Drug 6 start date	drug_6b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
194	Drug 7 Comercial name	drug_7_comercial_name	Cathegorical	Drug 7 comercial name	
195	Drug 7 Active drug	drug_7_PA	Cathegorical	Drug 7 Active drug	
196	Drug 7 ATC code	drug_7_ATC	Cathegorical	Drug 7 ATC code	
197	Drug 7 therapy type	drug_7a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
198	Drug 7 start date	drug_7b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
199	Drug 8 Comercial name	drug_8_comercial_name	Cathegorical	Drug 8 comercial name	
200	Drug 8 Active drug	drug_8_PA	Cathegorical	Drug 8 Active drug	
201	Drug 8 ATC code	drug_8_ATC	Cathegorical	Drug 8 ATC code	
202	Drug 8 therapy type	drug_8a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
203	Drug 8 start date	drug_8b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
204	Drug 9 Comercial name	drug_9_comercial_name	Cathegorical	Drug 9 comercial name	
205	Drug 9 Active drug	drug_9_PA	Cathegorical	Drug 9 Active drug	
206	Drug 9 ATC code	drug_9_ATC	Cathegorical	Drug 9 ATC code	
207	Drug 9 therapy type	drug_9a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
208	Drug 9 start date	drug_9b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
209	Drug 10 Comercial name	drug_10_comercial_name	Cathegorical	Drug 10 comercial name	
210	Drug 10 Active drug	drug_10_PA	Cathegorical	Drug 10 Active drug	
211	Drug 10 ATC code	drug_10_ATC	Cathegorical	Drug 10 ATC code	
212	Drug 10 therapy type	drug_10a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
213	Drug 10 start date	drug_10b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
214	Drug 11 Comercial name	drug_11_comercial_name	Cathegorical	Drug 11 comercial name	
215	Drug 11 Active drug	drug_11_PA	Cathegorical	Drug 11 Active drug	
216	Drug 11 ATC code	drug_11_ATC	Cathegorical	Drug 11 ATC code	
217	Drug 11 therapy type	drug_11a	Cathegorical	How do you take it? 1. continuous; 2. intermittent; 3. Sporadic; other NAN	
218	Drug 11 start date	drug_11b	Cathegorical	When did you start to take it? 1. less than 1 month; 2. from 1 month to 1 year; 3. more than 1 year; other. NAN	
219	@1_year_smoker	@1_year_smoker	Cathegorical	smoker for at least one year	1=yes; 0=no
220	current_smoker	current_smoker	Cathegorical	current smoker	1=yes; 0=no

List of Figures

List of Tables

References

Aagaard, Kjersti, Kevin Riehle, Jun Ma, Nicola Segata, Toni-Ann Mistretta, Cristian Coarfa, Sabeen Raza, Sean Rosenbaum, Ignatia Van den Veyver, Aleksandar Milosavljevic, et al. [2012]. *A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. PloS one*, 7(6), page e36466.

Abu-Mostafa, Yaser S, Malik Magdon-Ismail, and Hsuan-Tien Lin [2012]. *Learning from data*, volume 4. AMLBook Singapore.

Aggarwal, Charu C [2014]. *Data classification: algorithms and applications.* CRC Press.

Akaike, Hirotugu [1974]. *A new look at the statistical model identification. IEEE transactions on automatic control*, 19(6), pages 716–723.

Allison, Paul D [2000]. *Multiple imputation for missing data: A cautionary tale. Sociological methods & research*, 28(3), pages 301–309.

Ambler, Gareth, Rumana Z Omar, and Patrick Royston [2007]. *A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Statistical methods in medical research*, 16(3), pages 277–298.

Anbarasi, M, E Anupriya, and NCSN Iyengar [2010]. *Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology*, 2(10), pages 5370–5376.

Balakrishnama, Suresh and Aravind Ganapathiraju [1998]. *Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing*, 18.

Barnett, Vic and Toby Lewis [1998]. *Outliers in statistical data.* 3rd Edition. Wiley.

Baylis, D, DB Bartlett, HE Syddall, G Ntani, CR Gale, C Cooper, JM Lord, and AA Sayer [2013]. *Immune-endocrine biomarkers as predictors of frailty and mortality: a 10-year longitudinal study in community-dwelling older people. Age*, 35(3), pages 963–971.

Beaulieu-Jones, Brett K, Jason H Moore, The Pooled Resource Open-access Als, and Clinical Trials Consortium [2016]. *Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders.* In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 22, page 207. NIH Public Access.

Bellazzi, Riccardo, Fulvia Ferrazzi, and Lucia Sacchi [2011]. *Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5), pages 416–430.

Bellazzi, Riccardo and Blaz Zupan [2008]. *Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics*, 77(2), pages 81–97.

Berkhin, Pavel [2002]. *Survey Of Clustering Data Mining Techniques.* Technical Report.

Bishop, Christopher M [2006]. *Pattern Recognition and Machine Learning.*

Blum, Avrim L. and Pat Langley [1997]. *Selection of Relevant Features and Examples in Machine Learning. Artif. Intell.*, 97(1-2), pages 245–271. ISSN 0004-3702. doi:10.1016/S0004-3702(97)00063-5. `http://dx.doi.org/10.1016/S0004-3702(97)00063-5`.

Botsis, Taxiarchis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng [2010]. *Secondary use of EHR: data quality issues and informatics opportunities. AMIA Summits Transl Sci Proc*, 2010, pages 1–5.

Breiman, Leo [2001]. *Random forests. Machine learning*, 45(1), pages 5–32.

Breunig, Markus M, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander [2000]. *LOF: identifying density-based local outliers.* In *ACM sigmod record*, volume 29, pages 93–104. ACM.

Bright, Tiffani J, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R Coeytaux, Gregory Samsa, Vic Hasselblad, John W Williams, Michael D Musty, et al. [2012]. *Effect of clinical decision-support systemsa systematic review. Annals of internal medicine*, 157(1), pages 29–43.

Buuren, Stef and Karin Groothuis-Oudshoorn [2011]. *mice: Multivariate imputation by chained equations in R. Journal of statistical software*, 45(3).

Calvert, Jacob, Qingqing Mao, Jana L Hoffman, Melissa Jay, Thomas Desautels, Hamid Mohamadlou, Uli Chettipally, and Ritankar Das [2016]. *Using electronic health record collected clinical variables to predict medical intensive care unit mortality. Annals of Medicine and Surgery*, 11, pages 52–57.

Caruana, Rich and Alexandru Niculescu-Mizil [2006]. *An empirical comparison of supervised learning algorithms.* In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth [2000]. *CRISP-DM 1.0, Step-by-step data mining guide.*

Chen, Ming-Syan, Jiawei Han, and Philip S. Yu [1996]. *Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and data Engineering*, 8(6), pages 866–883.

Cheng, Yu, Fei Wang, Ping Zhang, and Jianying Hu [2016]. *Risk prediction with electronic health records: A deep learning approach.* In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 432–440. SIAM.

Cockrell, Joseph R and Marshal F Folstein [2002]. *Mini-mental state examination. Principles and practice of geriatric psychiatry*, pages 140–141.

Committee, Economic Policy et al. [2009]. *The 2009 Ageing Report: economic and budgetary projections for the EU-27 Member States (2008-2060). European Economy*, (2).

Cortes, Corinna and Vladimir Vapnik [1995]. *Support-Vector Networks. Mach. Learn.*, 20(3), pages 273–297. ISSN 0885-6125. doi:10.1023/A:1022627411411. `http://dx.doi.org/10.1023/A:1022627411411`.

Cousineau, Denis and Sylvain Chartier [2010]. *Outliers detection and treatment: a review. International Journal of Psychological Research*, 3(1), pages 58–67.

Delen, Dursun, Glenn Walker, and Amit Kadam [2005]. *Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine*, 34(2), pages 113–127.

Duchateau, Luc and Paul Janssen [2007]. *The frailty model*. Springer Science & Business Media.

Eapen, Zubin J, Li Liang, Gregg C Fonarow, Paul A Heidenreich, Lesley H Curtis, Eric D Peterson, and Adrian F Hernandez [2013]. *Validated, electronic health record deployable prediction models for assessing patient risk of 30-day rehospitalization and mortality in older heart failure patients. JACC: Heart Failure*, 1(3), pages 245–251.

Ester, Martin, Hans peter Kriegel, Jörg S, and Xiaowei Xu [1996]. *A density-based algorithm for discovering clusters in large spatial databases with noise*. pages 226–231. AAAI Press.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth [1996]. *From data mining to knowledge discovery in databases. AI magazine*, 17(3), page 37.

Folstein, Marshal F, Susan E Folstein, and Paul R McHugh [1975]. *"Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research*, 12(3), pages 189–198.

Frank, Eibe and Ian H Witten [1999]. *Making better use of global discretization*. In *16th International Conference on Machine Learning (ICML 99)*, pages 115–123. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Fried, Linda P, Catherine M Tangen, Jeremy Walston, Anne B Newman, Calvin Hirsch, John Gottdiener, Teresa Seeman, Russell Tracy, Willem J Kop, Gregory Burke, et al. [2001]. *Frailty in older adults evidence for a phenotype. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(3), pages M146–M157.

Futoma, Joseph, Jonathan Morris, and Joseph Lucas [2015]. *A comparison of models for predicting early hospital readmissions. Journal of biomedical informatics*, 56, pages 229–238.

Garcia-Garcia, Francisco José, G Gutierrez Avila, Ana Alfaro-Acha, MS Amor Andres, MV Escribano Aparicio, S Humanes Aparicio, JL Larrion Zugasti, M Gomez-Serranillo Reus, F Rodriguez-Artalejo, L Rodriguez-Manas, et al. [2011]. *The prevalence of frailty syndrome in an older population from Spain. The Toledo Study for Healthy Aging. The journal of nutrition, health & aging*, 15(10), pages 852–856.

Geethika Bhavya, Peddibhotla [2015]. *KDnuggets: Top 20 R Machine Learning and Data Science packages. http://www.kdnuggets.com/2015/06/top-20-r-machine-learning-packages.html.*

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot [2010]. *Variable selection using random forests. Pattern Recognition Letters*, 31(14), pages 2225–2236.

Goldstein, Benjamin A, Ann Marie Navar, Michael J Pencina, and John PA Ioannidis [2016]. *Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association*, page ocw042.

Gregory, Piatetsky [2014]. *KDnuggets Poll: What main methodology are you using for your analytics, data mining, or data science projects? http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html.* Last access 09/2016.

Gupta, Sunil, Truyen Tran, Wei Luo, Dinh Phung, Richard Lee Kennedy, Adam Broad, David Campbell, David Kipp, Madhu Singh, Mustafa Khasraw, et al. [2014]. *Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ open*, 4(3), page e004007.

Guyon, Isabelle and André Elisseeff [2003]. *An Introduction to Variable and Feature*

Selection. J. Mach. Learn. Res., 3, pages 1157–1182. ISSN 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944968.

Hagan, Martin T., Howard B. Demuth, and Mark Beale [1996]. *Neural Network Design.* PWS Publishing Co., Boston, MA, USA. ISBN 0-534-94332-2.

Hamerly, Greg and Charles Elkan [2004]. *Learning the k in A> means. Advances in neural information processing systems*, 16, page 281.

Han, Jiawei, Jian Pei, and Micheline Kamber [2011]. *Data mining: concepts and techniques.* Elsevier.

Hand, David J, Heikki Mannila, and Padhraic Smyth [2001]. *Principles of data mining.* MIT press.

Hartigan, J. A. and M. A. Wong [1979a]. *Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pages pp. 100–108. ISSN 00359254. http://www.jstor.org/stable/2346830.

Hartigan, John A and Manchek A Wong [1979b]. *Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pages 100–108.

Haux, Reinhold [2010]. *Medical informatics: past, present, future. International journal of medical informatics*, 79(9), pages 599–610.

Herland, Matthew, Taghi M Khoshgoftaar, and Randall Wald [2014]. *A review of data mining using big data in health informatics. Journal of Big Data*, 1(1), page 2.

Hilbe, Joseph M [2009]. *Logistic regression models.* CRC press.

Holzinger, Andreas [2016a]. *Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics*, 3(2), pages 119–131.

Holzinger, Andreas [2016b]. *Lecture Notes on Machine Learning in Health Informatics.* Graz University of Technology, Austria. http://hci-kdd.org/machine-learning-for-health-informatics-course/. Last access 09/2016.

Holzinger, Andreas, Matthias Dehmer, and Igor Jurisica [2014]. *Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. BMC bioinformatics*, 15(6), page I1.

Hosmer Jr, David W and Stanley Lemeshow [2004]. *Applied logistic regression.* John Wiley & Sons.

Jacobson, Olof and Hercules Dalianis [2016]. *Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. ACL 2016*, page 191.

Jain, A. K., M. N. Murty, and P. J. Flynn [1999]. *Data Clustering: A Review. ACM Comput. Surv.*, 31(3), pages 264–323. ISSN 0360-0300. doi:10.1145/331499.331504. `http://doi.acm.org/10.1145/331499.331504`.

Jaspers, Monique WM, Marian Smeulers, Hester Vermeulen, and Linda W Peute [2011]. *Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. Journal of the American Medical Informatics Association*, 18(3), pages 327–334.

Jensen, Peter B, Lars J Jensen, and Søren Brunak [2012]. *Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics*, 13(6), pages 395–405.

Jordan, MI and TM Mitchell [2015]. *Machine learning: Trends, perspectives, and prospects. Science*, 349(6245), pages 255–260.

Karlsson, Isak, Jing Zhao, Lars Asker, and Henrik Boström [2013]. *Predicting adverse drug events by analyzing electronic patient records.* In *Conference on Artificial Intelligence in Medicine in Europe*, pages 125–129. Springer.

Kaufman, L. and Peter J. Rousseeuw [1990]. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley-Interscience. ISBN 0471878766.

Keim, Daniel A [2002]. *Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics*, 8(1), pages 1–8.

Ketchen Jr, David J and Christopher L Shook [1996]. *The application of cluster analysis in strategic management research: an analysis and critique. Strategic management journal*, pages 441–458.

Kohavi, Ron and George H John [1997a]. *Wrappers for feature subset selection. Artificial intelligence*, 97(1), pages 273–324.

Kohavi, Ron and George H. John [1997b]. *Wrappers for Feature Subset Selection. Artif. Intell.*, 97(1-2), pages 273–324. ISSN 0004-3702. doi:10.1016/S0004-3702(97)00043-X. `http://dx.doi.org/10.1016/S0004-3702(97)00043-X`.

Kop, Reinier, Mark Hoogendoorn, Leon MG Moons, Mattijs E Numans, and Annette ten Teije [2015]. *On the advantage of using dedicated data mining techniques to predict colorectal cancer.* In *Conference on Artificial Intelligence in Medicine in Europe*, pages 133–142. Springer.

Kumari, Milan and Sunila Godara [2011]. *Comparative study of data mining classification methods in cardiovascular disease prediction 1.*

Kursa, Miron B, Aleksander Jankowski, and Witold R Rudnicki [2010]. *Boruta–a system for feature selection. Fundamenta Informaticae*, 101(4), pages 271–285.

Kursa, Miron B. and Witold R. Rudnicki [2010]. *Feature Selection with the Boruta Package. Journal of Statistical Software*, 36(11), pages 1–13. ISSN 1548-7660. `http://www.jstatsoft.org/v36/i11`.

Kurt, Imran, Mevlut Ture, and A Turhan Kurum [2008]. *Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications*, 34(1), pages 366–374.

Landwehr, Niels, Mark Hall, and Eibe Frank [2005]. *Logistic model trees. Machine Learning*, 59(1-2), pages 161–205.

Laurikkala, Jorma, Martti Juhola, Erna Kentala, N Lavrac, S Miksch, and B Kavsek [2000]. *Informal identification of outliers in medical data.* In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, volume 1, pages 20–24.

Lawton, MP and ELMNE M BRODY [1970]. *Assessment of older people: self-maintaining and instrumental activities of daily living. Nursing Research*, 19(3), page 278.

Lippi, Giuseppe, Pidder Jansen-Duerr, Jose Viña, Anna Durrance-Bagale, Imad Abugessaisa, David Gomez-Cabrero, Jesper Tegnér, Johannes Grillari, Jorge Erusalimsky, Alan Sinclair, et al. [2015]. *Laboratory biomarkers and frailty: presentation of the FRAILOMIC initiative. Clinical Chemistry and Laboratory Medicine (CCLM)*, 53(10), pages e253–e255.

Little, Roderick JA and Donald B Rubin [2014]. *Statistical analysis with missing data.* John Wiley & Sons.

Louppe, Gilles [2014]. *Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502.*

Mani, Subramani, Yukun Chen, Tom Elasy, Warren Clayton, and Joshua Denny [2012]. *Type 2 diabetes risk forecasting from EMR data using machine learning.* In *AMIA annual symposium proceedings*, volume 2012, page 606. American Medical Informatics Association.

McEvoy, Peter M, David M Erceg-Hurn, Lisa M Saulsman, and Michel A Thibodeau [2015]. *Imagery enhancements increase the effectiveness of cognitive behavioural group therapy for social anxiety disorder: A benchmarking study. Behaviour research and therapy*, 65, pages 42–51.

Murray, Christopher JL, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al. [2013]. *Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. The lancet*, 380(9859), pages 2197–2223.

Niaksu, Olegas [2015]. *CRISP Data Mining Methodology Extension for Medical Domain. Baltic Journal of Modern Computing*, 3(2), page 92.

Nilsson, Nils J [1996]. *Introduction to machine learning. An early draft of a proposed textbook.*

Nilsson, Roland, José M Peña, Johan Björkegren, and Jesper Tegnér [2007]. *Consistent feature selection for pattern recognition in polynomial time. Journal of Machine Learning Research*, 8(Mar), pages 589–612.

Ohno-Machado, Lucila et al. [2015]. *Mining electronic health record data: finding the gold nuggets. Journal of the American Medical Informatics Association*, 22(5), pages 937–937.

Oztekin, Asil, Dursun Delen, and Zhenyu James Kong [2009]. *Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. International journal of medical informatics*, 78(12), pages e84–e96.

Palaniappan, Sellappan and Rafiah Awang [2008]. *Intelligent heart disease prediction system using data mining techniques.* In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE.

Perer, Adam, Fei Wang, and Jianying Hu [2015]. *Mining and exploring care pathways from electronic medical records with visual analytics. Journal of biomedical informatics*, 56, pages 369–378.

Perlich, Claudia [2016]. *Top Data Scientist Claudia Perlich on Biggest Issues in Data Science.* `http://www.kdnuggets.com/2016/09/perlich-biggest-issues-data-science.html`. Last access 10/2016.

Pincus, Theodore, Jane A Summey, Salvatore A Soraci, Kenneth A Wallston, and Norman P Hummon [1983]. *Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis & Rheumatism*, 26(11), pages 1346–1353.

Prokosch, Hans-Ulrich, T Ganslandt, et al. [2009]. *Perspectives for medical informatics. Methods Inf Med*, 48(1), pages 38–44.

Pyle, Dorian [1999]. *Data preparation for data mining*, volume 1. morgan kaufmann.

Quetelet, Adolphe Lambert Jaques [1842]. *Recueil d'observations sur différents sujets des sciences physiques.* Hayez.

Ramakrishnan, Naren, David Hanauer, and Benjamin Keller [2010]. *Mining electronic health records. Computer*, 43(10), pages 77–81.

Ramaswamy, Sridhar and Todd R Golub [2002]. *DNA microarrays in clinical oncology. Journal of Clinical Oncology*, 20(7), pages 1932–1941.

Rasmussen, Carl Edward and Zoubin Ghahramani [2001]. *Occam's Razor*. In *In Advances in Neural Information Processing Systems 13*, pages 294–300. MIT Press.

Reunanen, Juha [2003]. *Overfitting in Making Comparisons Between Variable Selection Methods. J. Mach. Learn. Res.*, 3, pages 1371–1382. ISSN 1532-4435. `http://dl.acm.org/citation.cfm?id=944919.944978`.

Rosenblatt, Frank [1958]. *The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review*, 65(6), pages 386–408.

Royston, Patrick et al. [2004]. *Multiple imputation of missing values. Stata journal*, 4(3), pages 227–41.

Rubin, Donald B [1976]. *Inference and missing data. Biometrika*, pages 581–592.

Rubin, Donald B [1977]. *Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association*, 72(359), pages 538–543.

Rudnicki, Witold R, Marcin Kierczak, Jacek Koronacki, and Jan Komorowski [2006]. *A statistical method for determining importance of variables in an information system*. In *International Conference on Rough Sets and Current Trends in Computing*, pages 557–566. Springer.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga [2007]. *A review of feature selection techniques in bioinformatics. bioinformatics*, 23(19), pages 2507–2517.

Saeys, Yvan, Louis Wehenkel, Pierre Geurts, et al. [2012]. *Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics*, 28(13), pages 1766–1774.

Saulnier, Delphine M, Kevin Riehle, Toni-Ann Mistretta, Maria-Alejandra Diaz, Debasmita Mandal, Sabeen Raza, Erica M Weidler, Xiang Qin, Cristian Coarfa, Aleksandar Milosavljevic, et al. [2011]. *Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology*, 141(5), pages 1782–1791.

Schwarz, Gideon et al. [1978]. *Estimating the dimension of a model. The annals of statistics*, 6(2), pages 461–464.

Seigel, AF [1988]. *Statistics and data analysis: an introduction.* John Wiley & Sons, Inc.

Shah, Anoop D, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway [2014]. *Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology*, 179(6), pages 764–774.

Shouman, Mai, Tim Turner, and Rob Stocker [2012]. *Using data mining techniques in heart disease diagnosis and treatment.* In *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*, pages 173–177. IEEE.

Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni [2011]. *Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications*, 17(8), pages 43–48.

Srinivas, K, B Kavihta Rani, and A Govrdhan [2010]. *Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE)*, 2(02), pages 250–255.

Stoppiglia, Hervé, Gérard Dreyfus, Rémi Dubois, and Yacine Oussar [2003]. *Ranking a random feature for variable and feature selection. Journal of machine learning research*, 3(Mar), pages 1399–1414.

Sun, Jimeng, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Zahra Daar, and Walter F Stewart [2012]. *Combining knowledge and data driven insights for identifying risk factors using electronic health records.* In *AMIA*, volume 2012, pages 901–10.

Swindell, William R, Kristine E Ensrud, Peggy M Cawthon, Jane A Cauley, Steve R Cummings, and Richard A Miller [2010]. *Indicators of" Healthy Aging" in older women (65-69 years of age). A data-mining approach based on prediction of long-term survival. BMC geriatrics*, 10(1), page 55.

The R Foundation [2016]. *What is R?* `https://www.r-project.org/about.html`. Last access 09/2016.

Thompson, Bruce [2005]. *Canonical correlation analysis. Encyclopedia of statistics in behavioral science.*

Van Buuren, Stef [2012]. *Flexible imputation of missing data.* CRC press.

Van Buuren, Stef, Hendriek C Boshuizen, Dick L Knook, et al. [1999]. *Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine*, 18(6), pages 681–694.

van Buuren, Stef, Karin Groothuis-Oudshoorn, Alexander Robitzsch, Gerko Vink, Lisa Doove, and Shahab Jolani [2015]. *Package 'mice'.*

Verleysen, Michel and Damien François [2005]. *The curse of dimensionality in data mining and time series prediction.* In *International Work-Conference on Artificial Neural Networks*, pages 758–770. Springer.

Wartner, Sandra, Dominic Girardi, Manuela Wiesinger-Widi, Johannes Trenkler, Raimund Kleiser, and Andreas Holzinger [2016]. *Ontology-Guided Principal Component Analysis: Reaching the Limits of the Doctor-in-the-Loop.* In *International Conference on Information Technology in Bio-and Medical Informatics*, pages 22–33. Springer.

Washburn, Richard A, Kevin W Smith, Alan M Jette, and Carol A Janney [1993]. *The Physical Activity Scale for the Elderly (PASE): development and evaluation. Journal of clinical epidemiology*, 46(2), pages 153–162.

Weenink, David [2003]. *Canonical correlation analysis.* In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 25, pages 81–99. University of Amsterdam.

Weiss, Jeremy C, Sriraam Natarajan, Peggy L Peissig, Catherine A McCarty, and David Page [2012]. *Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. AI Magazine*, 33(4), page 33.

Werbos, P. [1974]. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* PhD Thesis, Harvard University, Cambridge, MA.

Witten, Ian H, Eibe Frank, Mark A Hall, and Christopher J Pal [2016]. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Wolpert, David H and William G Macready [1997]. *No free lunch theorems for optimization. IEEE transactions on evolutionary computation*, 1(1), pages 67–82.

Wu, Jionglin, Jason Roy, and Walter F Stewart [2010]. *Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care*, 48(6), pages S106–S113.

Ye, Jieping, Ravi Janardan, Qi Li, et al. [2004]. *Two-Dimensional Linear Discriminant Analysis.* In *NIPS*, volume 4, page 4.

Yesavage, Jerome A, Terence L Brink, Terence L Rose, Owen Lum, Virginia Huang, Michael Adey, and Von Otto Leirer [1983a]. *Development and validation of a geriatric depression screening scale: a preliminary report. Journal of psychiatric research*, 17(1), pages 37–49.

Yesavage, Jerome A, Terence L Brink, Terence L Rose, Owen Lum, Virginia Huang, Michael Adey, and Von Otto Leirer [1983b]. *Development and validation of a geriatric depression screening scale: a preliminary report. Journal of psychiatric research*, 17(1), pages 37–49.

Yesavage, Jerome A and Javaid I Sheikh [1986]. *9/Geriatric Depression Scale (GDS) recent evidence and development of a shorter violence. Clinical gerontologist*, 5(1-2), pages 165–173.

Yoo, Illhoi, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua [2012]. *Data mining in healthcare and*

biomedicine: a survey of the literature. Journal of medical systems, 36(4), pages 2431–2448.

Zanin, Massimiliano, David Papo, Pedro A Sousa, Ernestina Menasalvas, Andrea Nicchi, Elaine Kubik, and Stefano Boccaletti [2016]. *Combining complex networks and data mining: why and how. Physics Reports*, 635, pages 1–44.

Zhang, Zhongheng [2016]. *Multiple imputation with multivariate imputation by chained equation (MICE) package. Annals of translational medicine*, 4(2).

Zhu, Xiaojin [2011]. *Semi-supervised learning.* In *Encyclopedia of machine learning*, pages 892–897. Springer.

Zurada, Jacek M. [1992]. *Introduction to Artificial Neural Systems.* West Publishing Company.