TÉCNICO LISBOA

Imputation Techiques for Clinical Data of Ischemic Stroke Patients

Filipa Matos Marques

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Arlindo Manuel Limede de Oliveira
Prof. Alexandre Paulo Lourenço Francisco

Examination Committee

Chairperson: Prof. Daniel Jorge Viegas Gonçalves
Supervisor: Prof. Arlindo Manuel Limede de Oliveira
Member of the Committee: Prof. João Miguel Raposo Sanches

January 2021

Acknowledgments

I would like to thank my supervisors, professors Arlindo Oliveira and Alexandre Francisco, for guiding me through out what was the hardest task I ever had to do in my life.

To my friends, thank you for never letting me quit. A special acknowledgment goes to Diogo Lopes and Inês Silva for listening to all my whining and calling me out when needed.

Finally, a special thanks to my family and their enjoyment in buying houses which allowed me to isolate myself during quarantine for writing this thesis.

Resumo

No século XXI, todos os anos, na Europa, 880 mil pessoas sofrem um acidente vascular cerebral isquémico. Prever a evolução e desfecho do paciente é crucial aquando da escolha do tratamento. Nesta tese de mestrado foram criados vários modelos de modo a prever o desfecho do paciente através da versão binária da escala de Rankin modificada em dois momentos no tempo: três meses e um ano após a ocorrência do derrame.

É comum que os dados providenciados pelas entidades de saúde para a realização destes estudos estejam incompletos comprometendo os resultados. Torna-se então necessário escolher uma maneira apropriada de lidar com os dados omissos neste trabalho, optou-se por testar seis métodos de imputação diferentes e, com os dados completos, treinar classificadores com sete modelos de aprendizagem automática distintos.

Conclui-se que a área sob a curva característica de operação do recetor, para o melhor classificador, a prever a escala a três meses e um ano foi 0.8217 e 0.7537, respetivamente. Ademais, não foram encontradas diferenças, com significância estatística, no desempenho dos distintos métodos de imputação quando avaliados para cada um dos modelos de aprendizagem automática.

Palavras-chave: Acidente Vascular Cerebral Isquémico, Dados Omissos, Técnicas de Imputação, Aprendizagem Automática

Abstract

In the 21st century, every year, approximately 880 thousand people living in Europe suffer an ischemic stroke. Predicting the patient's outcome is key to choosing the course of treatment. In this master thesis, it was predicted the functional outcome, by the binary version, of the modified Rankin Scale at two points in time: three months and one year after the stroke took place.

Often, data provided by health organisations to conduct these studies is incomplete which can impair the results. Thus the need arises to choose a proper way to handle the missing data. Here missing values were imputed with six different methods and the classifiers were then trained with seven distinct machine learning models.

It was shown the area under the receiver operating characteristic curve for the best classifiers, at the three months and one-year marks, are 0.8217 and 0.7537, respectively. Moreover, it was not found a statistically significant difference between the performance of the distinct imputation methods for each machine learning model.

Keywords: Ischemic Stroke, Missing Data, Imputation Techniques, Machine Learning

Contents

List of Tables

List of Figures

Nomenclature

Acc Accuracy

AI Artificial Intelligence

ASTRAL Acute Stroke Registry and Analysis of Lausanne

AUC Area Under the Curve

CART Classification And Regression Trees

CVA Cerebrovascular Accident

DT Decision Trees

EHR Electronic Health Record

FN False Negatives

FP False Positives

FPR False Positive Rate

HADS Hospital Anxiety and Depression Scale

INESC-ID Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento

KNN k-Nearest Neighbours

Lasso Least Absolute Shrinkage and Selection Operator

LOOCV Leave One Out Cross-Validation

LR Logistic Regression

MAR Missing at Random

MCAR Missing Completely at Random

ML Machine Learning

MLP Multilayer Perceptron

MMSE Mini-Mental State Examination

MoCA Montreal Cognitive Assessment

mRS modified Rankin Scale

N Negative

NIHSS National Institutes of Health Stroke Scale/Score

NMAR Not Missing at Random

NN Neural Networks

PR Precision-Recall

RF Random Forest

ROC Receiver Operating Characteristic

SVM Support Vector Machines

THRIVE Totaled HealthRisks in Vascular Events

TN True Negatives

TP True Positives

TPR True Positive Rate

Xgboost Extreme Gradient Boosting

Y Positive

Chapter 1

Introduction

1.1 Motivation

A cerebrovascular accident (CVA), or stroke, results from ischemia caused by thrombosis, malformation, stenosis, or a haemorrhage from a ruptured aneurysm [32]. Roughly, 1.1 million people living in Europe suffer a stroke yearly in the 21st century, ischemic strokes accounting for approximately 80% of cases and this number is expected to rise to 1.5 million because of the ageing population [6, 101]. Furthermore, ischemic stroke incidence is also increasing in young adults in high-income countries, including Europe [5, 7, 27, 35, 67].

In the year 2013, 11.8% of all deaths were attributed to stroke making it the second main cause of death in the world (half of these were from ischemic strokes). Furthermore, in 2013, a CVA is also the third most common cause of disability (4.5%) being responsible for 113 million disability-adjusted life-years globally [26].

The primary goal of stroke rehabilitation is improving the survivors' independence, not only for the patients' well being, but because most indirect stroke-related costs come from compromised physical functioning and the need for a caregiver involvement [53].

The course of treatment is highly dependable on the predicted outcome of the patient meaning that any tool created to help predict the patients' functional outcome are immensely useful. Moreover, it is common for both the patient and the family to ask for a long term prognosis which is an answer that is neither immediate nor straightforward [69].

Over the last decade, the medical community has been searching for the best scores to predict the patients' functional outcome using data available at admission, making it possible to have a more informed treatment decision. Among them, the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) [75], the DRAGON [99] and the Totaled HealthRisks in Vascular Events (THRIVE) [29] scores.

Currently, the modified Rankin Scale (mRS) is the gold standard used scale for "measuring the degree of disability or dependence in the daily activities of people who have suffered a stroke or other causes of neurological disability" [88, 105]. The scale goes as follows (the physician should choose the best fit of the patients' ability) [88]:

i Score 0: No symptoms.

ii Score 1: No significant disability. Able to carry out all usual activities, despite some symptoms.

iii Score 2: Slight disability. Able to look after own affairs without assistance, but unable to carry out all previous activities.

iv Score 3: Moderate disability. Requires some help, but able to walk unassisted.

v Score 4: Moderately severe disability. Unable to attend to own bodily needs without assistance, and unable to walk unassisted.

vi Score 5: Severe disability. Requires constant nursing care and attention, bedridden, incontinent.

vii Score 6: Dead.

Some of this scale's major strengths are: i) it covers the full range of functional outcomes, from no symptoms to death [10] ii) its categorization is intuitive and easily understood by clinicians and patients [10] iii) its concurrent validity is demonstrated by strong correlation with measures of stroke pathology and agreement with other stroke scales [42]. The main criticisms from the scientific community to mRS has been its subjectivity when determining between categories and its reproducibility by examiners and patients [42].

When the goal is the prediction of patients' functional outcome, to simplify, a binary version of the mRS with only two classes (good outcome vs poor outcome) is usually used. Additionally, it is also common to measure the mRS at three months and one year after the stroke [69].

1.2 Topic Overview

The term Artificial Intelligence (AI) was used for the first time in a conference in 1956 at Dartmouth [85]. Scientists soon realized medicine was a promising application area leading to the development of many clinical decision support systems [68, 71]. Rule-based models had a lot of success in the next few years across several areas such as helping to make diagnoses [18], choosing appropriate treatments [95] and interpreting electrocardiograms [52]. Nonetheless, these systems are costly and require knowledge updates as new relations are discovered. Moreover, their performance is limited by the comprehensiveness of prior medical knowledge [8].

As stated in 1987 at a Sounding Board article, the field of medicine is "so broad and complex that it is difficult, if not impossible, to capture the relevant information in rules." [91]. Nowadays, AI is comprised, for example, of machine learning (ML) methods able to identify patterns and account for complex interactions within the data [21].

ML methods can be categorized as supervised or unsupervised. Supervised models require 'training' cases, *i.e.* which contain inputs and the correct output labels, and learn by analysing the patterns in the provided labelled input–output pairs and, afterwards, minimizing the difference between its predictions for the training cases and its observed outcomes. On the other hand, unsupervised models deduce the

underlying patterns of data not containing output labels trying to find outliers, identify sub-clusters or produce low-dimensional representations of the data [109].

The amount of large-scale annotated clinical data is increasing due to the adoption of electronic health record (EHR) systems, ML methods are also getting better every year and are readily available in opensource packages. These along with the rapidly growing computational power and cloud storage have contributed to the current growth in AI which, in turn, is expected to alter the landscape of medical practice in the close future [46, 109].

Nowadays, AI systems have already specialist-level performance in a wide range of medical tasks [25, 41]. They are applied across several areas: image-based diagnosis (radiology [55, 103], dermatology [25], ophthalmology [41], pathology [4, 58]), genome interpretation [77, 78], biomarker discovery [43], clinical outcome prediction and patient monitoring [12, 19], inferring health status with the help of wearable devices [56, 65] and autonomous robotic surgery [64, 92], to name a few. Furthermore, they also allow physicians to be in contact with areas they haven't been able to before, as AI enables remote healthcare services for rural and low-income zones [76].

In order for this change in the landscape of medical practice to happen, several challenges are yet to be met:

i Availability of Good Quality Data: ML requires clear data, preferably labelled, which isn't always available. Class imbalanced datasets are quite common and may create biased results. Data sparsity is another problem affecting the model's performance (some models can't handle missing values at all) [76].

ii Lack of Data Standardization and Exchange: healthcare information technology systems don't tend to follow data exchange standards for healthcare thus affecting data's efficacy and quality [76].

iii Safety Challenges: models are ought to perform well on subtle outliers, edge cases and hidden strata to ensure the safety of already implemented systems [76].

iv Privacy Challenges: the individuals' identification should not be possible as such the data should be anonymized to prevent privacy breaches [1]. This would enable secure data sharing through cloud services [72].

v Ethical Challenges: targeted user population and their sociological aspects should be understood before the collection of data as well as how this might impact a patient's well-being and dignity [76].

vi Understandable algorithms: several ML methods are still "black boxes" making it hard to interpret, both its conclusion and created biological insights, and identify its weaknesses [109].

vii Integration into clinical workflows: consequences such as interference in interpersonal communication styles, additional workloads for physicians, generation of new hazards [2] and alert fatigue [13] should be dealt with.

viii Regulatory Challenges: there is still a lot of questions regarding how models should be regulated by the competent agencies given its fast evolvement as it is fed with new data and user feedback. It is also unclear how updates should be evaluated [109].

ix Legal Challenges: it is yet to be decided how medical negligence attributed to complex decision support systems should be handled. When malpractice cases including AI systems arise, the legal system must provide clear guidance as to what entity holds the liability [94].

As said above, missing data is a problem present in EHR and, given that a lot of ML models only work on complete datasets, it is a problem that needs to be dealt with either by deleting incomplete observations or by imputing it, *i.e.*, replacing any values that are missing with a value estimated by the remaining available information [87].

When choosing between the various approaches one should take the source of missingness into evaluation [16]:

i Is the value lost or forgotten? The value was measured but not recorded due to *e.g* sensors disconnected, accidental human omission.

ii Is the value not applicable to the instance? The value was not measured due to an identifiable reason *e.g* physicians decided to disconnect the patient from the ventilator.

iii Is the value not of interest to the instance? The value was not measured due to its unusefulness in providing clinical information.

In the cases where missing data is a consequence of identifiable reasons, imputation might not be the best choice as it will add bias to the dataset. This data is then referred to as non-recoverable. On the other hand, data missing for unidentifiable reasons, also called recoverable data, is presumed to be missing as a result of unintended and random causes and can therefore be imputed [16].

1.3 Objectives

Our work aims to use machine learning techniques to predict the functional outcome of a patient using the binary version of the modified Rankin Scale: good outcome for scores 0 to 2 and poor outcome for scores 3 to 6. This is done at two points in time: three months and one year after the initial stroke. In addiction, we are also interested in studying the impact of missing data and the choice of the data imputation technique in machine learning models.

We start by performing the imputation using six distinct approaches: mode/median according to the quantitative/qualitative nature of the variables, mode/median according to the quantitative/qualitative nature of the variables and taking into account the dependence of a few variables, hotdeck, k-nearest neighbours, decision trees and multiple imputation with posterior predictive distribution/conditional mean imputation once again according to the quantitative/qualitative nature of the variables.

We then compare the impact of the different imputation methods in each machine learning model (L1 regression, Support Vector Machines, Random Forest, Xgboost, Neural Networks, Classification And Regression Trees and k-Nearest Neighbours).

To our knowledge, similar work has only been done by *Woźnica et al.* [107] who analysed different imputation methods for a collection of datasets and a collection of machine learning algorithms. Similarly, *Jadhav et al.* [44] and *Kyureghian et al.* [54] have evaluated some of the existing imputation techniques, yet not in the same way. They focused on the quality of imputed data, by assessing the accuracy of predicting the missing values to fully known simulated data.

1.4 Thesis Outline

This thesis is divided into the following chapters:

- Chapter 2. Background - In this section, the existent missing data mechanisms are presented as well as different ways to handle missing data. Basic concepts about some supervised machine learning models and cross validation are shown along with a few performance metrics.
- Chapter 3. Implementation - In this section, an overview of the dataset is presented. Afterwards, the methods used for data cleaning, manipulation and imputation are enumerated. Finally, the different machine learning models applied and how their evaluation and training was conducted is shown.
- Chapter 4. Results and Discussion - Here the results of the training and model evaluation for the different imputation approaches to predict the binary version of the mRS, at both three months and one year after the initial stroke, are shown.
- Chapter 5. Conclusions and Future Work- Here we present an overview of the results obtained and leave some remarks regarding future work.

Chapter 2

Background

2.1 Missing Data Mechanisms

It is important to understand the mechanisms by which the data is missing before addressing the issue of imputation since it will have an impact on some of the assumptions made. *Little et al.* [60, 82] formulated three possible missing data mechanisms taking into account the relation between the missing (unobserved) and the available (observed) data.

For simplicity's sake, let's consider missingness in the univariate case. To define missingness, in mathematical terms, a dataset X can be divided into two parts:

$$X = \{X_0, X_m\} \tag{2.1}$$

Being X_0 the observed data and X_m the missing data in the dataset.

For each observation we define a binary response:

$$R = \begin{cases} 1 & \textit{if X observed} \\ 0 & \textit{if X missing} \end{cases} \tag{2.2}$$

The missing value mechanism can be understood in terms of the probability that an observation is missing $Pr(R)$ given the observed and missing observations, in the form:

$$Pr(R \mid x_0, x_m) \tag{2.3}$$

The three possible missing data mechanisms are then defined as follows [87]:

i Missing Completely at Random (MCAR): the probability of an observation being missing depends only on itself, *i.e.*, equation (2.3) takes the form:

$$Pr(R \mid x_0, x_m) = Pr(R) \tag{2.4}$$

MCAR is the highest level of random given that the missingness does not depend on any information in the dataset. In a medical setting, this might correspond to a doctor forgetting to record the gender of a couple random patients that come in the emergency room - there is no hidden mechanism related to any variable and it does not depend on any characteristic of the patients.

ii Missing at Random (MAR): the probability of a value being missing is related only to the observable information, *i.e.*, some statistical relationship exists between the observed and the missing variables meaning the missing data may be traceable from the observed values in the dataset. Mathematically, the probability of missing, equation (2.3), reduces to:

$$Pr(R \mid x_0, x_m) = Pr(R \mid x_0) \tag{2.5}$$

As a medical example, let's assume elderly people are less probable to notify the physician they have had pneumonia before, the response rate of the variable "pneumonia" will be correlated to the variable "age".

iii Not Missing at Random (NMAR): the probability of a value being missing depends both on missing and observed values. It refers to the case when neither MCAR nor MAR holds, the pattern of missing data is not random and non predictable from available values.

NMAR is usually regarded as the worst type as it might lead to bias whereas MCAR and MAR may lead to loss of statistical power [37, 90]. Determining the missing mechanism is usually impossible, as it depends on unseen data. A t-test comparing the characteristics of the groups' missing values and observed values on a certain variable will yield different characteristics if the data is not MCAR yet the result is merely indicative since it always depends on the sample size of the data. Additionally, there is no method for distinguishing between MAR or NMAR data [60]. Given this impossibility we must rely on sensitivity analyses and testing how the inference holds under different conditions, *e.g.* diabetic patients will have their blood sugar measured more often than non diabetic patients meaning the variable "blood sugar" depends on the variable "diabetic" [44].

Brown et al. [11] add outliers treated as missing data to the above list of standard types of missing data. The authors deem necessary the deletion of set values as their stay may skew test results.

2.2 Handling Missing Values

Various machine learning models cannot handle missing data so when presented with it the analyst has a choice: ignoring it or using imputation, *i.e.*, replacing the missing value with plausible ones.

The goal of the various imputation methods is the accurate estimation of population parameters in order to keep the power of the following data analysis and data mining techniques. There is no rule as to what method should be chosen to handle the missing values of a given dataset yet there is a common agreement that imputation should be used with care in datasets with over 25% of the data missing [44].

2.2.1 Ignoring Missing Values

The easiest way to handle missing data is to omit the observations or cases that have missing values. Although this is often the standard method, it reduces the dataset. Therefore should only be used when a small amount of missing values is present [44]. Moreover, usually, deletion methods lead to valid inferences only for MCAR data [89]. There are two general approaches:

- Complete-Case Analysis / Listwise Deletion: observations with one or more missing values are discarded. It is assumed that the sample is representative of the whole population meaning the analysis will not be biased towards a subgroup. Indeed, it has been shown that when the relationships within a dataset are strong and do not include the outcome variable, the method often produces unbiased regression slope estimates [87].

  Therefore listwise deletion might be used for large datasets having a low percentage of missing data yet always keeping in mind it reduces statistical power (a smaller number of samples lead to estimates with larger standard errors), wastes information and it can create a possible bias of the analysis especially if data is not MCAR [90].

- Available-Case Analysis/ Pairwise Deletion: observations with one or more missing values are only discarded if they are being analysed, *e.g.* two analysis are being conducted: the first with variables A and B and the second with variables A and C; if pairwise deletion is chosen then the observations in variable C with missing data will be kept for the first analysis and the observations in variable B with missing data will be kept for the second analysis. Consequently, sample sizes will be different making it impossible to make a statistical comparison of the results [87].

2.2.2 Single Imputation

Single imputation fills missing values with a predicted value all the while ignoring uncertainty resulting often in the underestimation of variance [87]. Similarly to the deletion methods, several approaches can be taken. Below is a non-comprehensive list:

- Imputation with a constant: the missing values are replaced with a constant. When dealing with a categorical variable one might replace it with “Missing” or a value of no significance, *e.g.*, “999”.

- Mode, Mean and Median Imputation: the categorical and numerical missing values are replaced by the variables mode or mean/median, respectively. The mean should only be used for populations which have a normal distribution, otherwise the median should be used [87]. The latter is also more robust to outliers.

  There are disadvantages [11]: i) The new variance understates the true variance. ii) The new distribution has more values under the category containing mean/median/mode than the true population. iii) The correlations between variables are diminished.

  Studies show this imputation method yields highly biased parameter estimates [38–40], yet this effect is almost dissipated if less than 10% of the data is missing and the correlations between the

variables are low [79, 102].

A special case can be used, conditional mode/mean/median. Here a variable is grouped according to a second variable and the mode/mean/median is computed for every unique value of the second variable. It might be useful when a known relation exists, *e.g.*, sportsmen usually have lower blood pressure thus we impute the missing observations in ”blood pressure” with two different medians depending on whether or not our patient is a ”sportsman”. The method is not recommended when analyses of correlations or covariances are to take place given that one of the consequences of this is the overstatement of the strength of the relationship between the two variables [90].

- Hot Deck Imputation: the missing values are replaced with a value from the known data's estimated distribution. The implementation is done in two steps, first the data is grouped in clusters and each missing value is attributed to a cluster. Then a distribution for the variable with the missing values is created for each cluster and the missing value is filled. This simple approach allows for the variable distribution's preservation however it underestimates the variability [81]. Moreover, the definition of ”similar” for the creation of clusters is not straightforward, several metrics can be used which will result in different imputations [11].

- Model-Based Imputation: the missing values are replaced by values estimated by a predictive model. The complete data will be used to create a model, *e.g.*, regression, logistic regression, neural networks or other (non) parametric modelling techniques. Due to its characteristics, the model won't have high accuracy when the data is MCAR. When rightly applied, its estimated values are usually more well-behaved than the true values [87].

- Regression Imputation: the missing values are replaced by values estimated by a regression model (a particular case of a Model-Based Imputation). This imputation method, like the Hot Deck, is able to preserve the distribution shape however it might produce biased results when applied to NMAR and MAR data.

  There are disadvantages since it does not take into account the uncertainty in the missing data [11]: i) It assumes the estimated variable correlates with the remaining variables in the dataset. ii) It reinforces relationships already existent in the dataset reducing its generalization capability. iii) It understates the distribution's variance. iv) The estimated value is not constrained and may consequently be outside predetermined boundaries for set variable thus requiring additional adjustment.

- k-Nearest Neighbours Imputation: the missing values are replaced by the mean of the k values coming from the k most similar complete observations. There are several ways to compute this similarity (distance functions, *e.g.*, Euclidean, Manhattan, Mahalanobis, Pearson, etc) notwithstanding it is very time consuming for a large dataset. Moreover, the value given to k should be thoroughly investigated, the value should be large enough to encompass all significant attributes yet not as large that would include attributes which significantly differ from our target observation [87]. Its

main advantage is the fact that the correlation structure of the data is taken into consideration. Additionally, it can handle both discrete and continuous variables [87].

2.2.3 Multiple Imputation

Single imputation tends to underestimate the variance and ignores uncertainty [87] while multiple imputation incorporates uncertainty into its methods [59]. *Rubin* [83] created a method which takes the average of the outcome across multiple imputed datasets. The imputation of multiple plausible values allows the model to account for uncertainty. This Monte Carlo technique consists of three steps:

i Imputation: missing values are replaced, using a method of choice, M times (5–10 is generally sufficient) [89].

ii Analysis: every M completed dataset is analysed (*e.g.* it is built a logistic regression classifier for outcome prediction), resulting in M analyses [87].

iii Pooling: the M analysis and results are consolidated into one final one, *e.g.*, by computing the mean and the 95 % CI of the M analyses [87].

The above three steps make multiple imputation a very time consuming step, which is why many analyst opt not to choose set imputation.

2.3 Machine Learning Models

Over the last few decades dozens of different machine learning models have been developed by scientists for multiple applications. Below are seven supervised models briefly explained.

2.3.1 Logistic Regression: L1-regularised

Belonging to the generalized linear models' family, logistic regressions (LR) are commonly used in datasets with both numerical and categorical features where the outcome is a binary categorical variable. It works by assigning weights to each input feature and outputting a value between 0 and 1, which can be understood as a probability of "success" regarding the target variable. Any probability over 0.5 is considered a "success" [3].

The Least Absolute Shrinkage and Selection Operator (Lasso) regression, regression with L1 regularization, shrinks data values towards a central point, *e.g.* mean, and is particularly relevant for models with high levels of multicollinearity or when the automation of variable selection or parameter elimination is the goal [49]. Given a set of instance-label pairs (x_i, y_i) of size p, where $x \in \mathbb{R}^n$, $i = 1, 2, ...l$ and $y \in \{+1, -1\}$, the goal of the logistic regression (L1-regularised) algorithm is to minimize [31]:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.6}$$

The parameter λ controls the strength of the penalty, the amount of shrinkage: i) λ = 0, no parameters are eliminated and the estimate coincides with the one by linear regression. ii) λ increases, an increasing number of coefficients are set to zero/eliminated (theoretically, when $\lambda = \infty$, all coefficients are eliminated). Bias also increases. iii) λ decreases, variance increases.

This regularization may result in sparse models which have few coefficients, as some coefficients, λ_j may reduce to zero and be eliminated from the model [49].

2.3.2 Support Vector Machines

A Support Vector Machines (SVM) classifier uses a collection of the training points, named support vectors, in the decision function in order to define the decision boundary between classes. Its goal, as training happens, is to find the optimal hyperplane for classifying the testing dataset based on the referred support vectors and some constraints. Given a training dataset, $D = (x_i, y_i)$ of size p with $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ and label $y_i = -1$ or $+1$, formally, the classifier can be defined as a quadratic optimization problem solving the following equation:

$$\min \|w\|^2 \;\; s.t. \;\; y_i(w^T x_i + b) \geq 1 \;\; for\ all\ i \tag{2.7}$$

where $w = (w_1, w_2, ..., w_p)$ represents the weight vector and b the bias. An important point to be noted when training an SVM model is the parameter C which controls the trade-off between having a wide margin and correctly classifying training data.

$$\min \|w\|^2 + C \sum_{1}^{m} \xi_i \;\; s.t. \;\; y_i(w^T x_i + b) \geq (1 - \xi_i), \xi_i \geq 0 \;\; for\ all\ i \tag{2.8}$$

A larger value of C corresponds to a smaller number of misclassified training samples and is susceptible to overfitting [57, 108].

2.3.3 Random Forest

A Random Forest (RF) is a supervised ensemble method consisting of a number of decision trees [9]. A Decision Tree (DT) consists of the following structures, as shown in fig. 2.1: i) Root node, has no incoming edges and one to multiple outgoing edges. ii) Internal nodes, have exactly one incoming edge and multiple outgoing edges. iii) Edges/Branches, connect between nodes. Typically represent conjunctions of features that lead to those class labels. iv) Leaf/Terminal Nodes, have exactly one incoming edge and no outgoing edges. Typically represent class labels.

The non-terminal nodes (root and internal nodes) contain attribute test conditions based on which data is split or separated due to different characteristics. The criteria for separation can be calculated through information gain or entropy calculation, *e.g.*, using the following equations:

$$Entropy = \sum_{j=1}^{C} p_j \; log \frac{1}{p_j} \tag{2.9}$$

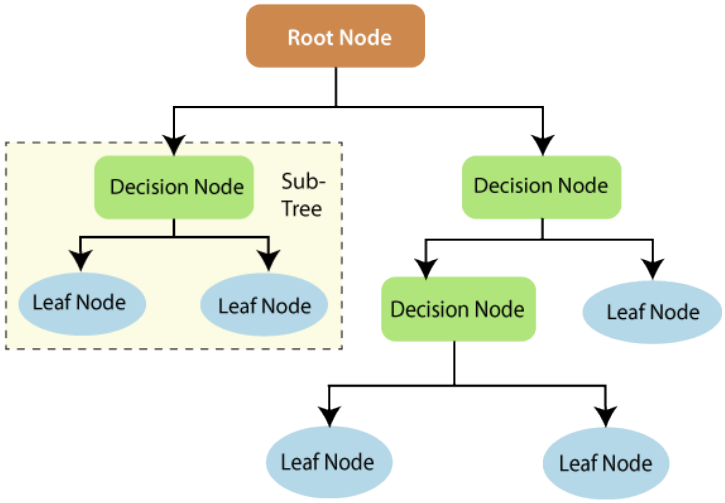


Figure 2.1: Diagram showing the three types of nodes in a decision tree [66].

$$Gini\ Index = 1 - \sum_{j=1}^{C} p_j^2 \tag{2.10}$$

where C is the set of classes and p_j is the fraction of items labelled with class j.

Random Forests aggregate the votes from several decision trees to determine the final classes of the test dataset as to improve the accuracy and control the well known tendency of decision trees to overfit. Individual DT commonly tend to overfit and show high variance, in order to control these, each tree is built from a sample drawn with replacement from the training set. Additionally, when splitting the nodes when building a tree, the best split is found either from all input features or a random subset of maximum features. This process of injecting randomness yields individual decision trees with disassociated prediction errors and by taking an average of the predictions, some errors cancel out, providing a better model [3, 9].

2.3.4 Extreme Gradient Boosting

Contrary to RF, where an ensemble of independent recursive trees of unlimited depth is built, Gradient Boosting creates a sequential series of smaller trees by correcting the residuals from one tree to the next [93].

The Extreme Gradient Boosting (Xgboost) also works by an additive strategy. Given a training dataset, $D = (x_i, y_i)$ with $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,q})$ and label $y_i = -1$ or $+1$, a tree ensemble model uses K additive functions to predict the output:

$$\widehat{y_l} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F}, \tag{2.11}$$

being $\mathcal{F}$ the set of every possible regression trees. Xgboost aims to minimize the regularized objective using the following equations:

$$\mathcal{L}(\phi) = \sum_i l(\widehat{y_l}, y_i) + \sum_k \Omega(f_k) \tag{2.12}$$

where

$$\Omega(f) = \gamma\, T + \frac{1}{2}\, \lambda\, \|\omega\|^2, \tag{2.13}$$

being l a differentiable and convex loss function. The second term in equation (2.13) is used to penalize the model's complexity (number of leaves in the tree T and vector of scores on leaves ω) thus helping to smooth the final learned weights, avoiding overfitting.

Furthermore, after every step, the newly added weights are scaled by a factor η. This technique, named shrinkage, helps to reduce each tree's impact and to slow the learning stage (in an assumption that this will result in a better model) [15, 61, 62].

2.3.5 Neural Networks

Neural Networks (NN) were made to imitate the neuronal structure of the human brain being the most popular the feed-forward multilayer perceptron (MLP). Normally, the MLP network contains multiple layers that define the information transfer between the input and the response layers. Every hidden layer is composed of nodes having weighted connections defining the strength of the information which flows between layers. The training goal is to identify the optimal weights that result in a lower prediction error for a training dataset. The MLP is usually trained by the backpropagation algorithm [84], it starts by assigning a random weighting scheme to all the network's connections and then the following steps are repeated: i) Estimating the output in the response variable via the forward-propagation of information through the network. ii) Calculating the difference between the response's predicted value and true value. iii) Changing the weights through a backpropagation step beginning at the output layer followed by the remaining hidden layers.

A fitted MLP neural network with two layers can be represented as [104]:

$$y_k = f_0 \left(\sum_h w_{hk} f_h \left(\sum_i w_{ih} x_i \right) \right) \tag{2.14}$$

where the estimated value of the response variable y_k is a sum of products between the respective weights w for i input variables x and h hidden nodes, mediated by the activation functions f_h and f_o for each hidden and output node, respectively.

2.3.6 Classification And Regression Trees

Classification And Regression Trees, commonly known as CART, is a term introduced by Breiman [36] to refer to DT algorithms that can be used for classification problems. The representation for the CART model is a binary tree, *i.e.*, the algorithm works by repeatedly finding the best predictor variable to split the data into two subsets. To select the input variable to be used and the specific split point, a greedy algorithm (e.g. the Gini index in equation (2.10)) is used to minimize a cost function (e.g. sum squared error). Tree construction ends utilizing a predefined stopping criterion, such as a minimum number of training instances assigned to each leaf node of the tree [36].

2.3.7 k-Nearest Neighbours

The k-Nearest Neighbours (KNN) algorithm [80, 104] used for classification is the same as the one described above for imputation. The algorithm works at it follows: i) Initialize k, *i.e.*, the number of neighbours to be considered ii) For each observation in the test dataset, calculate the distance between the test observation and the remaining observations in the train dataset. Then, for each observation in the test dataset: iii) Sort the distances from smallest to largest iv) Pick the first k entries v) Get the labels of the selected k entries vi) If regression, return the mean of the k labels vii) If classification, return the mode of the k labels.

2.4 Cross Validation

In ML it is customary to divide the dataset into three [45]: i) Training dataset: used to train the model, estimate the model parameters, typically composed of 60% of the available data. ii) Validation dataset: used to provide an unbiased evaluation of the model's fitness. The model which performs well on training data is then run on the validation dataset and if the error on this dataset increases then we can conclude we have an overfitting model. It is typically composed of 10% to 20% of the available data. iii) Test dataset/ Holdout dataset: used to perform the final model evaluation. This dataset contains data that has never been used in training. Typically composed of 5% to 20% of the available data.

When a dataset is large enough it is possible to split it accordingly. Unfortunately, this is not always the case, when enough data is not available one might opt for having only a training and a test dataset and no validation set yet this comes with a few downsides. Sample variability between the train and the test dataset leads to the models yielding a good prediction on the train dataset but failing to generalize on the test dataset, *i.e.*, having a low training error rate and a high test error rate. Moreover, when splitting the dataset into three (train, test and validation datasets), only a subset of the data is used to train the model meaning the model is likely to not perform well and overestimate the test error rate for the model to fit on the complete dataset [45].

Therefore, when the dataset is not large enough to deem the last two disadvantages irrelevant, one may choose to use an approach called cross validation. This statistical technique consists of dividing the data into two subsets, one for training and the other for evaluating the model's performance. In order to reduce variability, this technique is performed multiple times using different subsets of the same dataset. In the end, the validation results of the multiple rounds are combined and a more accurate estimate of the model's predictive performance is obtained [45].

There are several cross validation approaches, the two most used are:

- Leave one out cross validation (LOOCV): when dividing the data into two subsets, the test dataset has only one observation while the training dataset has the remaining. Thus if a dataset has n observations, the technique is repeated n times and n performance metrics are generated which makes execution expensive. Additionally, being the test dataset composed of a single observation, if this observation is an outlier then the variability will be much higher. On the other hand, it is a

better approach than the validation set where only a subset of the data is used for training, and it is obtained a less biased model [45].

- k Fold: the dataset is randomly divided into k groups/folds of equal size. The model is trained on k-1 folds and afterwards tested on the first one, this is repeated k times using a different test fold every time. The value of k is usually 10 as extensive studies show this value gives the best estimate of error [106]. Although being computationally intensive, it presents improvements in time when compared to LOOCV. Likewise, the bias is reduced and, with the increase of k, it is obtained a reduction in variance [45].

2.5 Performance Metrics

In order to evaluate the classification performance, several metrics must be known:

(i) Accuracy (Acc) - the proportion of true results, both true positives (TP) and true negatives (TN), among the total number of examined cases (n).

$$Accuracy = \frac{TP + TN}{n} \tag{2.15}$$

(ii) Sensitivity/ Recall/ True Positive Rate (TPR) - the proportion of positives, both TP and false negatives (FN), that are correctly identified.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.16}$$

(iii) Specificity - the proportion of negatives, both TN and false positives (FP), that are correctly identified.

$$Specificity = \frac{TN}{TN + FP} \tag{2.17}$$

(iv) Precision - ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{2.18}$$

(v) $F1_{score}$ - weighted average of Precision and Recall. Should be used in the case of an uneven class distribution.

$$F1_{score} = 2\,\frac{Precision\ .\ Recall}{Precision + Recall} \tag{2.19}$$

2.5.1 Receiver Operating Characteristic Curves

Binary imbalanced classifications often make use of Receiver Operating Characteristic (ROC) Curves and Precision-Recall curves to help with its interpretation, namely, to understand the trade-off in performance for different threshold values. These curves are particularly useful in imbalanced datasets

because they are not biased to the majority or minority class. In order to directly compare classification models, it is common to use the area under the ROC curve (AUC) which can be interpreted as "the probability that the scores given by a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one" [28].

The ROC curve is plotted with true positive rate (as in equation (2.16)) against the false positive rate (FPR, as in equation (2.20)) where TPR is on y-axis and FPR is on the x-axis. It is built by calculating the TPR and FPR for different thresholds that separate two binary classes at different points.

$$FPR = 1 - Specificity = \frac{FP}{FP + TN} \tag{2.20}$$

AUC

The AUC is a measure of separability, it gives information regarding how well a model is able to distinguish between classes. An AUC of 0.5 translates to an incapacity of the model to distinguish between classes and, from this value up, the higher the AUC the better the distinction between positive class and negative class [73].

In order to better understand these concepts, a ROC curve and its underlying probability distributions were plotted, being the red distribution curve the positive class and the green distribution curve the negative class. Fig. 2.2 shows a perfect situation where there is no overlap between the two curves meaning the models is able to distinguish perfectly between the two classes [73].

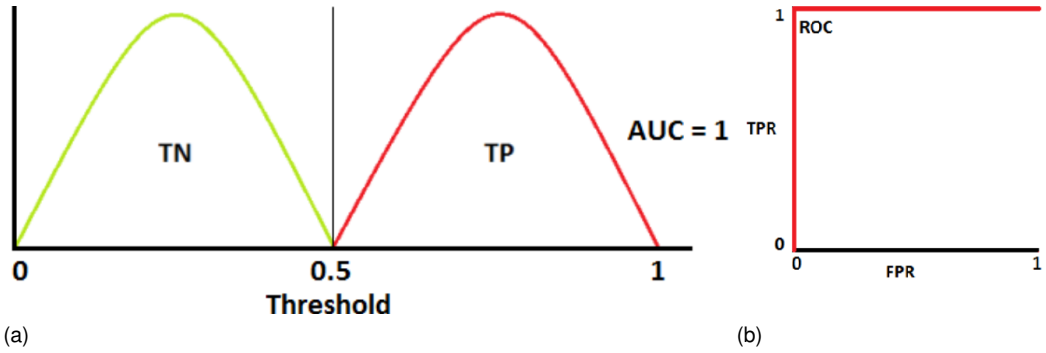


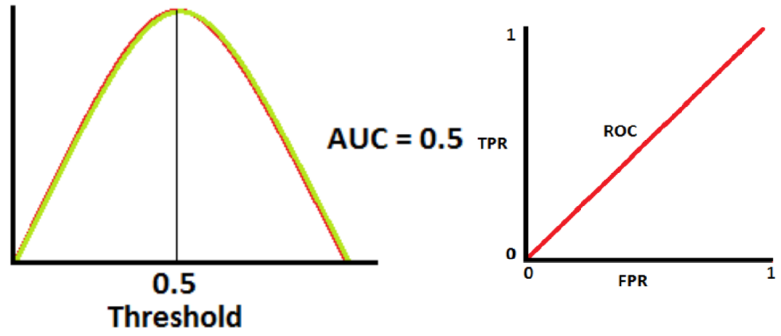
Figure 2.2: Receiver Operating Characteristic Curve (a) and its underlying probability distributions (b) for an ideal model, AUC=1 [73].

The contrary case to Fig. 2.2 is Fig. 2.3, which shows the worst situation, *i.e.*, when the AUC is close or equal to 0.5. Here the model has no discriminatory power thus there is a total overlap of the positive and the negative distributions [73].

Typically, the two distributions partially overlap. In Fig. 2.4 we see an example where the AUC is 0.7 meaning there is a 70% chance that the model will be able to distinguish between both classes.

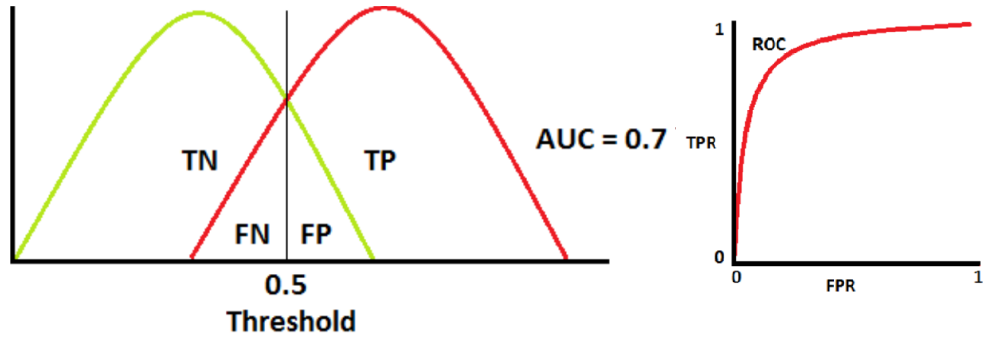
Theoretically, the AUC can be zero which would mean the model was predicting negative class as a positive class and vice versa. This situation is depicted in Fig. 2.5.

The partial AUC can also be computed for a given sensitivity or specificity interval [min,max]. It is common for a standardized version to be presented being the McClish one of the most common



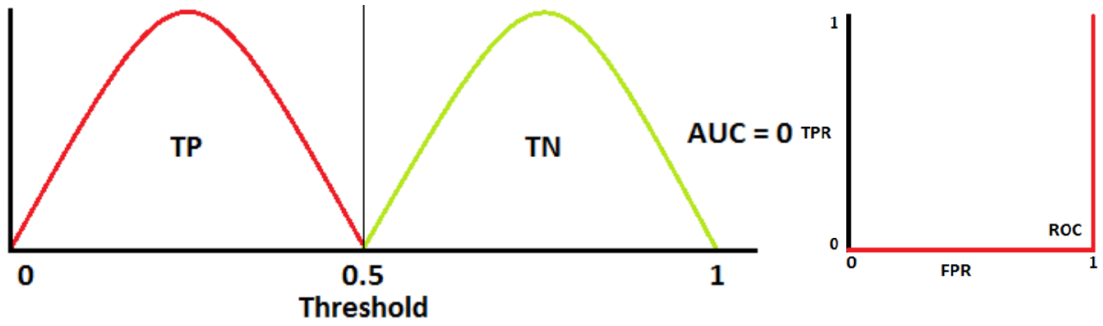
(a) (b)

Figure 2.3: Receiver Operating Characteristic Curve (a) and its underlying probability distributions (b) for the worst model, AUC=0.5 [73].



(a) (b)

Figure 2.4: Receiver Operating Characteristic Curve (a) and its underlying probability distributions (b) for a typical model, AUC=0.7 [73].



(a) (b)

Figure 2.5: Receiver Operating Characteristic Curve (a) and its underlying probability distributions (b) for a model reciprocating classes, AUC=0 [73].

corrections:

$$Corrected\ Partial\ AUC = \frac{1 + \frac{AUC - min}{max - min}}{2} \tag{2.21}$$

Fernández et al. [28] defend ROC AUC may be misleading when computed for an imbalanced dataset with a small minority class and a severe skew. They have shown how a small variation in the number of correct and incorrect predictions results in a large change in the ROC curve and, consequently, in the ROC AUC score providing an excessively optimistic value for performance. It is proposed the use of the precision-recall (PR) curve and PR AUC when these conditions apply. Similar studies

were conducted by Saito [86] which support this conclusion.

2.5.2 Precision-Recall Curves

A precision-recall curve, as the name suggests, is plotted with precision (as in equation (2.18)) against recall (as in equation (2.16)) where precision is on y-axis and recall is on the x-axis. Both these metrics focus on the minority (positive) class and are indifferent to the majority (negative) class.

Fig. 2.6 shows the PR curve for the perfect model, *i.e.*, 100 % precision and 100 % recall, every observation is correctly classified.

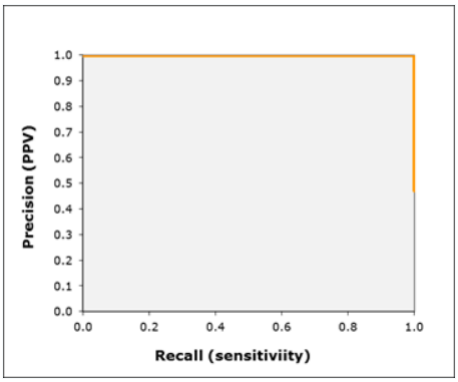


Figure 2.6: Precision-Recall Curve for the perfect model [86].

The contrary case to Fig. 2.6 is Fig. 2.7, the worst situation, *i.e.*, there is a complete overlap of results between the two classes, where the PR curve lies over the ratio between the positive (Y) and the negative (N) groups. Fig. 2.7 shows an imbalanced dataset with the ratio Y:N equal to 1:9 therefore, being X the number of people in the positive group, equation (2.22) shows the precision will remain the same through every value of recall as 0.1 [86].

$$Precision = \frac{TP}{TP + FP} = \frac{X}{X + 9X} = 0.1 \tag{2.22}$$

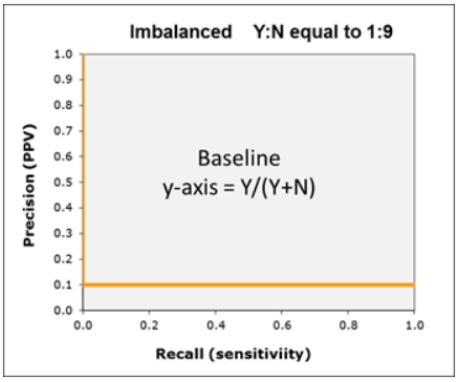


Figure 2.7: Precision-Recall Curve for the worst model - imbalanced distribution Y:N equal to 1:9 [86].

Typically, there is a partial overlap of results between the two classes, as seen in curves represented in Fig. 2.8 The rule is, the closer the curve is to the upper right corner, the better the model [86].

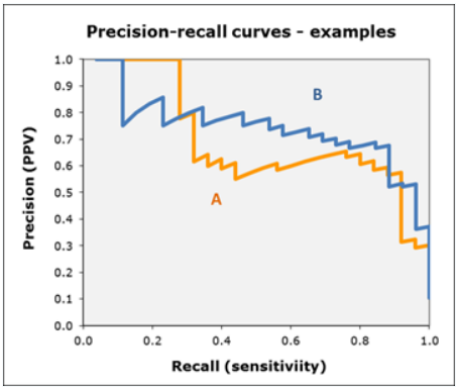


Figure 2.8: Precision-Recall Curve for the worst model [86].

2.5.3 DeLong's Test

DeLong's test [20] is a method for comparing the AUC of different machine learning models.

Being $\hat{\theta}^{(A)}$ the AUC of model A, $\hat{\theta}^{(B)}$ the AUC of model B, $\mathbb{V}$ and $\mathbb{C}$ the variance and covariance, respectively, the z-score is calculated as described in equation (2.23).

$$z \triangleq \frac{\hat{\theta}^{(A)} - \hat{\theta}^{(B)}}{\sqrt{\mathbb{V}[\hat{\theta}^{(A)} - \hat{\theta}^{(B)}]}} = \frac{\hat{\theta}^{(A)} - \hat{\theta}^{(B)}}{\sqrt{\mathbb{V}[\hat{\theta}^{(A)}] + \mathbb{V}[\hat{\theta}^{(B)}] - 2\,\mathbb{C}\,[\hat{\theta}^{(A)}, \hat{\theta}^{(B)}]}} \tag{2.23}$$

This value, z-score, will correspond to a p-value in a table for two-tailed test for z statistics [20, 100].

Commonly, a t-test is used to make the comparison between two machine learning models however the use of k-fold cross validation invalidated that option. A key assumption of the paired Student's t-test is that the observations in each sample are independent yet when using k-fold cross validation a given observation is used in the training dataset k-1 times meaning the estimated skill scores are dependent [22].

Chapter 3

Implementation

3.1 Database: Precise Stroke

The dataset used in this study results from a collaboration, between investigators from Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento in Lisbon (INESC-ID) and from the Santa Maria Hospital in Lisbon, in the project Precise. Our original dataset was comprised of 536 patients however the mRS three months and one year after the event was only recorded for 243 and 234 patients, respectively. This database has data collected at admission, follow-up data, data collected on discharge, three months and one year after the initial stroke.

Before data pre-processing 93% of the dataset features had more than 30% of its data missing, 90% of the dataset features had more than 50% of its data missing and 64% of the dataset features had more than 70% of its data missing. Further research was not done given the elevated number of features in the dataset, 393 and 466 for the mRS three months and one year, respectively. For more information regarding missing data exploration in R the work by *Ghazali et al.* should be consulted [34].

3.2 Data Cleaning and Manipulation

Data cleaning was performed for both predicted outcomes in the same manner by deleting features that contained more than 90% of missing values and features which were meta-data, *e.g.* record number. Variables that record times were converted into time differences between variables, *e.g.* time of the initial event and time of arrival at the hospital becomes time between the event and arrival. Variables consisting of a true/false list were used to create a column for each list entry. Furthermore, observations having a field ”Unknown” or ”Untested” were set to ”NA” whereas the field ”Not Applicable” was kept. Patients for which the mRS was not recorded or were dead by the time of its assessment were removed. Moreover, using the caret package [51], features with zero variance and near-zero variance (the feature must have a ratio of the most common value to the second most common value lower than 95:5) were removed as well as features with a correlation higher than 85%. These last features were removed in order to enable the use of multiple imputation.

After cleaning the data, the resulting dataset had 138 features and 243 patients for the mRS three months and 192 features and 234 patients for the mRS one year.

The target variable was the mRS three months and one year after the event. To turn the problem into a binary classification problem the mRS was discretized into two classes:

- Good outcome: defined by mRS ≤ 2
- Poor outcome: defined by mRS > 2

This particular discretization is of medical relevance because it separates the patients who will be able to live a rather normal independent life from the ones who will require significant assistance.

3.3 Data Imputation

The database Precise Stroke has a number of dependent fields, *i.e.* fields which can only be filled when a third field has a pre-determined answer, as is shown in Fig. 3.1: ”Idade” can only be filled when the previous field’s (”Hipertensão Arterial” or ”Diabetes Mellitus”) answer is ”Sim”. In this cases primal data imputation was performed, single value imputation was used with the value ”9999” or ”0” depending on the variable’s quantitative or qualitative nature, respectively.

Figure 3.1: Example of two dependent fields on the database Precise Stroke.

After data cleaning, manipulation and a primal data imputation, 6 experiments were designed that aimed to assess with what precision we could predict the patient’s mRS three months and the mRS one year after admission. Each experiment corresponds to a different imputation method:

- Mode/Median Imputation (Imp. 1), according to the quantitative/qualitative nature of the variables. Done by using the default settings of the imputeTS package in R [70].
- Mode/Median Imputation (Imp. 2), according to the quantitative/qualitative nature of the variable and taking into account the dependence of a few variables. Done by using the default settings of the imputeTS package in R for each pair dependent/independent variable [70].
- Hotdeck Imputation (Imp. 3). Done by using the default settings of the VIM package in R [50].
- k-Nearest Neighbours Imputation (Imp. 4). Done by using the default settings (k=5 and a variation of the Gower Distance) of the VIM package in R [50].

- Decision Trees Imputation (Imp. 5). Done by using the default settings of the missForest package in R [97, 98].
- Multiple Imputation (Imp. 6) with posterior predictive distribution or conditional mean-imputation, according to the quantitative or qualitative nature of the variables. Done by using the default settings (maximum number of iterations=30 and number of chains=4) of the mi package in R [33].

3.4 Machine Learning Models

For each experiment we used the following classifiers:

- Logistic Regression: L1-regularised. Done by using the default settings of the caret method "regLogistic" provided by the R package LiblineaR [14, 31].
- Support Vector Machines. Done by using the default settings of the caret method "svmPoly" provided by the R package kernlab [57, 108].
- Random Forest. Done by using the default settings of the caret method "rf" provided by the R package randomForest [9].
- Extreme Gradient Boosting. Done by using the default settings of the caret method "xgbLinear" provided by the R package xgboost [15, 61, 62].
- Neural Network. Done by using the default settings of the caret method "nnet" provided by the R package nnet [80, 104].
- Classification And Regression Trees. Done by using the default settings of the caret method "rpart" provided by the R package rpart [36].
- k-Nearest Neighbours. Done by using the default settings (distance = Euclidean) of the caret method "knn" provided by R itself [80, 104].

3.5 Evaluation and Training

To measure the performance of the models it was used the AUC. To train and validate the model it was used 10-fold cross validation, using the caret package [51]. ROC and PR curves were created, using the mLeval R package [47], to further compare the models.

In order to determine if the differences observed between the different classifiers' AUC were statistically significant the DeLong's test was applied, using the pROC R package, and a p-value threshold of 0.05 was chosen.

To determine the best parameterization for each model a grid search was performed over a set of reasonable values. For the LR classifier, the search was over the regularization cost parameter and tolerance. For the SVM classifier, the search was over the cost parameter, the scale and polynomial

degree. For the RF classifier, the search was over the number of randomly selected predictors. For the Xgboost classifier, the search was over the learning rate, the L1 and L2 regularization cost parameter and the number of boosting iterations. For the NN classifier, the search was over the number of hidden units and the weight decay. For the CART classifier, the search was over the complexity parameter. For the KNN classifier, the search was over the number of neighbours.

Chapter 4

Results and Discussion

4.1 Modified Rankin Scale at Three Months

From table 4.1 it can be seen how there is no one better imputation method, it greatly depends on the model being used to train the classifier. *Woźnica et al.* [107] arrived at the same conclusion, more complex methods aren't always the best option. Here the combination which achieved the best results was performing hotdeck imputation and using neural networks as the classification model with an AUC of 0.8217.

In Figure 4.1 can be found the twenty most important variables and its relative importance (scale of 100%) for the best modified Rankin Scale classifier at three months (the most important features for the remaining classifiers can be found in Appendix A). 13 out of the 20 variables are all known predictors that are used by traditional scores (National Institutes of Health Stroke Scale/Score (NIHSS) [63], Hospital Anxiety and Depression Scale (HADS) [96], Mini-Mental State Examination (MMSE) [30] and the Montreal Cognitive Assessment (MoCA) [74]) also appear in the most important features list of the classifiers. The major presence of the NIHSS score comes as no surprise given its metrics measure the symptoms' severity and there is a direct correlation between the severity of the symptoms and the patient's likelihood to recover [69]. Interestingly, features related to recovery are also represented, and ranked fourth and fifth nonetheless, highlighting the importance of physical therapy and speech therapy.

In a medical context, a classifier which has an 80% sensitivity, *i.e.* is able to predict eight in every ten patients who will require significant assistance (positive class), is considered a good model [23]. Looking at table 4.2, the partial AUC values for an 80% sensitivity were computed and the correction by McClish was applied. The best imputation method is the same as when the total AUC is computed, table 4.1, as well as the best pair imputation method/classification model.

Furthermore, the different metrics AUC, AUC-PR, $F1_{score}$ and accuracy in tables 4.1, 4.3, 4.4 and 4.5, respectively, are not in agreement when electing the best imputation method and classification model pair. For AUC-PR, $F1_{score}$ and accuracy the best imputation method would be decision trees paired with a random forest classifier for AUC-PR, extreme gradient boosting for $F1_{score}$ and support vector machines for accuracy. These results show how important it is the choice of the evaluation

metric.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.7665 + 0.0218	0.7615 + 0.0219	**0.8004 + 0.0209**	0.7754 + 0.0216	0.7979 + 0.0208	0.6250 + 0.0230	0.7110 + 0.0219
Imp. 2	0.7705 + 0.0217	0.7589 + 0.0219	0.8001 + 0.0209	0.7566 + 0.0220	0.7984 + 0.0207	0.6433 + 0.0235	0.6919 + 0.0218
Imp. 3	0.7815 + 0.0214	0.7730 + 0.0216	0.7998 + 0.0209	0.7737 + 0.0216	**0.8217 + 0.0198**	0.6456 + 0.0225	0.7061 + 0.0218
Imp. 4	0.7760 + 0.0216	**0.7811 + 0.0214**	0.7916 + 0.0212	**0.7818 + 0.0214**	0.7945 + 0.0209	0.6218 + 0.0233	0.7071 + 0.0220
Imp. 5	0.7891 + 0.0212	0.7672 + 0.0217	0.7876 + 0.0213	-	0.7365 + 0.0224	**0.7087 + 0.0230**	0.7303 + 0.0220
Imp. 6	**0.8014 + 0.0209**	0.7791 + 0.0214	0.7972 + 0.0210	-	0.7768 + 0.0215	0.6768 + 0.0228	**0.7314 + 0.0220**

Table 4.1: AUC results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.8708	0.8670	**0.9114**	0.8819	0.8979	0.6819	0.8045
Imp. 2	0.8759	0.8640	0.9104	0.8590	0.8967	0.7065	0.7783
Imp. 3	0.8885	0.8805	0.9110	0.8791	**0.9151**	0.7119	0.7981
Imp. 4	0.8825	**0.8918**	0.9027	**0.8895**	0.8930	0.6769	0.7989
Imp. 5	0.9006	0.8739	0.8978	-	0.8320	**0.7948**	0.8295
Imp. 6	**0.9129**	0.8896	0.9086	-	0.8794	0.7548	**0.8311**

Table 4.2: Partial AUC results, 80% sensitivity, for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

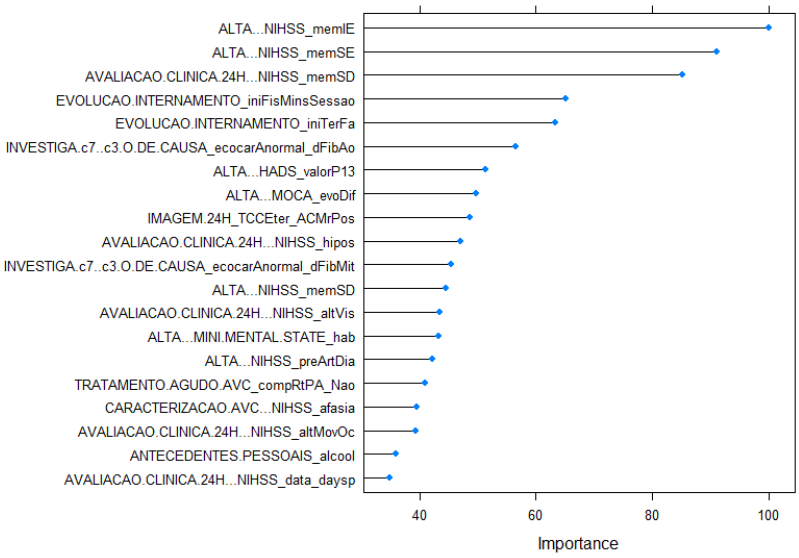


Figure 4.1: The twenty most important variables and its relative importance (scale of 100%) for the best model predicting the modified Rankin Scale at three months: hotdeck imputation and neural network classification.

In chapter 2 it was discussed how the metric AUC could be misleading when computed for an imbalanced dataset as a small variation in the number of correct and incorrect predictions resulted in a large change in the ROC curve and, consequently, in the AUC score providing an excessively optimistic value for performance [28]. *Fernández et al.* advise the reader to, in this situation, use the precision-recall curve and AUC-PR.

Figure 4.2 shows the ROC curves for the six imputation methods and seven different models predicting the modified Rankin Scale at three months. Analysing each model individually it is possible to see the curves overlap several times and there is no clear distinction between them which is expected considering the close AUC values in table 4.1 and the statistical tests performed, table 4.8.

Coherently, the precision-recall curves are zigzagged, figure 4.3. It is common to have noisy curves for small recall values however when this tendency persists for higher recalls, curves for different classifiers crossing each other very frequently, it makes it hard to choose the best classifier by analyzing set curves. It is, however, possible to see two uncommon situations in subfigures 4.3b and 4.3e, a curve not starting at coordinates (0,1) and a curve not starting at recall zero.

Regarding the uncommon start place, the first point of the precision-recall curve is usually estimated based on its second point since the precision value is undefined when the number of positive predictions is 0. This estimation is done taking into consideration the second point's true positive. Whenever the number of true positives, for the second point, is zero then the first point is also (0,0). However, if TP is not zero the first point is estimated by drawing a horizontal line from the second point to the y-axis. Here, the classifier with the decision tree imputation, represented in red in subfigure 4.3b, is not able to avoid false positives at any level of probability threshold, at the lowest level of probability there is one false positive, meaning the first point does not need to be estimated and corresponds to (0.077, 0.500). This situation happens when a classifier is trained with imbalanced data and isn't able to achieve good results, *i.e.*, it is considered a poor classifier [17, 48].

On the other hand, when a precision-recall curve does not start at recall zero then the classifier with, for example, the decision tree imputation, represented in red in subfigure 4.3e, is not able to avoid true positives at any level of probability threshold, at the lowest level of probability there are nine true positives meaning there are no values for precision at lower recalls than 0.692. No reasons were found to explain set situation, one can only assume a larger number of patients would allow the model to develop better distinctions thus improving the precision-recall curve.

Looking at tables 4.1 and 4.3, AUC and AUC-PR, the range of values is bigger for AUC-PR than for AUC, as expected, thus helping to better distinguish between the classification models and imputation methods. Contrary to what happens when using the AUC metric, table 4.1, when the AUC-PR is chosen as an evaluation metric there is clearly one best imputation method, decision trees, yielding best result for four out of the seven models. However, the best pair imputation method/classification model is hotdeck/random forest supporting the first conclusion: there is no one better imputation method, it greatly depends on the model being used to train the classifier.

Kendall's correlation between the different evaluation metrics can be seen in table 4.7. Contrary to what was expected, AUC-PR and $F1_{score}$ have a smaller correlation than AUC-PR with the remaining evaluation metrics. Additionally, the correlation between accuracy and AUC was also smaller than the correlation of accuracy with the remaining evaluation metrics which was also unexpected. Knowing the pair AUC and accuracy give equal importance to the positive and the negative classes while the pair AUC-PR and $F1_{score}$ give greater importance to the positive class, it was expected to see a greater correlation among the metrics in these two pairs which has shown not to be true.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.470	**0.640**	0.630	**0.650**	0.550	0.140	0.500
Imp. 2	0.440	0.630	0.630	0.600	0.550	0.100	0.520
Imp. 3	0.400	0.600	0.610	0.600	0.550	**0.400**	0.400
Imp. 4	0.480	0.600	0.600	0.600	0.530	0.130	0.480
Imp. 5	**0.600**	**0.640**	**0.660**	-	0.530	0.130	**0.600**
Imp. 6	**0.600**	0.620	0.590	-	**0.570**	0.390	**0.600**

Table 4.3: AUC-PR results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.538	0.667	0.688	**0.750**	0.571	0.645	0.621
Imp. 2	0.562	0.667	**0.710**	0.636	0.571	0.645	0.621
Imp. 3	0.583	0.643	0.690	0.595	0.571	0.621	0.583
Imp. 4	**0.667**	0.621	0.688	0.667	0.583	0.621	**0.667**
Imp. 5	0.588	**0.720**	**0.710**	-	**0.640**	0.625	0.588
Imp. 6	0.600	0.636	0.667	-	0.621	**0.667**	0.600

Table 4.4: $F1_{score}$ results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.7292	0.8333	**0.8333**	0.8125	0.6667	0.7917	**0.8125**
Imp. 2	0.7292	0.8125	**0.8333**	0.8125	0.6667	0.7708	**0.8125**
Imp. 3	0.6875	0.8125	0.8125	0.7708	0.7292	0.7500	0.7917
Imp. 4	0.7708	0.7917	0.7917	**0.8333**	0.6875	0.7708	0.7708
Imp. 5	**0.7917**	**0.8542**	**0.8333**	-	**0.8125**	0.7708	0.7708
Imp. 6	0.7500	0.8333	0.7917	-	0.7917	**0.8333**	0.7917

Table 4.5: Accuracy results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.538	0.615	**0.846**	0.692	0.462	**0.769**	0.692
Imp. 2	0.692	0.615	**0.846**	0.538	0.462	**0.769**	0.692
Imp. 3	0.538	**0.692**	0.769	**0.846**	0.615	0.692	0.538
Imp. 4	0.692	**0.692**	**0.846**	**0.846**	0.538	0.692	0.692
Imp. 5	**0.769**	**0.692**	**0.846**	-	0.615	**0.769**	**0.769**
Imp. 6	0.462	0.538	0.769	-	**0.692**	0.615	0.462

Table 4.6: Sensitivity results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

For the same reasons, the correlation between AUC-PR and $F1_{score}$ with sensitivity was also expected to be bigger than the correlation between AUC and accuracy with sensitivity. While this was true to $F1_{score}$, AUC-PR had the worst correlation with sensitivity.

Given that the observed performances were so close to each other, paired DeLong's tests using a p-value of 0.05 were performed to determine whether the observed differences were statistically significant. Table 4.8 and 4.9 show the results among each classification model and imputation method, respectively. For readability reasons, the p-value was omitted and a check-mark was placed instead when the difference between the classifiers are statistically significant. Models which did not show any

	F1 score	Accuracy	Precision-Recall AUC	AUC	Sensitivity
F1 score		0.728	0.233	0.025	0.525
Accuracy			0.260	0.111	0.485
Precision-Recall AUC				0.774	0.036
AUC					0.070
Sensitivity					

Table 4.7: Kendall correlation between evaluation metrics calculated for six imputation methods and seven different models predicting the modified Rankin Scale at three months.

significant difference were also omitted.

Taking a closer look at each model in table 4.8 it can be concluded that the great majority of imputation methods are statistically equivalent. To the best of our knowledge, this can be explained by two facts: the elevated amount of missing values (93% of the dataset features had more than 30% of its data missing) and the small number of patients used to conduct this study. Previous studies [44, 54, 107] have shown the difference between the performance of the several imputation methods is small and small sample sizes only allow large differences to be detected [24]. Moreover, it is known imputation should be used carefully in datasets with over 25% of the data missing [44], the high proportion of missing data may introduce considerable bias resulting in too similar imputations.

On the other hand, in table 4.9, it can be concluded that the great majority of classification models are statistically different.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1						✓
Imp. 2						✓
Imp. 3						
Imp. 4						
Imp. 5						
Imp. 6						

(a) Logistic Regression L1-regularised.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1					✓	
Imp. 2					✓	
Imp. 3					✓	✓
Imp. 4					✓	
Imp. 5						✓
Imp. 6						

(b) Neural Network.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1					✓	✓
Imp. 2					✓	✓
Imp. 3					✓	
Imp. 4					✓	✓
Imp. 5						
Imp. 6						

(c) Classification And Regression Trees.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1						
Imp. 2					✓	✓
Imp. 3						
Imp. 4						
Imp. 5						
Imp. 6						

(d) k-Nearest Neighbours.

Table 4.8: Paired DeLong's test results for each model - prediction of modified Rankin Scale at three months. Note: The following models are not presented since there were no significant differences present: Support Vector Machine, Random Forest, Extreme Gradient Boosting. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation.

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN						✓	✓
Xgboost							
RF						✓	✓
SVM							
LR							

(a) Mode/Median Imputation (Imp. 1).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN				✓		✓	
Xgboost					✓		
RF						✓	
SVM							
LR							

(b) Mode/Median Imputation taking into account the dependence of a few variables (Imp. 2).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN				✓		✓	✓
Xgboost							
RF							
SVM							
LR							

(c) Hotdeck Imputation (Imp. 3).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN							
Xgboost							
RF							
SVM							
LR							

(d) K-Nearest Neighbours Imputation (Imp. 4).

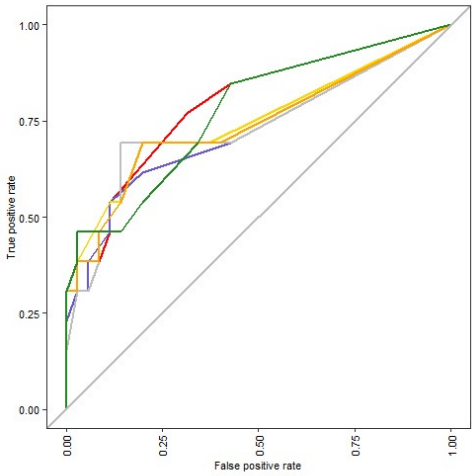
	KNN	CART	NN	RF	SVM	LR
KNN				✓	✓	✓
CART				✓	✓	✓
NN				✓		✓
RF						
SVM						
LR						

(e) Decision Trees Imputation (Imp. 5).

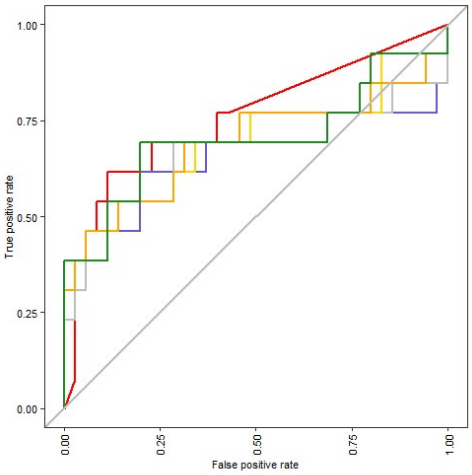
	KNN	CART	NN	RF	SVM	LR
KNN		✓	✓	✓	✓	✓
CART			✓	✓	✓	✓
NN						
RF						
SVM						
LR						

(f) Multiple Imputation (Imp. 6).

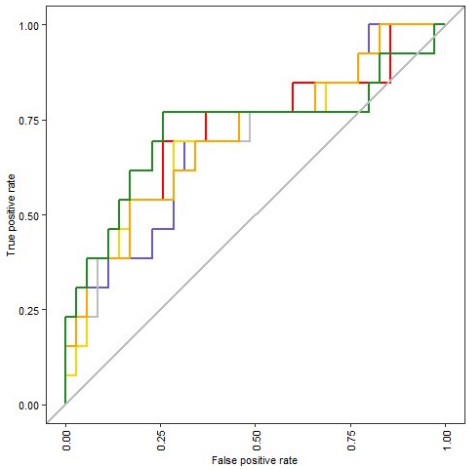
Table 4.9: Paired DeLong's test results for each imputation method - prediction of modified Rankin Scale at three months. The seven different methods used: KNN - k-Nearest Neighbours, CART - Classification And Regression Trees, NN - Neural Network, Xgboost - Extreme Gradient Boosting, RF - Random Forest, SVM - Support Vector Machines, LR - Logistic Regression.



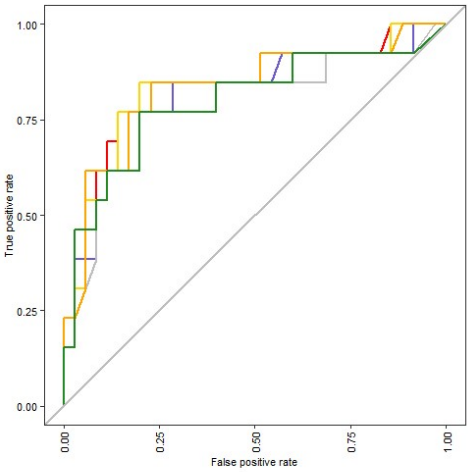
(a) k-Nearest Neighbours.



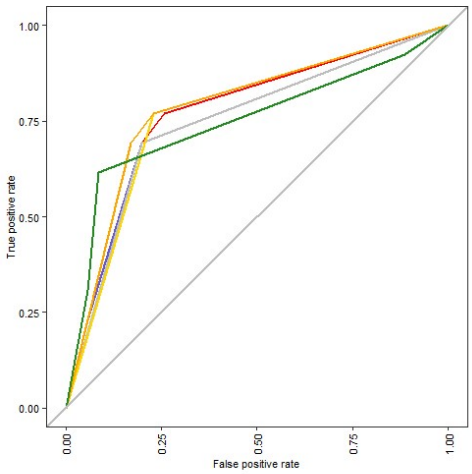
(b) Neural Network.



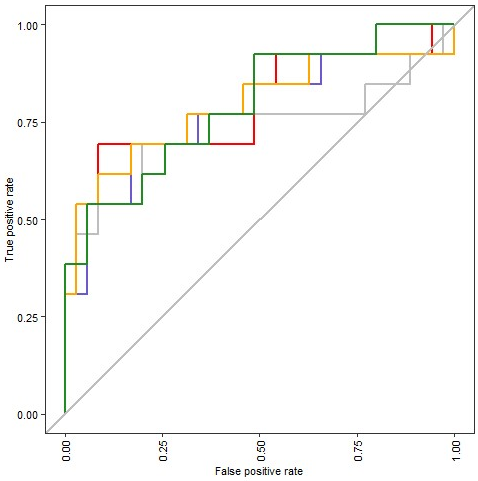
(c) Logistic Regression L1-regularised.



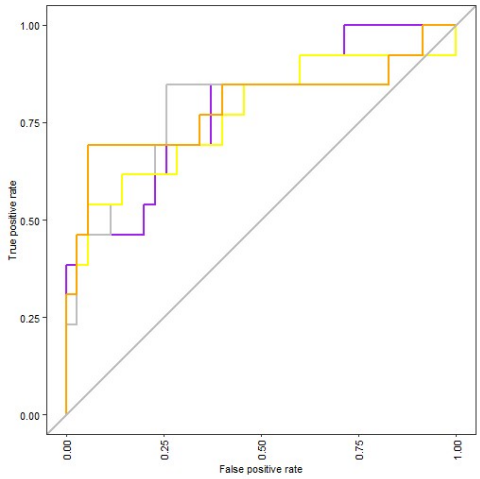
(d) Random Forest.



(e) Classification And Regression Trees.

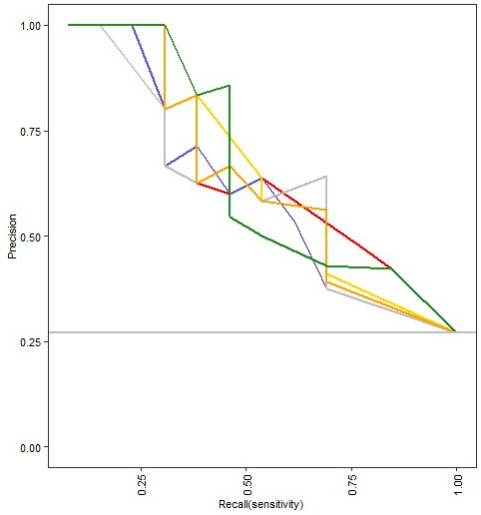


(f) Support Vector Machines.

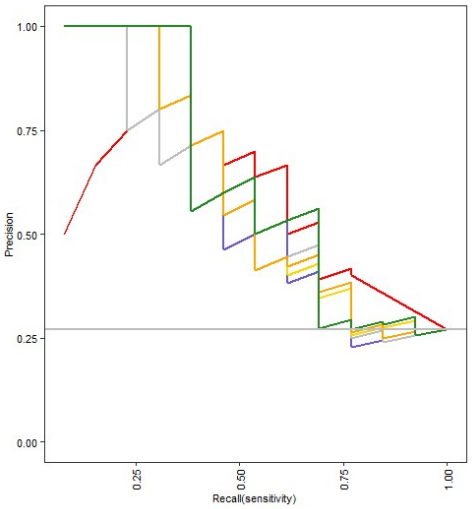


(g) Extreme Gradient Boosting.

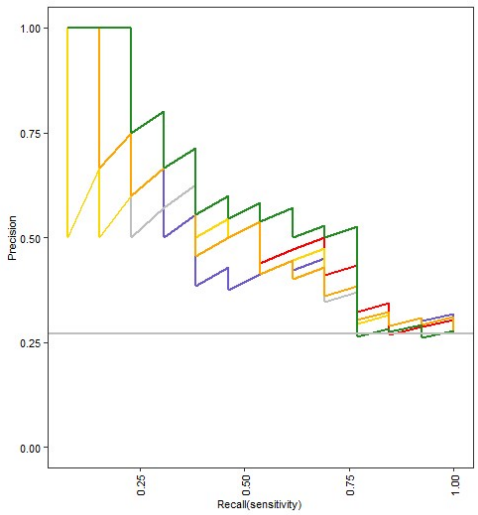
Figure 4.2: Receiver Operating Characteristic Curves presenting six imputation methods for seven different models predicting the modified Rankin Scale at three months. The colour scheme for the six different imputations used: Mode/Median Imputation - Yellow, Mode/Median Imputation taking into account the dependence of a few variables - Orange, Hotdeck Imputation - Purple, K-Nearest Neighbours Imputation - Grey, Decision Trees Imputation - Red, Multiple Imputation - Green.



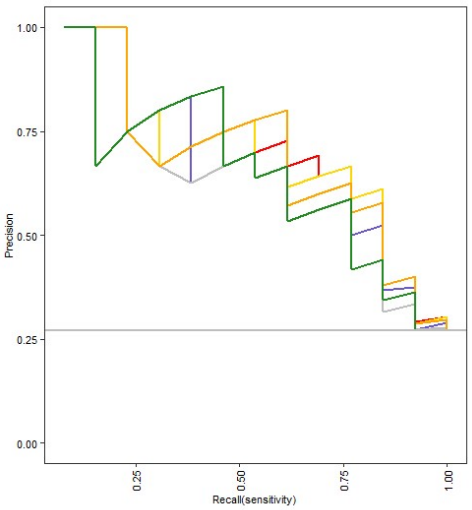
(a) k-Nearest Neighbours.



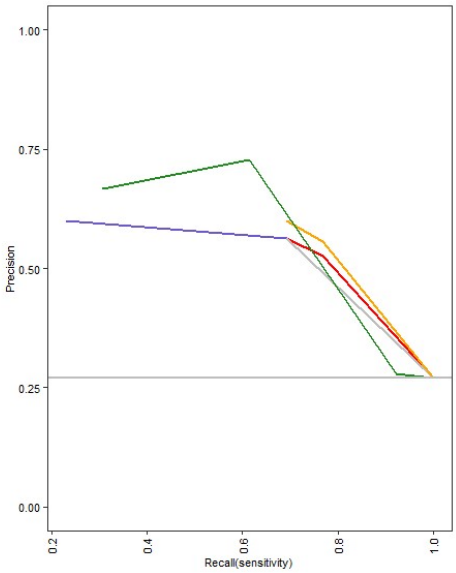
(b) Neural Network.



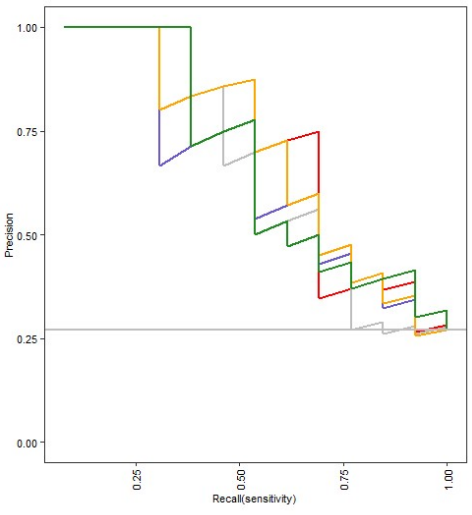
(c) Logistic Regression L1-regularised.



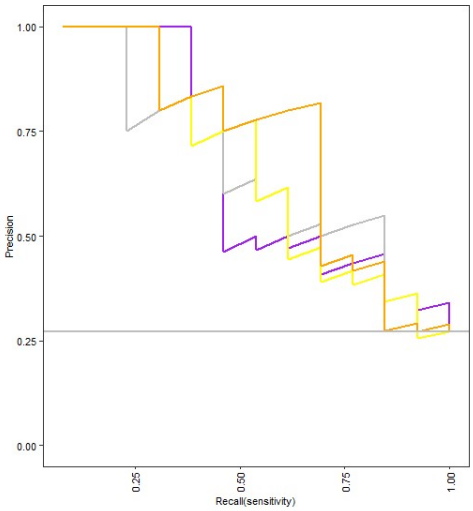
(d) Random Forest.



(e) Classification And Regression Trees.



(f) Support Vector Machines.



(g) Extreme Gradient Boosting.

Figure 4.3: Precision-Recall Curves presenting six imputation methods for seven different models predicting the modified Rankin Scale at three months. The colour scheme for the six different imputations used: Mode/Median Imputation - Yellow, Mode/Median Imputation taking into account the dependence of a few variables - Orange, Hotdeck Imputation - Purple, K-Nearest Neighbours Imputation - Grey, Decision Trees Imputation - Red, Multiple Imputation - Green.

4.2 Modified Rankin Scale at One Year

Contrarily to the results for the models predicting the modified Rankin Scale at three months, from table 4.10 the best imputation model can be chosen, hotdeck, which yields the best results for four out of the seven models supporting the previous conclusion that more complex methods aren't always the best option [107]. Here, similarly to the results at three months, the combination which achieved the best results was performing hotdeck imputation and using neural networks as the classification model with an AUC of 0.7537.

It was expected an increase in performance for the models predicting the modified Rankin Scale at one year when compared to predicting the modified Rankin Scale at three months since stroke symptoms are maximal on onset and decrease in severity with time. Moreover, as time goes by, the patients' state is less likely to significantly change, meaning, it was expected to be easier to predict the patients' functional outcome at one year based on their state at three months than their functional outcome at three months based on the patients' state a few days after stroke (when the symptoms are more likely to vary greatly on a daily basis) [69]. Comparing the AUC for both three months and one year, tables 4.1 and 4.10 respectively, it is seen this is not true. A possible explanation is that the added features have a portion of missing data too big that, when completed with the different imputation techniques, add noise rather than any relevant information resulting in the worsening of the AUC.

In Figure 4.4 can be found the twenty most important features and its relative importance (scale of 100%) for the best modified Rankin Scale classifier at one year (the most important features for the remaining classifiers can be found in Appendix B). The tendency seen at three months of variables corresponding to known predictors that are used by traditional scores is still present as anticipated. Only 4 of the 20 features were recorded at three months which is a lower number than expected given that, as mentioned above, the patients' situation is less likely to significantly change the more time has passed since the stroke occurred [69]. This corroborates the above hypothesis, data recorded at three months added more noise than information. Although the majority of the most important variables at one year were present at three months, the overlap between the variables at the two dates is very small.

Looking at table 4.11, the partial AUC values for an 80% sensitivity were computed and the correction by McClish was applied. As for the three months mark, the best imputation method is the same as when the total AUC is computed, table 4.10, as well as the best pair imputation method/classification model.

Furthermore, the different metrics AUC, AUC-PR, $F1_{score}$ and accuracy in tables 4.10, 4.12, 4.13 and 4.14, respectively, are again not in agreement when electing the best imputation method and classification model pair. For AUC-PR, $F1_{score}$ and accuracy the best imputation method would be k-nearest neighbours.

Looking at the ROC curves for the six imputation methods and seven different models predicting the modified Rankin Scale at one year, in figure 4.5, the same conclusion can be drawn as at three months: there is no clear distinction between them, the different classifier for each classification method are equivalent which is supported by the AUC values in table 4.10 and the statistical tests performed, table 4.17.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.7140 + 0.0244	0.7166 + 0.0239	0.6756 + 0.0231	0.7092 + 0.0244	0.7254 + 0.0241	0.6680 + 0.0247	0.6982 + 0.0235
Imp. 2	0.7103 + 0.0244	0.7188 + 0.0239	0.6632 + 0.0230	0.7067 + 0.0243	0.7488 + 0.0235	0.6341 + 0.0242	**0.6989 + 0.0235**
Imp. 3	**0.7295 + 0.0242**	**0.7296 + 0.0239**	0.6683 + 0.0230	**0.7320 + 0.0242**	**0.7537 + 0.0232**	0.6739 + 0.0247	0.6966 + 0.0233
Imp. 4	0.7236 + 0.0243	0.7236 + 0.0239	0.6802 + 0.0233	0.7075 + 0.0244	0.7358 + 0.0238	0.6297 + 0.0241	**0.6989 + 0.0235**
Imp. 5	0.7278 + 0.0242	0.7233 + 0.0241	**0.6826 + 0.0232**	-	0.6857 + 0.0248	0.6782 + 0.0247	0.6885 + 0.0238
Imp. 6	0.7171 + 0.0242	0.6959 + 0.0239	0.6687 + 0.0231	-	0.7146 + 0.0244	**0.6861 + 0.0245**	0.6950 + 0.0240

Table 4.10: AUC results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.8062	0.8117	0.7564	0.7999	0.8078	0.7427	0.7877
Imp. 2	0.8013	0.8147	0.7388	0.7970	0.8357	0.6957	**0.7888**
Imp. 3	**0.8264**	**0.8288**	0.7460	**0.8298**	**0.8374**	0.7508	0.7857
Imp. 4	0.8187	0.8211	0.7626	0.7973	0.8186	0.6894	**0.7888**
Imp. 5	0.8247	0.8198	**0.7662**	-	0.7663	0.7569	0.7735
Imp. 6	0.8108	0.7837	0.7465	-	0.8065	**0.7679**	0.7822

Table 4.11: Partial AUC results, 80% sensitivity, for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

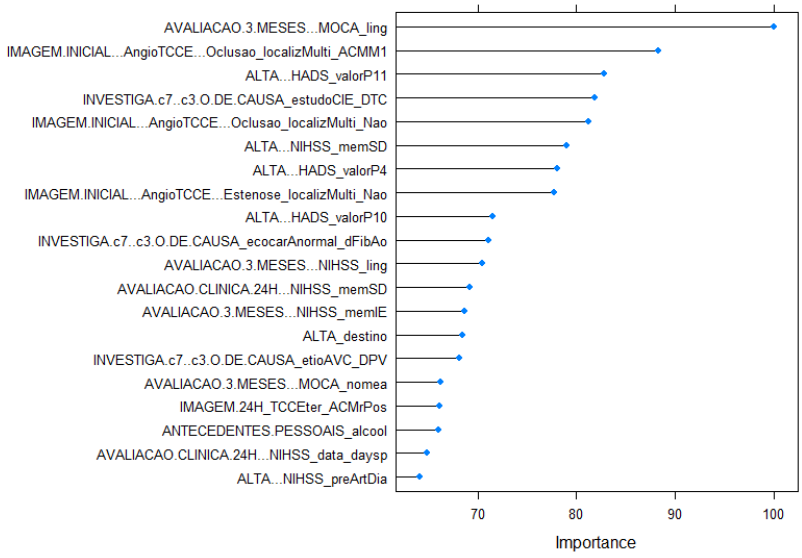


Figure 4.4: The twenty most important variables and its relative importance (scale of 100%) for the best model predicting the modified Rankin Scale at one year: hotdeck imputation and neural network classification.

As at the three-month timeline, the precision-recall curves are zigzagged, figure 4.6, complicating the choice of the best classifier. As before, uncommon situations are presented, for example, in subfigure 4.6a where a curve starts at coordinates (0,0). The situation is similar to the one at three months, again the classifier with the decision tree imputation, represented in red in subfigure 4.6a, is not able to avoid false positives at any level of probability threshold, at the lowest level of probability there is one false positive, meaning the first point does not need to be estimated and corresponds to (0.0, 0.0) [17, 48].

Looking at tables 4.10 and 4.12, AUC and AUC-PR, the range of values is bigger for AUC-PR than

for AUC, as expected alike the results at three months. Contrary to what happens when using the AUC metric, table 4.10, which has hotdeck as the best imputation method, when the AUC-PR is chosen multiple imputation is the best method, yielding best results for three out of the seven models.

Like at three months, the best pair imputation method/classification model also differ. Here, for the AUC-PR evaluation metric, there is a tie for the best pair: multiple imputation/logistic regression and mode/median imputation with dependent variables/extreme gradient boosting.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.560	0.610	0.670	0.590	0.080	0.340	0.590
Imp. 2	0.610	0.600	0.690	**0.710**	0.100	0.340	0.580
Imp. 3	0.680	0.540	0.580	0.620	0.120	**0.430**	0.460
Imp. 4	0.630	0.630	**0.700**	0.600	0.250	0.340	**0.650**
Imp. 5	0.520	0.520	0.610	-	0.480	0.250	0.530
Imp. 6	**0.710**	**0.640**	0.670	-	**0.550**	0.380	0.370

Table 4.12: AUC-PR results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	**0.769**	**0.750**	**0.727**	**0.769**	**0.741**	0.513	0.571
Imp. 2	0.741	**0.750**	0.667	0.759	0.696	0.513	0.571
Imp. 3	0.688	0.720	0.690	0.710	0.645	0.593	0.600
Imp. 4	0.741	0.720	0.690	**0.769**	**0.741**	**0.621**	**0.643**
Imp. 5	0.667	0.692	0.688	-	0.640	0.571	0.516
Imp. 6	0.733	0.611	**0.727**	-	0.571	0.583	0.552

Table 4.13: $F1_{score}$ results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.8478	**0.8261**	0.8261	**0.8696**	0.7609	0.7174	0.8043
Imp. 2	0.8696	**0.8261**	0.8261	0.8261	0.6957	0.7174	0.8043
Imp. 3	0.8478	0.8043	0.8261	0.8043	0.7609	0.7826	**0.8261**
Imp. 4	**0.8913**	**0.8261**	**0.8478**	0.8043	0.8043	**0.8261**	**0.8261**
Imp. 5	0.8261	**0.8261**	0.8261	-	0.7826	0.8043	0.7826
Imp. 6	0.8478	**0.8261**	0.8261	-	**0.8478**	0.7174	0.8043

Table 4.14: Accuracy results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.909	0.818	0.727	0.909	**0.909**	**0.909**	**0.909**
Imp. 2	0.909	0.818	0.909	**1.000**	0.727	**0.909**	**0.909**
Imp. 3	**1.000**	0.818	0.909	**1.000**	**0.909**	0.727	0.545
Imp. 4	0.909	0.818	0.909	0.909	**0.909**	0.818	0.818
Imp. 5	**1.000**	0.818	**1.000**	-	0.727	0.727	0.727
Imp. 6	**1.000**	**1.000**	0.727	-	0.545	0.636	0.727

Table 4.15: Sensitivity results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

Kendall's correlation between the different evaluation metrics can be seen in table 4.16. As expected AUC-PR and $F1_{score}$ have a higher correlation than AUC-PR and AUC, however, an unanticipated lower correlation than AUC-PR and accuracy. Additionally, as in the results for the models predicting the modified Rankin Scale at three months, the correlation between accuracy and AUC was also smaller than the correlation of accuracy with the remaining evaluation metrics which was also unexpected.

Finally, the correlation between AUC-PR and $F1_{score}$ with sensitivity was also expected to be bigger than the correlation between AUC and accuracy with sensitivity. While this was not true for the results at three months, at one year these correlations are now as anticipated.

	F1 score	Accuracy	Precision-Recall AUC	AUC	Sensitivity
F1 score		0.488	0.307	0.478	0.362
Accuracy			0.720	0.115	0.211
Precision-Recall AUC				0.100	0.286
AUC					0.167
Sensitivity					

Table 4.16: Kendall correlation between metrics calculated for six imputation methods and seven different models predicting the modified Rankin Scale at one year.

Table 4.17 and 4.18 show the results of the paired DeLong's tests, using a p-value of 0.05, among each classification model and imputation method, respectively.

Taking a closer look at each model in table 4.17 it can be concluded that the great majority of imputation methods are statistically equivalent. The same happened for the results at three months and both outcomes can be explained by the facts enumerated before: the elevated amount of missing values and the small number of patients used to conduct this study.

On the other hand, in table 4.18, contrarily to what happens for the results at three months, it is seen greater statistical equivalence between classification models. Again, a possible explanation lies in the added features: adding it, with its big portion of missing data might have resulted in the addiction of noise rather than any relevant information.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1					✓	
Imp. 2					✓	✓
Imp. 3					✓	✓
Imp. 4					✓	
Imp. 5						
Imp. 6						

(a) Neural Network

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1				✓		
Imp. 2			✓		✓	✓
Imp. 3				✓		
Imp. 4					✓	✓
Imp. 5						
Imp. 6						

(b) Classification And Regression Trees

Table 4.17: Paired DeLong's test results for each model - prediction of modified Rankin Scale at one year. Note: The following models are not presented since there were no significant differences present: Logistic Regression, Support Vector Machine, Random Forest, Extreme Gradient Boosting and k-Nearest Neighbours. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation.

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN							
CART			✓	✓		✓	✓
NN					✓		
Xgboost							
RF						✓	✓
SVM							
LR							

(a) Mode/Median Imputation (Imp. 1).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓		✓		
CART			✓	✓		✓	✓
NN				✓	✓		✓
Xgboost					✓		
RF						✓	✓
SVM							
LR							

(b) Mode/Median Imputation taking into account the dependence of a few variables (Imp. 2).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN			✓	✓			
CART			✓	✓		✓	✓
NN					✓		
Xgboost					✓		
RF						✓	✓
SVM							
LR							

(c) Hotdeck Imputation (Imp. 3).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓				
CART			✓	✓	✓	✓	✓
NN					✓		
Xgboost							
RF						✓	✓
SVM							
LR							

(d) K-Nearest Neighbours Imputation (Imp. 4).

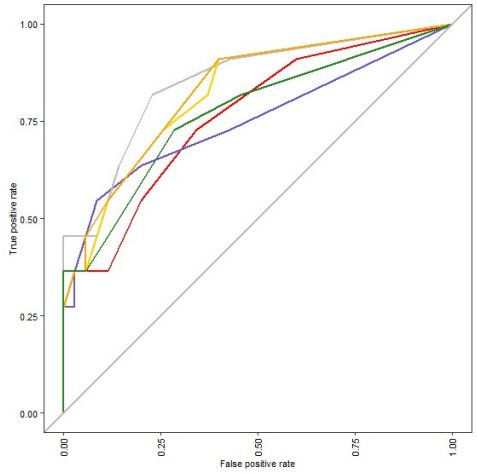
	KNN	CART	NN	RF	SVM	LR
KNN					✓	✓
CART					✓	✓
NN					✓	✓
RF					✓	✓
SVM						
LR						

(e) Decision Trees Imputation (Imp. 5).

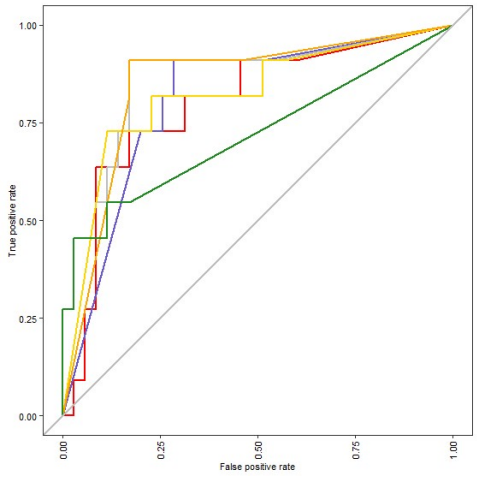
	KNN	CART	NN	RF	SVM	LR
KNN						
CART						
NN				✓		
RF					✓	
SVM						
LR						

(f) Multiple Imputation (Imp. 6).

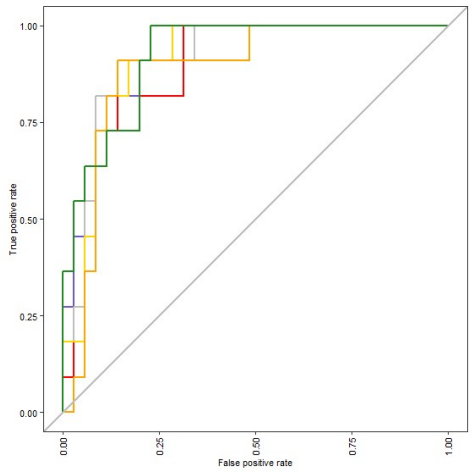
Table 4.18: Paired DeLong's test results for each imputation method - prediction of modified Rankin Scale at one year. The seven different methods used: KNN - k-Nearest Neighbours, CART - Classification And Regression Trees, NN - Neural Network, Xgboost - Extreme Gradient Boosting, RF - Random Forest, SVM - Support Vector Machines, LR - Logistic Regression.



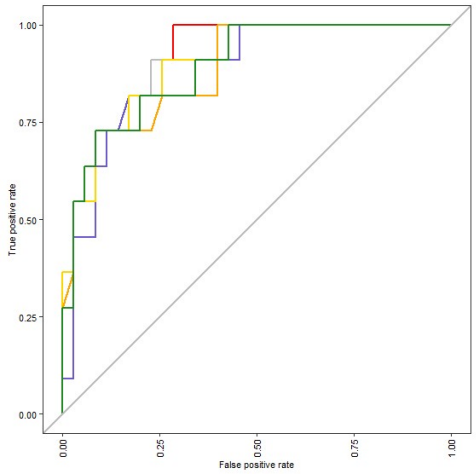
(a) k-Nearest Neighbours.



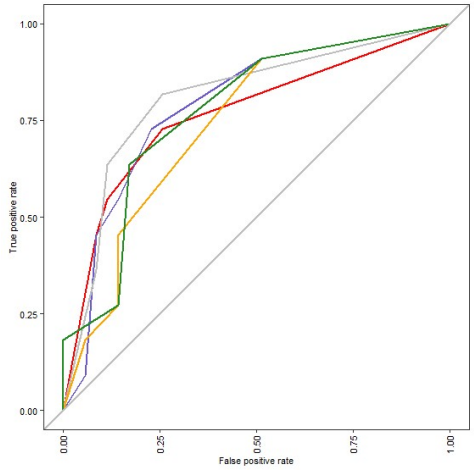
(b) Neural Network.



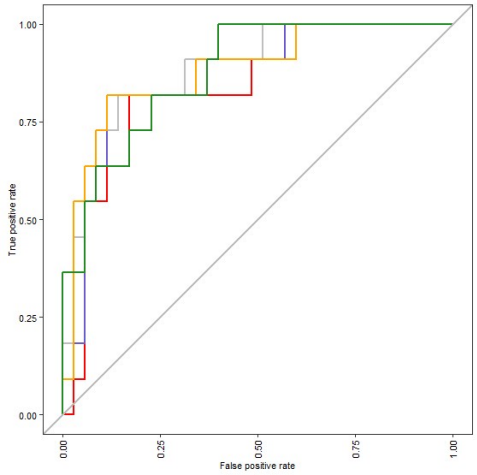
(c) Logistic Regression L1-regularised.



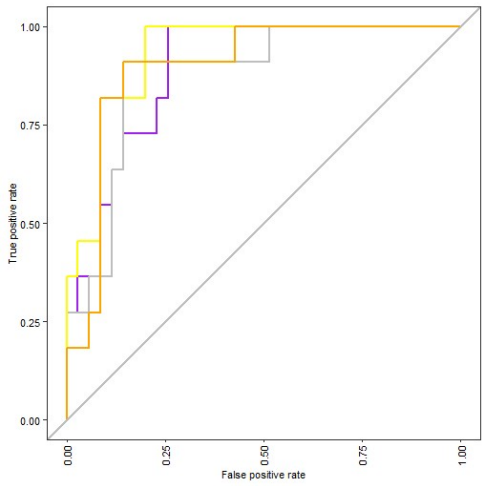
(d) Random Forest.



(e) Classification And Regression Trees.

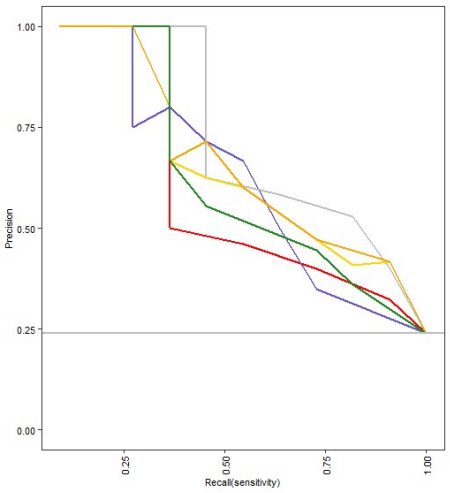


(f) Support Vector Machines.

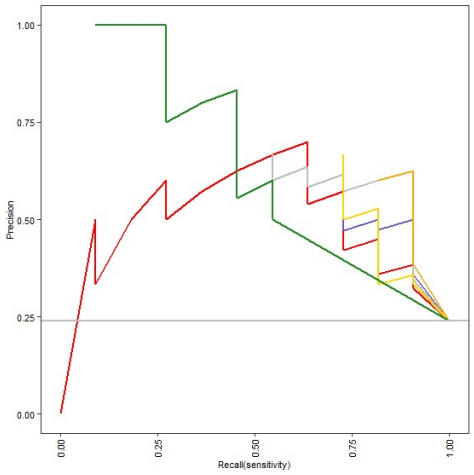


(g) Extreme Gradient Boosting.

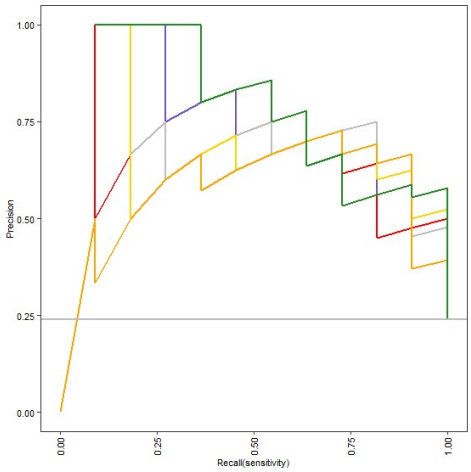
Figure 4.5: Receiver Operating Characteristic Curves presenting six imputation methods for seven different models predicting the modified Rankin Scale at one year. The colour scheme for the six different imputations used: Mode/Median Imputation - Yellow, Mode/Median Imputation taking into account the dependence of a few variables - Orange, Hotdeck Imputation - Purple, K-Nearest Neighbours Imputation - Grey, Decision Trees Imputation - Red, Multiple Imputation - Green.



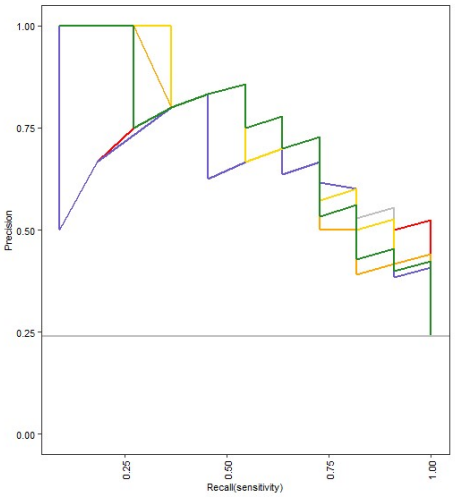
(a) k-Nearest Neighbours.



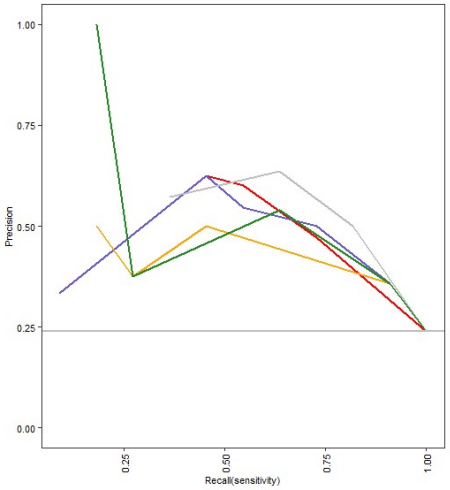
(b) Neural Network.



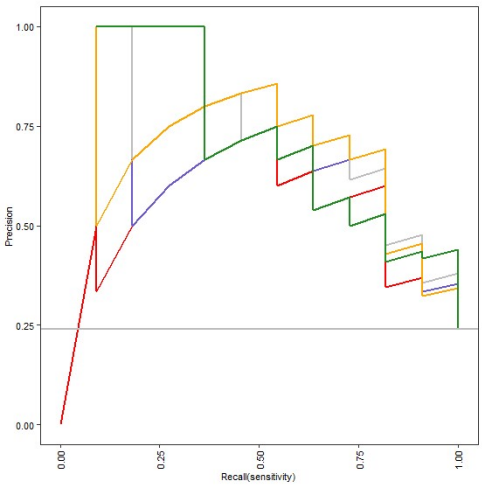
(c) Logistic Regression L1-regularised.



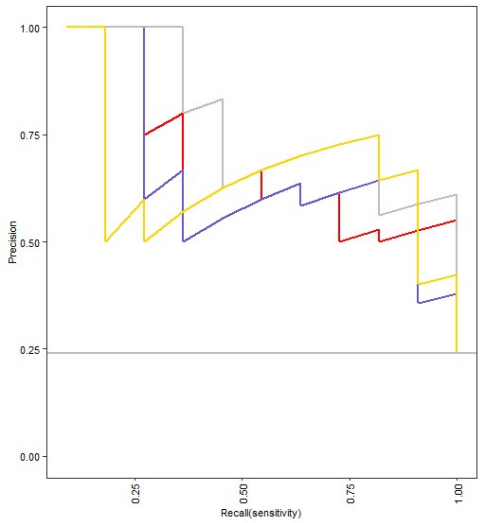
(d) Random Forest.



(e) Classification And Regression Trees.



(f) Support Vector Machines.



(g) Extreme Gradient Boosting.

Figure 4.6: Precision-Recall Curves presenting six imputation methods for seven different models predicting the modified Rankin Scale at one year. The colour scheme for the six different imputations used: Mode/Median Imputation - Yellow, Mode/Median Imputation taking into account the dependence of a few variables - Orange, Hotdeck Imputation - Purple, K-Nearest Neighbours Imputation - Grey, Decision Trees Imputation - Red, Multiple Imputation - Green.

Chapter 5

Conclusions and Future Work

The quality of data is one of the main concerns of data scientists as the quality of the results of a machine learning algorithm depend on it. Missing data introduces ambiguity and most algorithms are not robust enough to handle it thus producing misleading conclusions. A common technique when dealing with missing values is data imputation, *i.e.*, replacing it with most plausible values.

In this master thesis it was predicted the functional outcome, by the binary version of the mRS at two points in time. Missing data was imputed with six different methods and the classifier was trained with seven distinct machine learning models.

It was possible to conclude that machine learning can indeed effectively predict the functional outcome of an ischemic stroke patient. The AUC for the three months and one-year mark are 0.8217 and 0.7537, respectively meaning more data does not necessarily imply better results. Moreover, although there is a clear distinction between classifiers trained with only different machine learning methods, the same cannot be said for classifiers trained with only different imputation methods. Finally, it was highlighted how important it is to choose the right evaluation metric according to the problem's specificity.

In the future we wish to improve our performances by using more and richer records from the Precise Stroke Database. By adding more patients and with less missing data we hope to be able to answer the question: which imputation method results better for electronic health records and does this answer depend on the machine learning method being used for training.

Bibliography

[1] Al-Rubaie, M. and J. M. Chang

2019. Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security and Privacy*, 17(2):49–58.

[2] Ash, J. S., M. Berg, and E. Coiera

2004. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. *Journal of the American Medical Informatics Association*, 11(2):104–112.

[3] Banerjee, S., L. Alvey, P. Brown, S. Yue, L. Li, and W. J. Scheirer

2020. An Assistive Computer Vision Tool to Automatically Detect Changes in Fish Behavior In Response to Ambient Odor. *bioRxiv*, P. 2020.09.01.277657.

[4] Bejnordi, B. E., M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. Van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. J. Lin, P. A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y. W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. A. Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio

2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association*, 318(22):2199–2210.

[5] Béjot, Y., A. Aouba, C. de Peretti, O. Grimaud, C. Aboa-Eboulé, F. Chin, F. Woimant, E. Jougla, and M. Giroud

2010. Time trends in hospital-referred stroke and transient ischemic attack: results of a 7-year nationwide survey in France. *Cerebrovascular diseases (Basel, Switzerland)*, 30(4):346–54.

[6] Béjot, Y., H. Bailly, J. Durier, and M. Giroud

2016. Epidemiology of stroke in Europe and trends for the 21st century. *La Presse Médicale*, 45(12):e391–e398.

[7] Béjot, Y., B. Daubail, A. Jacquin, J. Durier, G.-V. Osseby, O. Rouaud, and M. Giroud
2014. Trends in the incidence of ischaemic stroke in young adults between 1985 and 2011: the Dijon Stroke Registry. *Journal of neurology, neurosurgery, and psychiatry*, 85(5):509–13.

[8] Berner, E. S., G. D. Webster, A. A. Shugerman, J. R. Jackson, J. Algina, A. L. Baker, E. V. Ball, C. G. Cobbs, V. W. Dennis, E. P. Frenkel, L. D. Hudson, E. L. Mancall, C. E. Rackley, and O. D. Taunton
1994. Performance of Four Computer-Based Diagnostic Systems. *New England Journal of Medicine*, 330(25):1792–1796.

[9] Breiman, L.
2001. Random forests. *Machine Learning*, 45(1):5–32.

[10] Broderick, J. P., O. Adeoye, and J. Elm
2017. Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials.

[11] Brown, M. L. and J. F. Kros
2003. Data mining and the impact of missing data. *Industrial Management and Data Systems*, 103(8-9):611–621.

[12] Cai, X., O. Perez-Concha, E. Coiera, F. Martin-Sanchez, R. Day, D. Roffe, and B. Gallego
2016. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, 23(3):553–561.

[13] Carspecken, C. W., P. J. Sharek, C. Longhurst, and N. M. Pageler
2013. A clinical case of electronic health record drug alert fatigue: Consequences for patient outcome. *Pediatrics*, 131(6).

[14] Chang, K.-W., C.-J. Hsieh, and C.-J. Lin
2008. LIBLINEAR: A Library for Large Linear Classification Rong-En Fan Xiang-Rui Wang. Technical report.

[15] Chen, T. and C. Guestrin
2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, Pp. 785–794. Association for Computing Machinery.

[16] Cismondi, F., A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. N. Finkelstein
2013. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*, 58(1):63–72.

[17] Davis, J. and M. Goadrich
2006. The relationship between precision-recall and ROC curves. In *ACM International Conference Proceeding Series*, volume 148, Pp. 233–240.

[18] De Dombal, F. T., D. J. Leaper, J. R. Staniland, A. P. Mccann, and J. C. Horrocks
1972. Computer-aided Diagnosis of Acute Abdominal Pain. *British Medical Journal*, 2(5804):9–13.

[19] Delen, D., A. Oztekin, and Z. Kong

2010. A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artificial Intelligence in Medicine*, 49(1):33–42.

[20] DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson

1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837.

[21] Deo, R. C.

2015. Machine learning in medicine. *Circulation*, 132(20):1920–1930.

[22] Dietterich, T. G.

1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923.

[23] Drozdowska, B. A., S. Singh, and T. J. Quinn

2019. Thinking About the Future: A Review of Prognostic Scales Used in Acute Stroke. *Frontiers in Neurology*, 10(March).

[24] Ellis, P. D.

2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*, 1 edition.

[25] Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun

2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[26] Feigin, V. L., B. Norrving, and G. A. Mensah

2017. Global Burden of Stroke. *Circulation Research*, 120(3):439–448.

[27] Feigin, V. L., D. O. Wiebers, Y. P. Nikitin, W. M. O'Fallon, and J. P. Whisnant

1995. Stroke Epidemiology in Novosibirsk, Russia: A Population-Based Study. *Mayo Clinic Proceedings*, 70(9):847–852.

[28] Fernández, A., S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera

2018. *Learning from Imbalanced Data Sets*, 1 edition. Springer International Publishing.

[29] Flint, A. C., S. P. Cullen, B. S. Faigeles, and V. A. Rao

2010. Predicting long-term outcome after endovascular stroke treatment: The totaled health risks in vascular events score. *American Journal of Neuroradiology*, 31(7):1192–1196.

[30] Folstein, M. F., S. E. Folstein, and P. 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):189–198.

[31] Friedman, J., T. Hastie, and R. Tibshirani
2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[32] Gao, X., Y. Uchiyama, X. Zhou, T. Hara, T. Asano, and H. Fujita
2011. A fast and fully automatic method for cerebrovascular segmentation on time-of-flight (TOF) MRA image. *Journal of digital imaging*, 24(4):609–25.

[33] Gelman, A. and J. Hill
2011. Opening Windows to the Black Box. *Journal of Statistical Software*, 40.

[34] Ghazali, S. M., N. Shaadan, and Z. Idrus
2020. Missing data exploration in air quality data set using r-package data visualisation tools. *Bulletin of Electrical Engineering and Informatics*, 9(2):755–763.

[35] Giroud, M., M. Lemesle, C. Quantin, M. Vourch, F. Becker, C. Milan, P. Brunet-Lecomte, and R. Dumas
1997. A hospital-based and a population-based stroke registry yield different results: the experience in Dijon, France. *Neuroepidemiology*, 16(1):15–21.

[36] Gordon, A. D., L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone
1984. Classification and Regression Trees. *Biometrics*, 40(3):874.

[37] Graham, J. W.
2009. Missing data analysis: Making it work in the real world.

[38] Graham, J. W., S. M. Hofer, S. I. Donaldson, D. P. MacKinnon, and J. L. Schafer
2004. Analysis with missing data in prevention research. In *The science of prevention: Methodological advances from alcohol and substance abuse research.*, Pp. 325–366. American Psychological Association.

[39] Graham, J. W., S. M. Hofer, and D. P. MacKinnon
1996. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31(2):197–218.

[40] Graham, J. W., S. M. Hofer, and A. M. Piccinin
1994. Analysis with missing data in drug prevention research. In *NIDA Research Monograph Series*, volume 142, Pp. 13–63. NIDA Res Monogr.

[41] Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster
2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - Journal of the American Medical Association*, 316(22):2402–2410.

[42] Harrison, J. K., K. S. McArthur, and T. J. Quinn

2013. Assessment scales in stroke: Clinimetric and clinical considerations.

[43] He, Z. and W. Yu

2010. Stable feature selection for biomarker discovery.

[44] Jadhav, A., D. Pramod, and K. Ramanathan

2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10):913–933.

[45] James, G., D. Witten, T. Hastie, and R. Tibshirani

2000. *An introduction to Statistical Learning*, volume 7.

[46] Jha, S. and E. J. Topol

2016. Adapting to artificial intelligence: Radiologists and pathologists as information specialists.

[47] John, C. R.

2020. MLeval: Machine Learning Model Evaluation.

[48] Keilwagen, J., I. Grosse, and J. Grau

2014. Area under precision-recall curves for weighted and unweighted data. *PLoS ONE*, 9(3).

[49] Kotz, S., C. B. Read, N. Balakrishnan, and B. Vidakovic

2005. *Encyclopedia of Statistical Sciences*, volume 16, 2 edition.

[50] Kowarik, A. and M. Templ

2016. Imputation with the R package VIM. *Journal of Statistical Software*, 74(1):1–16.

[51] Kuhn, M.

2008. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26.

[52] Kundu, M., M. Nasipuri, and D. K. Basu

. Knowledge-based ECG interpretation: a critical review. *Pattern Recognition*, 33(3):351–373.

[53] Kwon, S., A. G. Hartzema, P. W. Duncan, and S. M. Lai

2004. Disability Measures in Stroke: Relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. *Stroke*, 35(4):918–923.

[54] Kyureghian, G., O. Capps, and R. M. Nayga

2011. A missing variable imputation methodology with an empirical application. *Advances in Econometrics*, 27 A:313–337.

[55] Lakhani, P. and B. Sundaram

2017. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582.

[56] Li, X., J. Dunn, D. Salins, G. Zhou, W. Zhou, S. M. Schüssler-Fiorenza Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, R. Sonecha, S. Datta, T. McLaughlin, and M. P. Snyder
2017. Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. *PLOS Biology*, 15(1):e2001402.

[57] Lin, H. T., C. J. Lin, and R. C. Weng
2007. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276.

[58] Litjens, G., C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak
2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6(1):1–11.

[59] LITTLE, R. J. A. and D. B. RUBIN
1989. The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, 18(2-3):292–326.

[60] Little, R. J. A. and D. B. Rubin
2002. *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

[61] Lundberg, S. and S.-I. Lee
2017a. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-December:4766–4775.

[62] Lundberg, S. M. and S.-I. Lee
2017b. Consistent feature attribution for tree ensembles.

[63] Lyden, P., T. Brott, B. Tilley, K. M. Welch, E. J. Mascha, S. Levine, E. C. Haley, J. Grotta, and J. Marler
1994. Improved reliability of the NIH stroke scale using video training. *Stroke*, 25(11):2220–2226.

[64] Majdani, O., T. S. Rau, S. Baron, H. Eilers, C. Baier, B. Heimann, T. Ortmaier, S. Bartling, T. Lenarz, and M. Leinung
2009. A robot-guided minimally invasive approach for cochlear implant surgery: Preliminary results of a temporal bone study. *International Journal of Computer Assisted Radiology and Surgery*, 4(5):475–486.

[65] Majumder, S., T. Mondal, and M. J. Deen
2017. Wearable sensors for remote health monitoring.

[66] Mbaabu, O.
2020. Introduction to Random Forest in Machine Learning.

[67] Medin, J., A. Nordlund, and K. Ekberg

2004. Increasing Stroke Incidence in Sweden Between 1989 and 2000 Among Persons Aged 30 to 65 Years: Evidence From the Swedish Hospital Discharge Register. *Stroke*, 35(5):1047–1051.

[68] Miller, R. A.

1994. Medical diagnostic decision support systems- Past, present, and future: A threaded bibliography and brief commentary.

[69] Monteiro, M., A. C. Fonseca, A. T. Freitas, T. Pinho E Melo, A. P. Francisco, J. M. Ferro, and A. L. Oliveira

2018. Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6):1953–1959.

[70] Moritz, S. and T. Bartz-Beielstein

2017. imputeTS: Time series missing value imputation in R. *R Journal*, 9(1):207–218.

[71] Musen, M. A., B. Middleton, and R. A. Greenes

2014. Clinical decision-support systems. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine: Fourth Edition*, Pp. 643–674. Springer London.

[72] Narayan, S., M. Gagné, and R. Safavi-Naini

2010. Privacy preserving ehr system using attribute-based infrastructure. In *Proceedings of the ACM Conference on Computer and Communications Security*, Pp. 47–52.

[73] Narkhede, S.

2018. Understanding AUC - ROC Curve.

[74] Nasreddine, Z. S., N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow

2005. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

[75] Ntaios, G., M. Faouzi, J. Ferrari, W. Lang, K. Vemmos, and P. Michel

2012. An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL score. *Neurology*, 78(24):1916–1922.

[76] Qayyum, A., J. Qadir, M. Bilal, and A. Al-Fuqaha

2020. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering*.

[77] Quang, D., Y. Chen, and X. Xie

2015. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763.

[78] Quang, D. and X. Xie

2016. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11).

[79] Raymond, M. R.

1986. Missing Data in Evaluation Research. *Evaluation & the Health Professions*, 9(4):395–420.

[80] Ripley, B. D.

2014. *Pattern recognition and neural networks*. Cambridge University Press.

[81] ROTH, P. L.

1994. MISSING DATA: A CONCEPTUAL REVIEW FOR APPLIED PSYCHOLOGISTS. *Personnel Psychology*, 47(3):537–560.

[82] Rubin, D. B.

1976. Inference and missing data. *Biometrika*, 63(3):581–592.

[83] Rubin, D. B., ed.

1987. *Multiple Imputation for Nonresponse in Surveys*, Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.

[84] Rumelhart, D. E., G. E. Hinton, and R. J. Williams

1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[85] Russell, S. J.

2010. *Artificial intelligence : a modern approach*. Prentice Hall.

[86] Saito, T. and M. Rehmsmeier

2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432.

[87] Salgado, C. M., C. Azevedo, H. Proença, and S. M. Vieira

2016. Missing data. In *Secondary Analysis of Electronic Health Records*, Pp. 143–162. Springer International Publishing.

[88] Saver, J. L., B. Filip, S. Hamilton, A. Yanes, S. Craig, M. Cho, R. Conwit, and S. Starkman

2010. Improving the reliability of stroke disability grading in clinical trials and clinical practice: The rankin focused assessment (RFA). *Stroke*, 41(5):992–995.

[89] Schafer, J. L.

1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.

[90] Schafer, J. L. and J. W. Graham

2002. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

[91] Schwartz, W. B., R. S. Patil, and P. Szolovits

1987. Artificial Intelligence in Medicine.

[92] Shademan, A., R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim
2016. Supervised autonomous robotic soft tissue surgery. *Science Translational Medicine*, 8(337):64–337.

[93] Sheridan, R. P., W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford
2016. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Modeling*, 56(12):2353–2360.

[94] Shortliffe, E. H.
1987. Computer Programs to Support Clinical Decision Making. *JAMA: The Journal of the American Medical Association*, 258(1):61–66.

[95] Shortliffe, E. H., R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen
1975. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, 8(4):303–320.

[96] Snaith, R. P.
2003. The hospital anxiety and depression scale. *Health and Quality of Life Outcomes*, 1:29.

[97] Stekhoven, D. J.
2013. missForest: Nonparametric Missing Value Imputation using Random Forest.

[98] Stekhoven, D. J. and P. Buehlmann
2012. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

[99] Strbian, D., A. Meretoja, F. J. Ahlhelm, J. Pitkäniemi, P. Lyrer, M. Kaste, S. Engelter, and T. Tatlisumak
2012. Predicting outcome of IV thrombolysis - Treated ischemic stroke patients: The DRAGON score. *Neurology*, 78(6):427–432.

[100] Sun, X. and W. Xu
2014. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393.

[101] Truelsen, T., B. Piechowski-Jozwiak, R. Bonita, C. Mathers, J. Bogousslavsky, and G. Boysen
2006. Stroke incidence and prevalence in Europe: a review of available data. *European Journal of Neurology*, 13(6):581–598.

[102] Tsikriktsis, N.
2005. A review of techniques for treating missing data in OM survey research. *Journal of Operations Management*, 24(1):53–62.

[103] van Ginneken, B., A. A. A. Setio, C. Jacobs, and F. Ciompi
2015. Off-the-shelf convolutional neural network features for pulmonary nodule detection in com-

puted tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Pp. 286–289. IEEE.

[104] Venables, W. and B. Ripley
2002. *Modern Applied Statistics with S*, 4 edition. Springer-Verlag New York.

[105] Wilson, J. T., A. Hareendran, M. Grant, T. Baird, U. G. Schulz, K. W. Muir, and I. Bone
2002. Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin Scale. *Stroke*, 33(9):2243–2246.

[106] Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal
2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Inc.

[107] Woźnica, K. and P. Biecek
2020. Does imputation matter? Benchmark for predictive models.

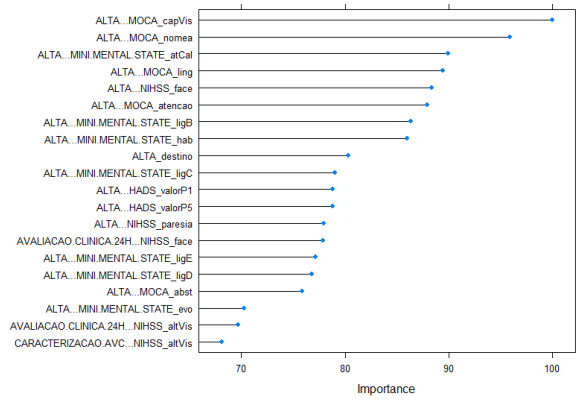
[108] Wu, T.-F., C.-J. Lin, and R. C. Weng
2004. Probability Estimates for Multi-class Classification by Pairwise Coupling. Technical report.

[109] Yu, K. H., A. L. Beam, and I. S. Kohane
2018. Artificial intelligence in healthcare.

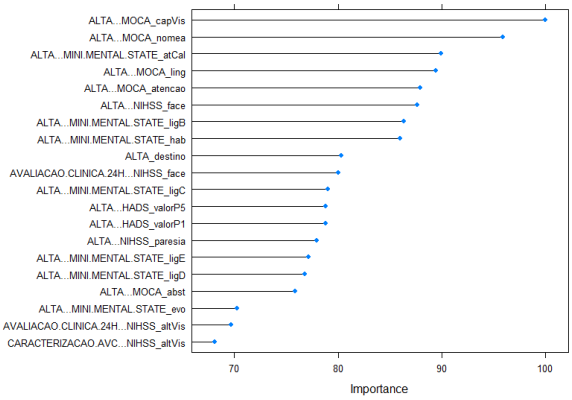
Appendix A

Modified Rankin Scale at Three Months

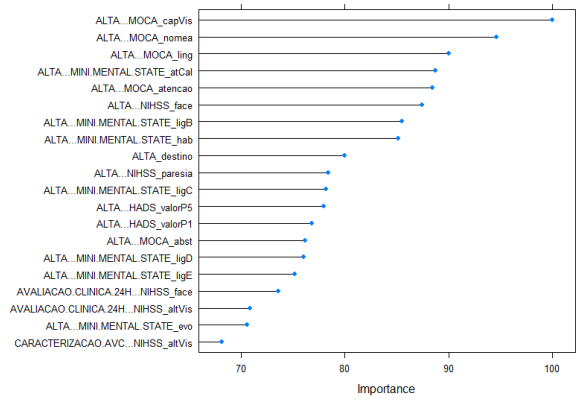
The twenty most important variables and its relative importance (scale of 100%) for each imputation method and for each model predicting the modified Rankin Scale at three months.



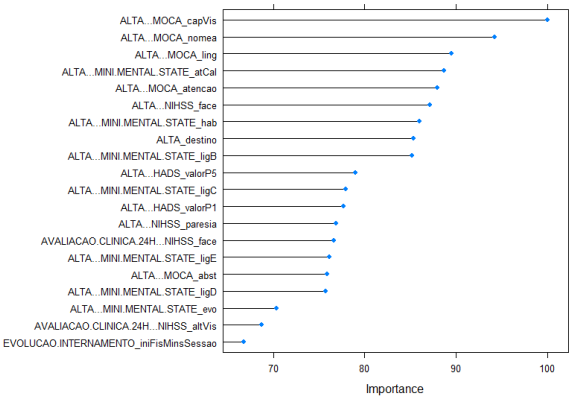
(a) Mode/Median Imputation



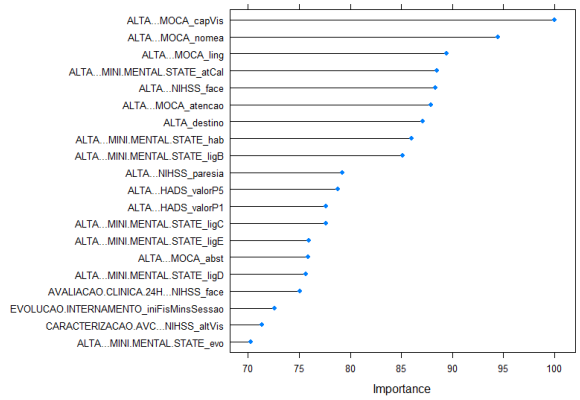
(b) Mode/Median Imputation taking into account the dependence of a few variables



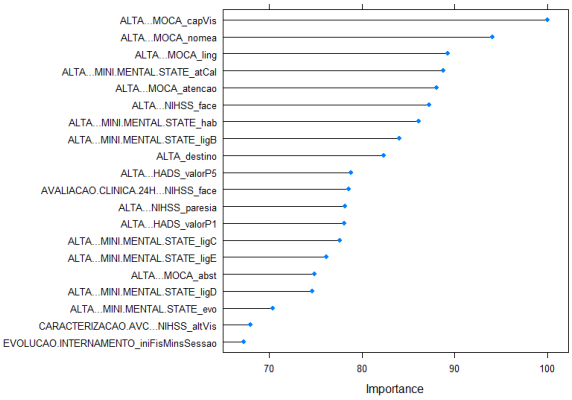
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

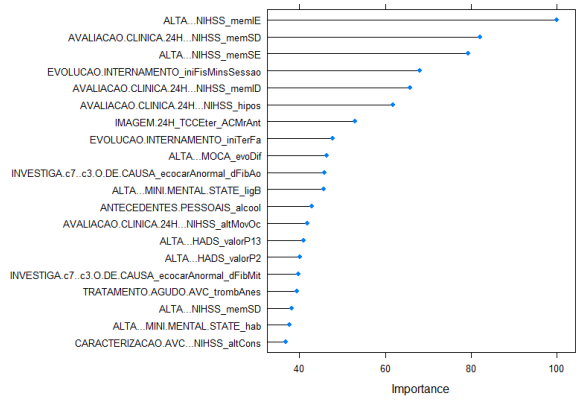


(e) Decision Trees Imputation

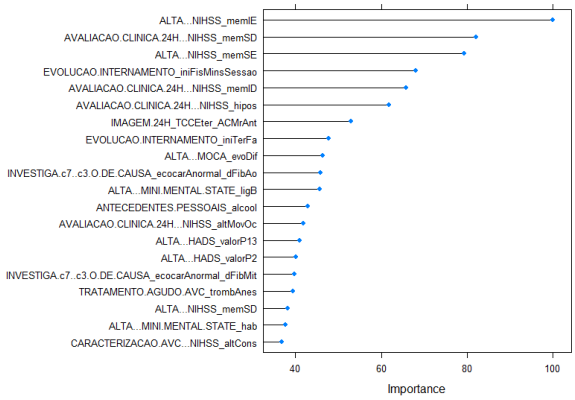


(f) Multiple Imputation

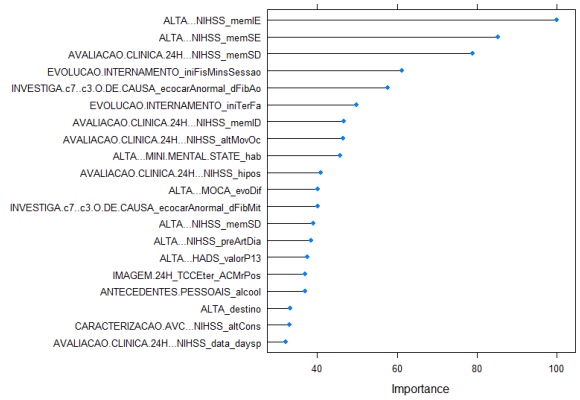
Figure A.1: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model k-Nearest Neighbours predicting the modified Rankin Scale at three months.



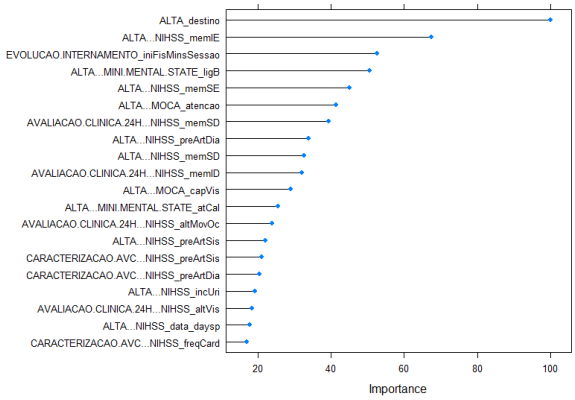
(a) Mode/Median Imputation



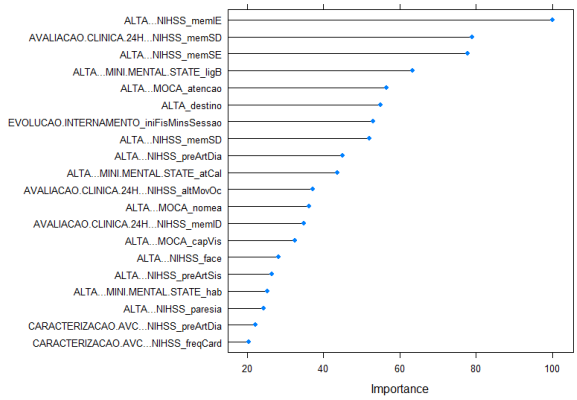
(b) Mode/Median Imputation taking into account the dependence of a few variables



(c) K-Nearest Neighbours Imputation

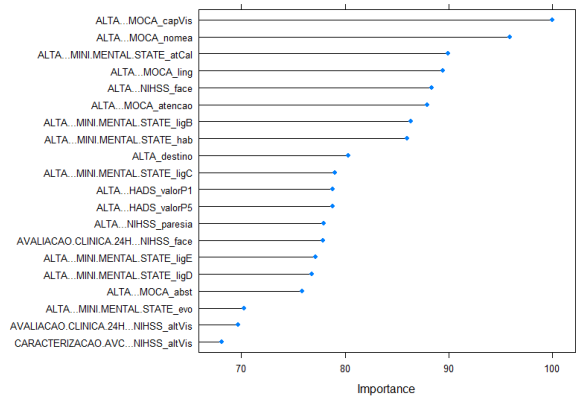


(d) Decision Trees Imputation

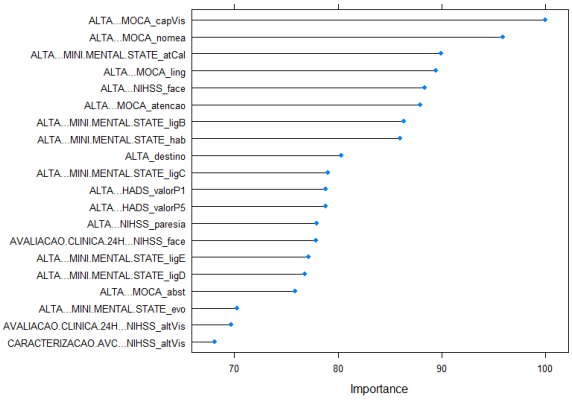


(e) Multiple Imputation

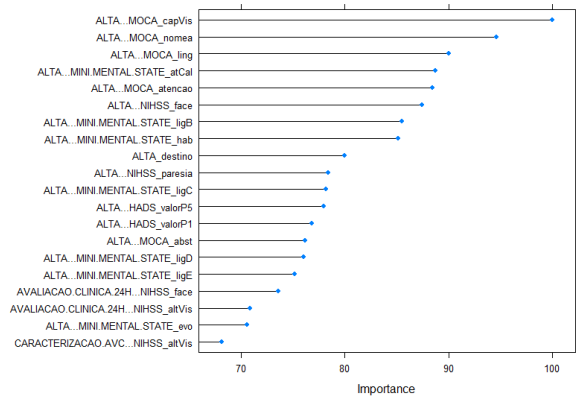
Figure A.2: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Neural Network predicting the modified Rankin Scale at three months.



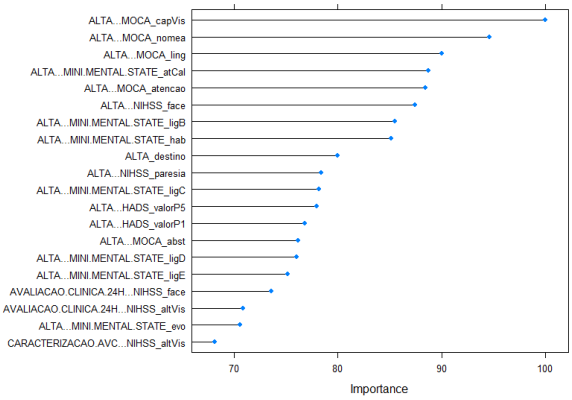
(a) Mode/Median Imputation



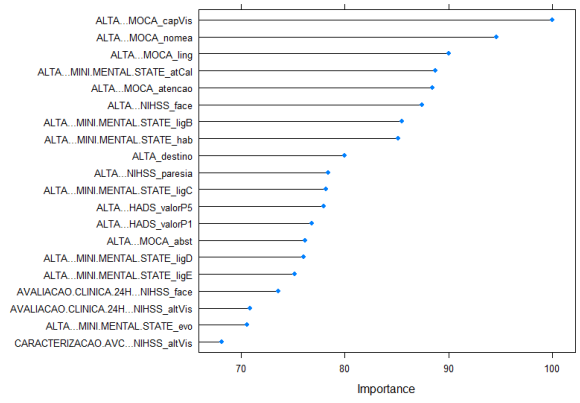
(b) Mode/Median Imputation taking into account the dependence of a few variables



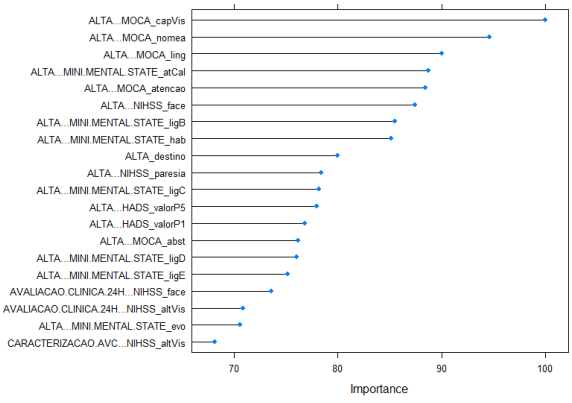
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

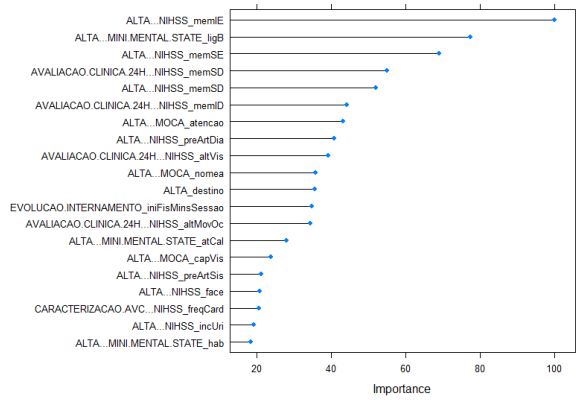


(e) Decision Trees Imputation

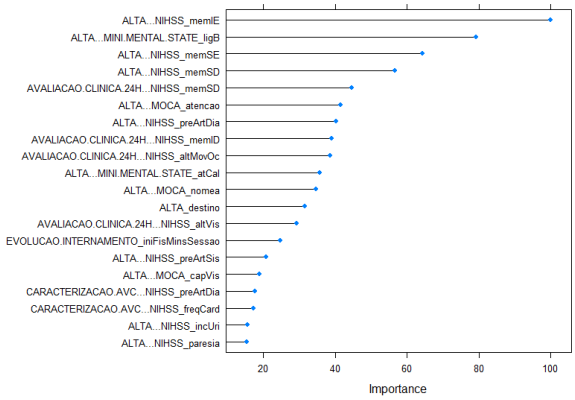


(f) Multiple Imputation

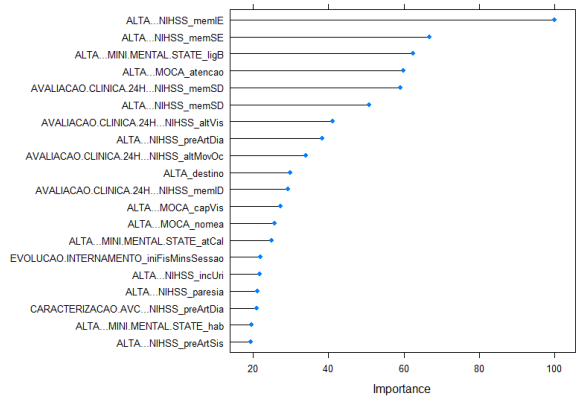
Figure A.3: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Logistic Regression L1-regularised predicting the modified Rankin Scale at three months.



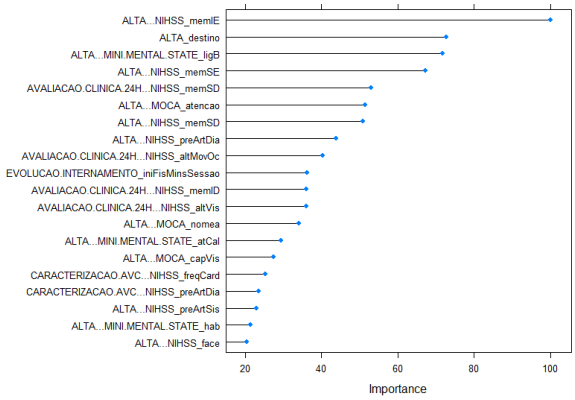
(a) Mode/Median Imputation



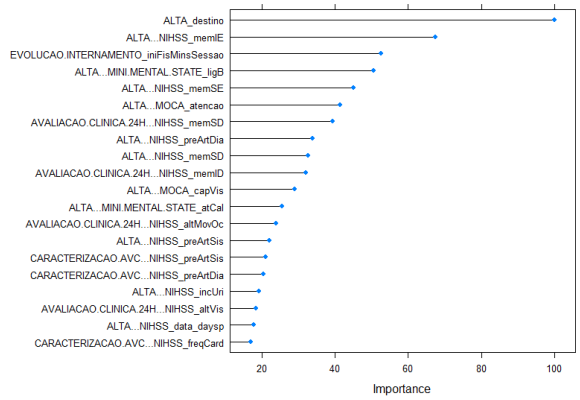
(b) Mode/Median Imputation taking into account the dependence of a few variables



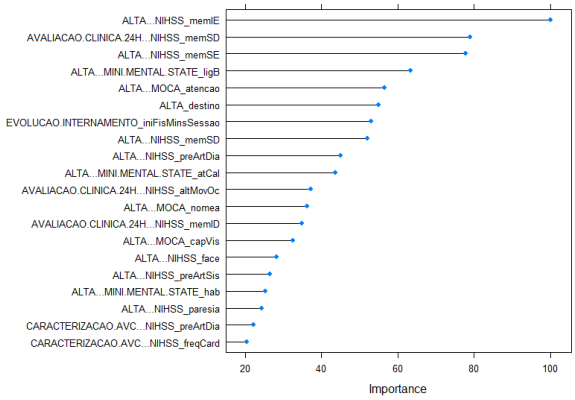
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

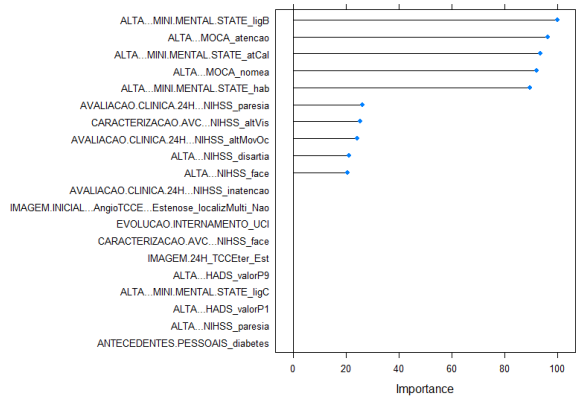


(e) Decision Trees Imputation

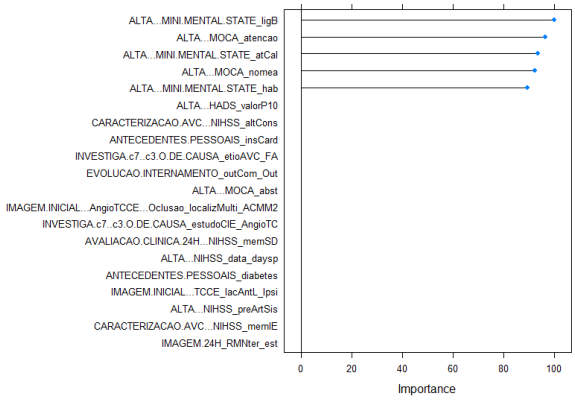


(f) Multiple Imputation

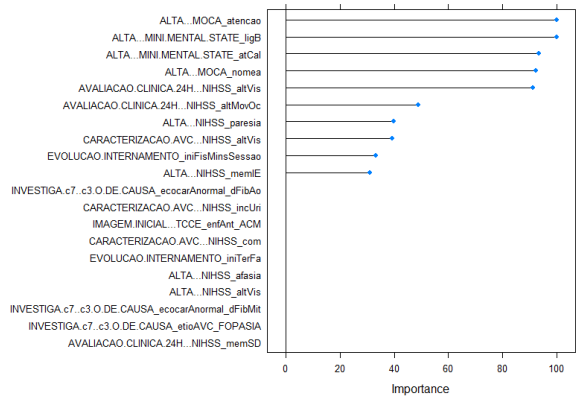
Figure A.4: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Random Forest predicting the modified Rankin Scale at three months.



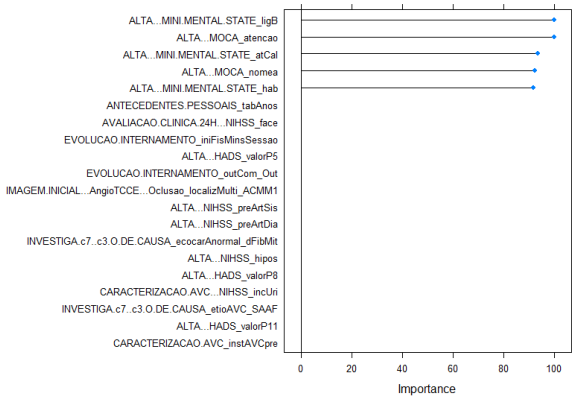
(a) Mode/Median Imputation



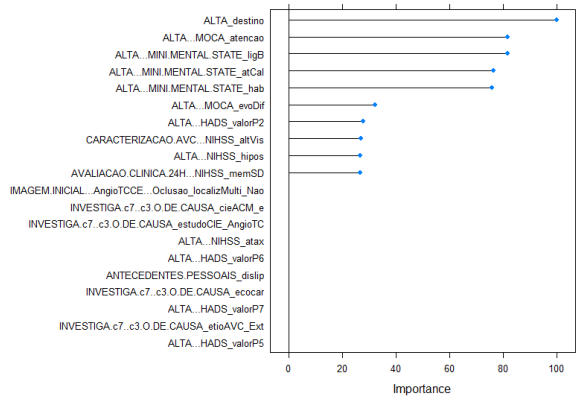
(b) Mode/Median Imputation taking into account the dependence of a few variables



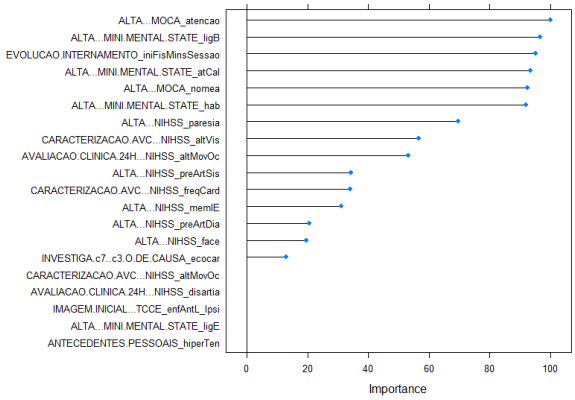
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

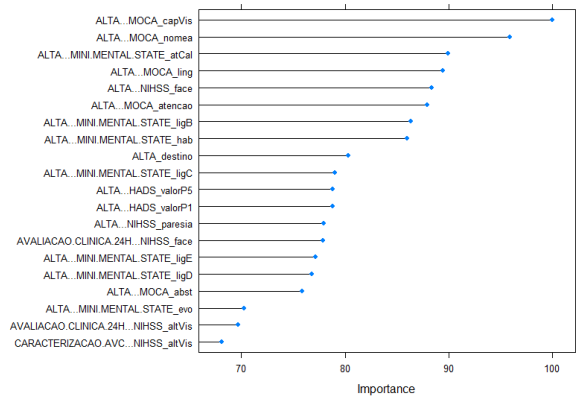


(e) Decision Trees Imputation

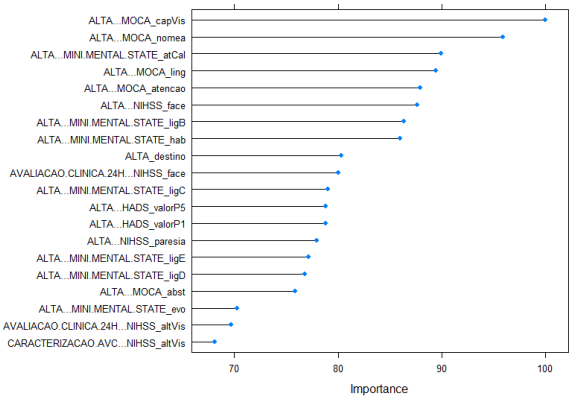


(f) Multiple Imputation

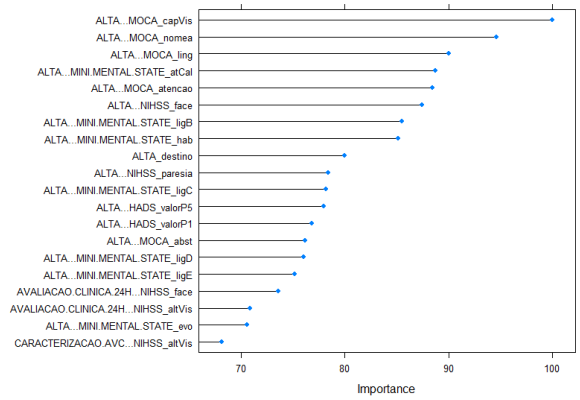
Figure A.5: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model CART predicting the modified Rankin Scale at three months.



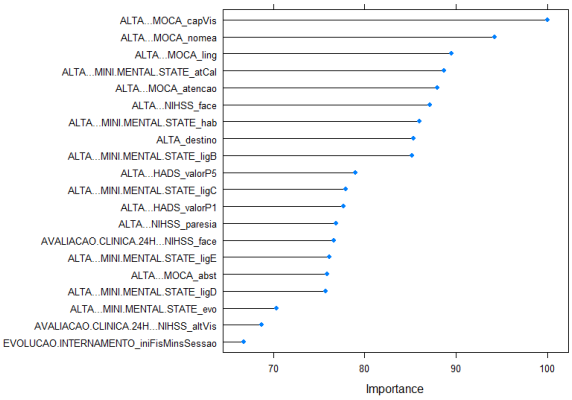
(a) Mode/Median Imputation



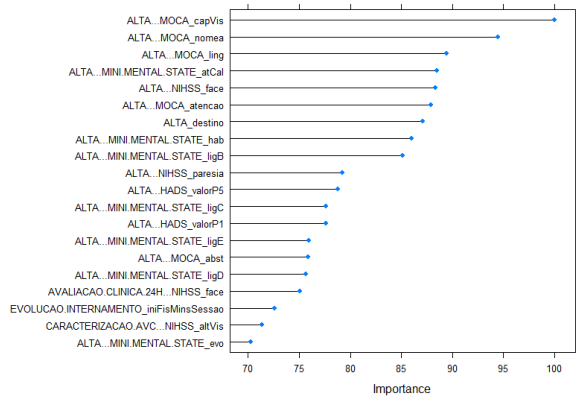
(b) Mode/Median Imputation taking into account the dependence of a few variables



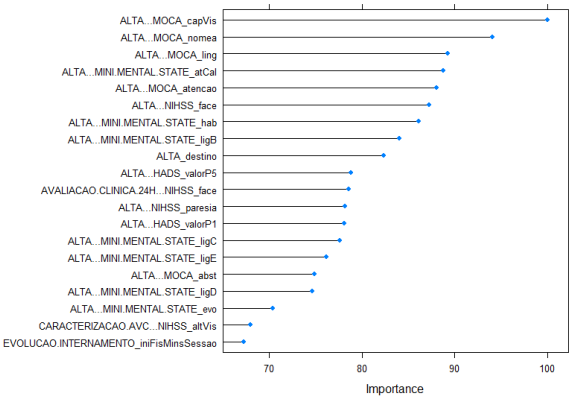
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

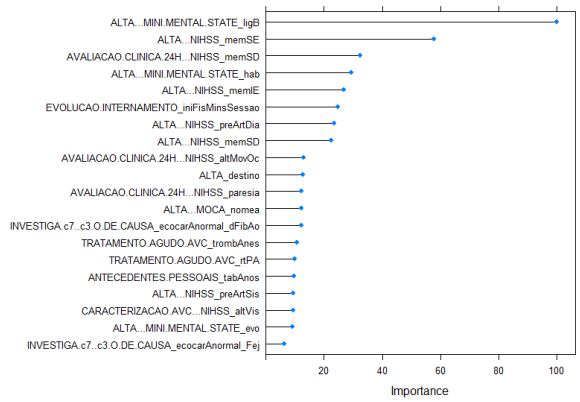


(e) Decision Trees Imputation

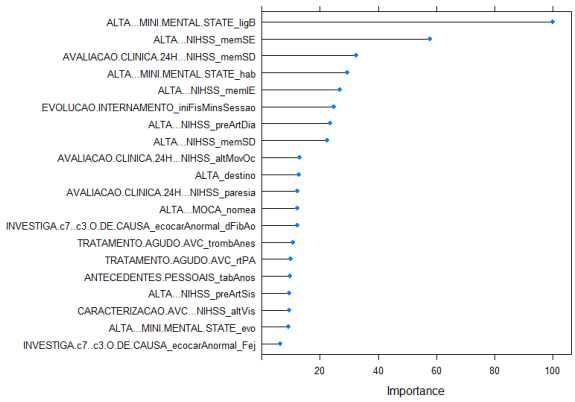


(f) Multiple Imputation

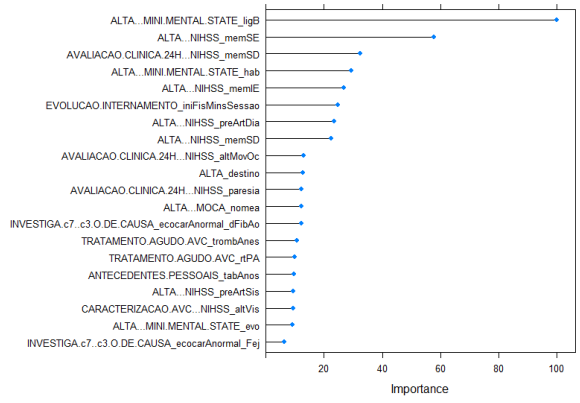
Figure A.6: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Support Vector Machines predicting the modified Rankin Scale at three months.



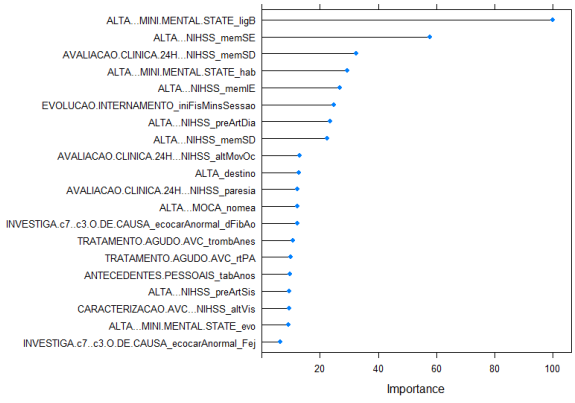
(a) Mode/Median Imputation



(b) Mode/Median Imputation taking into account the dependence of a few variables



(c) Hotdeck Imputation



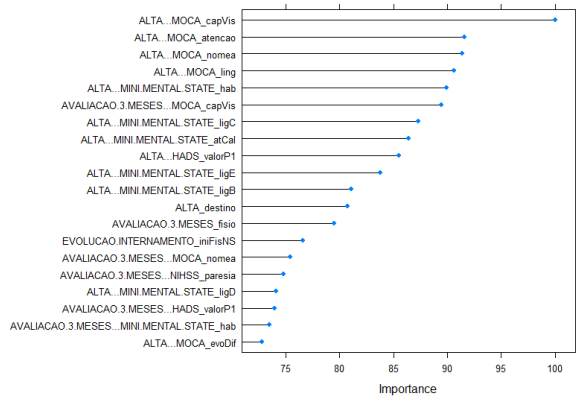
(d) K-Nearest Neighbours Imputation

Figure A.7: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Extreme Gradient Boosting predicting the modified Rankin Scale at three months.

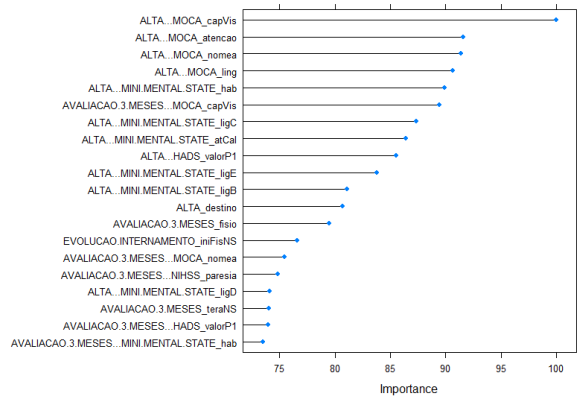
Appendix B

Modified Rankin Scale at One Year

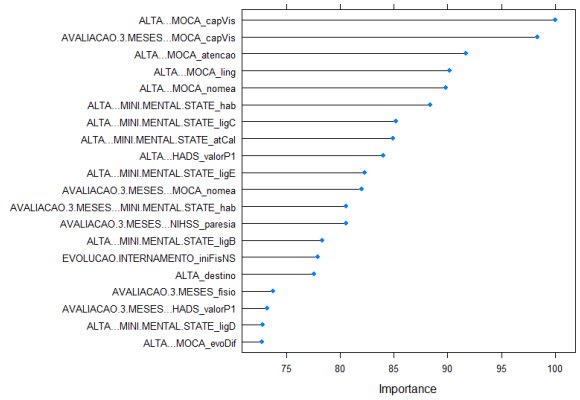
The twenty most important variables and its relative importance (scale of 100%) for each imputation method and for each model predicting the modified Rankin Scale at one year.



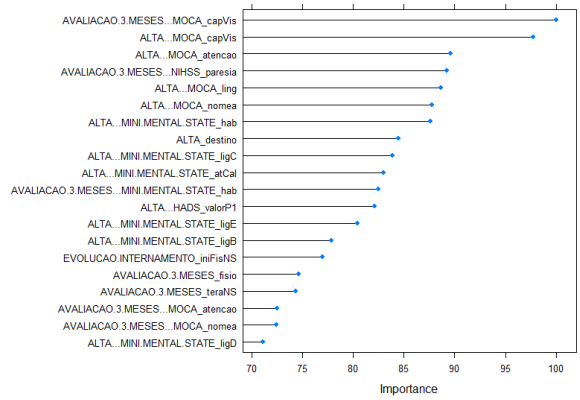
(a) Mode/Median Imputation



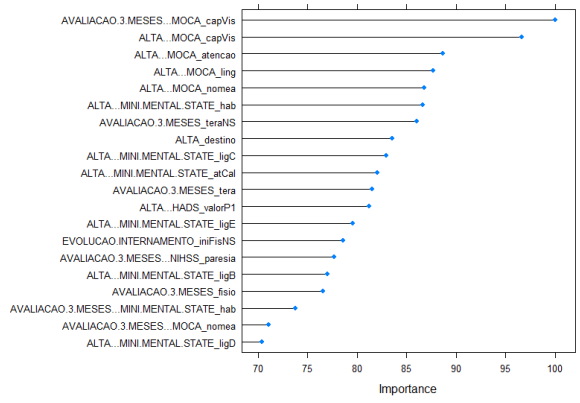
(b) Mode/Median Imputation taking into account the dependence of a few variables



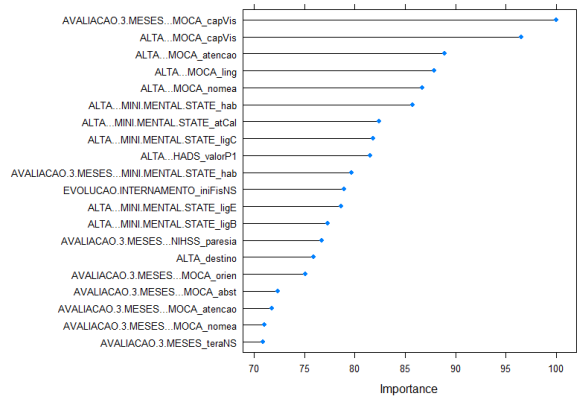
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

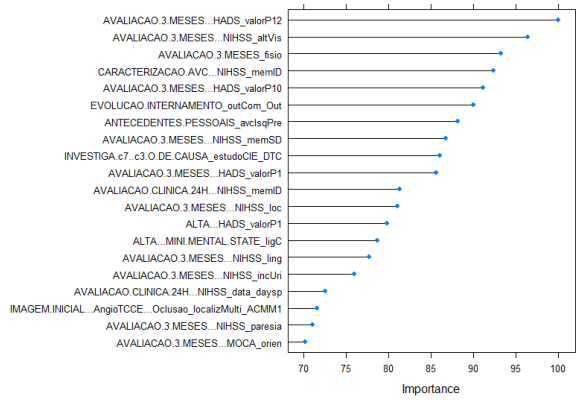


(e) Decision Trees Imputation

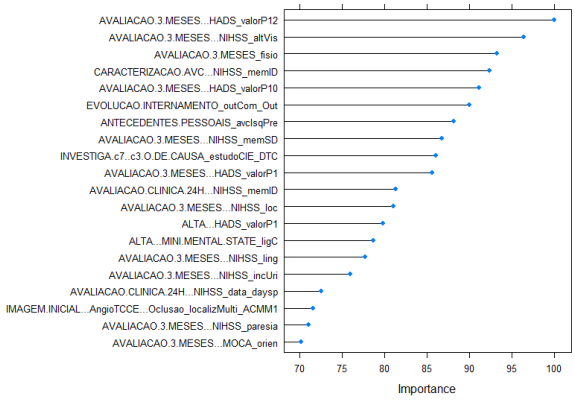


(f) Multiple Imputation

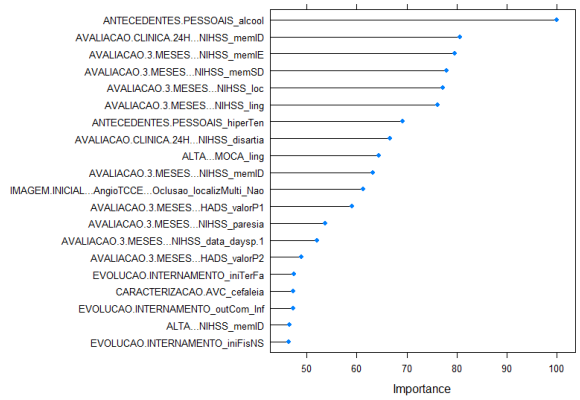
Figure B.1: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model k-Nearest Neighbours predicting the modified Rankin Scale at one year.



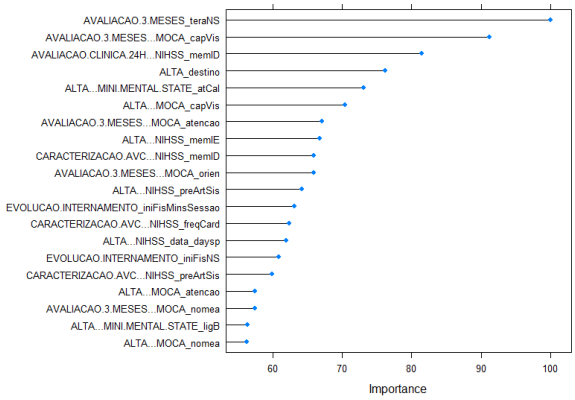
(a) Mode/Median Imputation



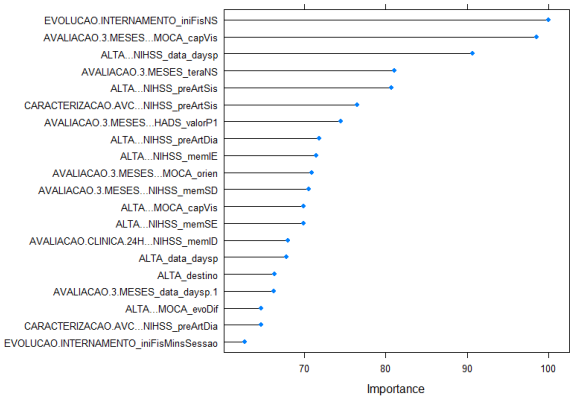
(b) Mode/Median Imputation taking into account the dependence of a few variables



(c) K-Nearest Neighbours Imputation

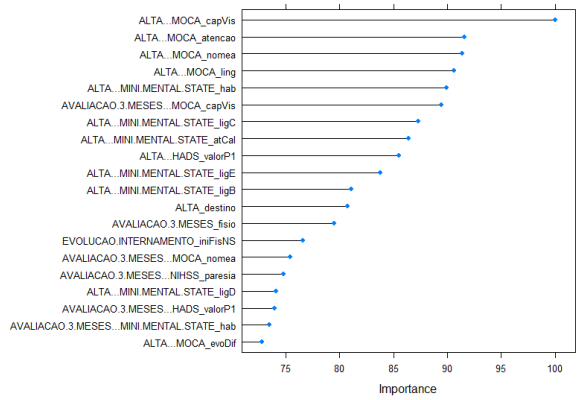


(d) Decision Trees Imputation

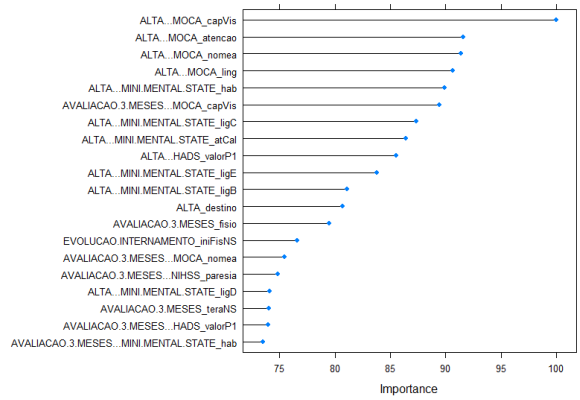


(e) Multiple Imputation

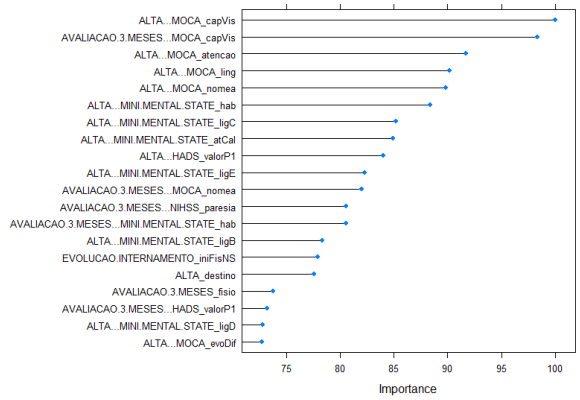
Figure B.2: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Neural Network predicting the modified Rankin Scale at one year.



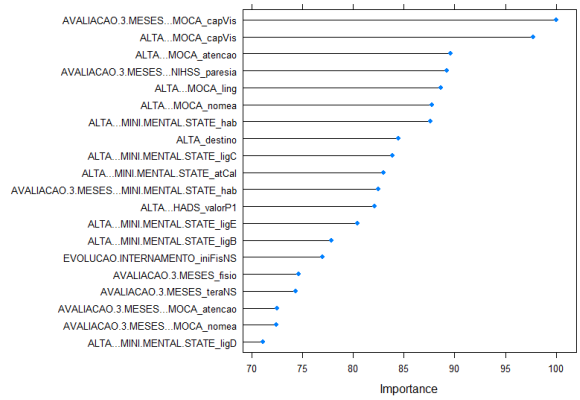
(a) Mode/Median Imputation



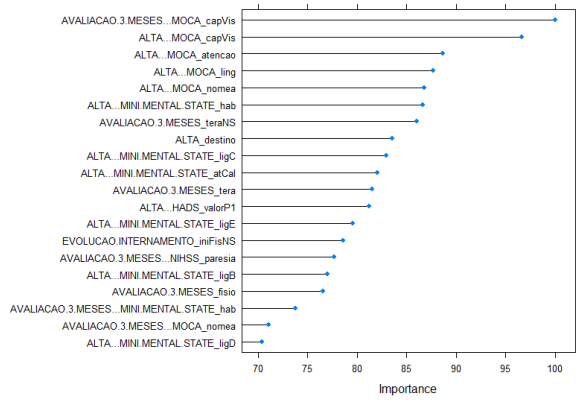
(b) Mode/Median Imputation taking into account the dependence of a few variables



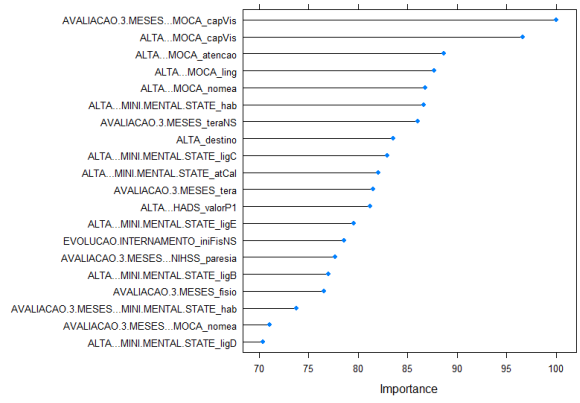
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

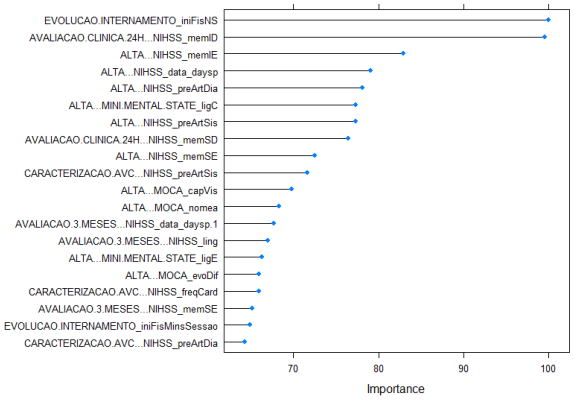


(e) Decision Trees Imputation

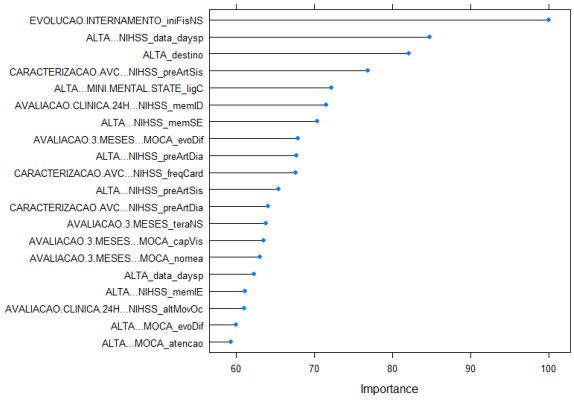


(f) Multiple Imputation

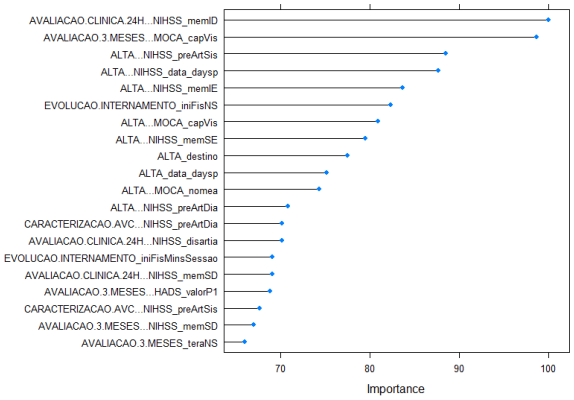
Figure B.3: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Logistic Regression L1-regularised predicting the modified Rankin Scale at one year.



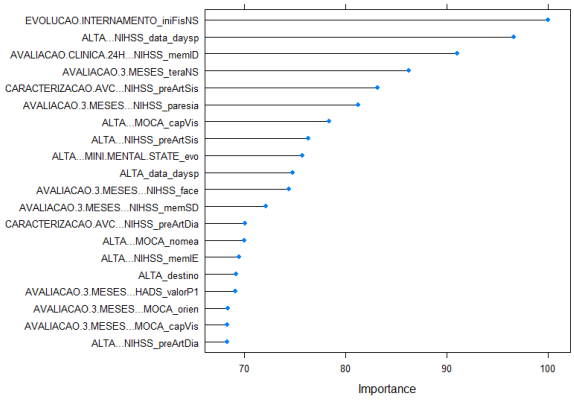
(a) Mode/Median Imputation



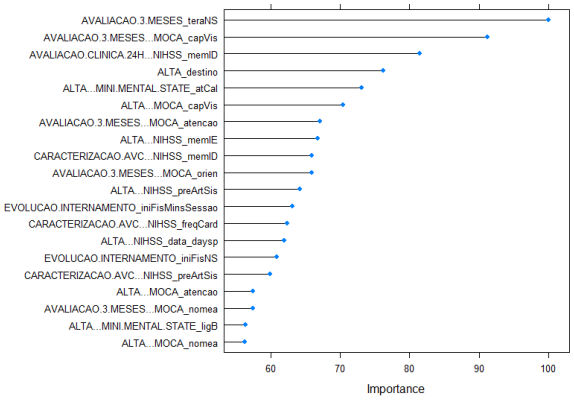
(b) Mode/Median Imputation taking into account the dependence of a few variables



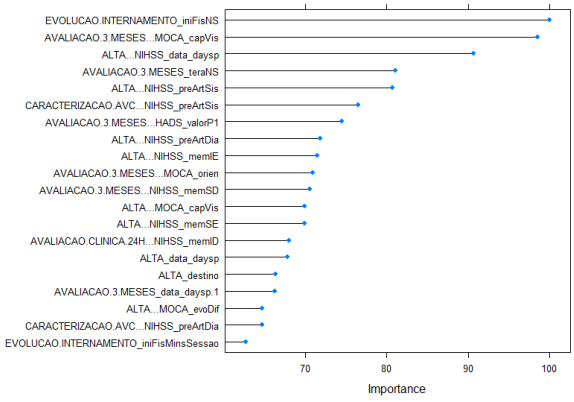
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

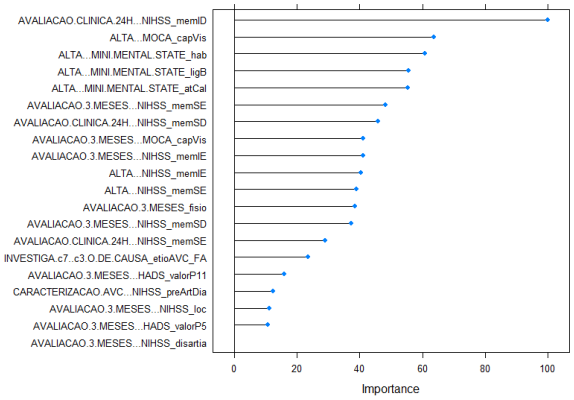


(e) Decision Trees Imputation

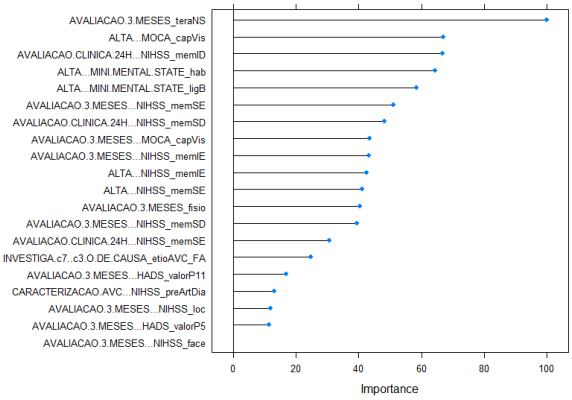


(f) Multiple Imputation

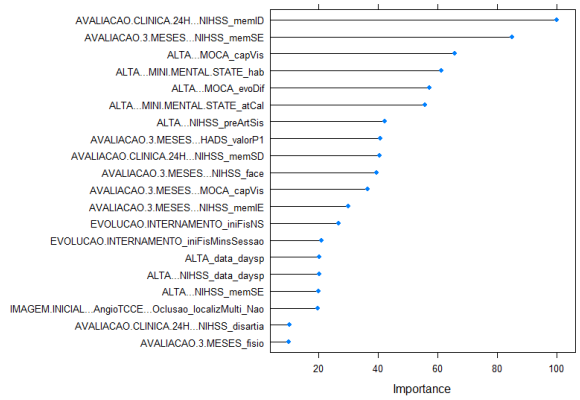
Figure B.4: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Random Forest predicting the modified Rankin Scale at one year.



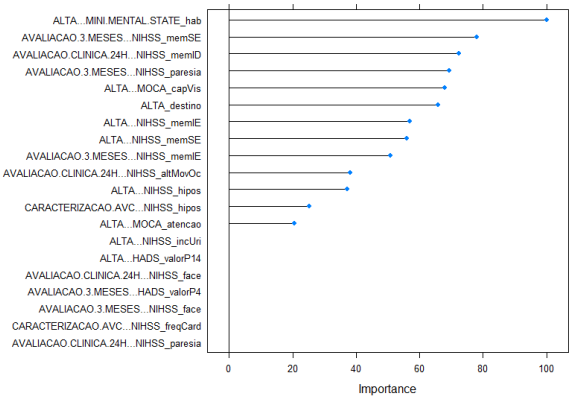
(a) Mode/Median Imputation



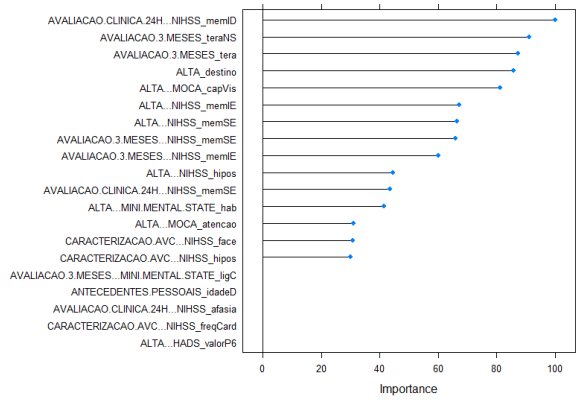
(b) Mode/Median Imputation taking into account the dependence of a few variables



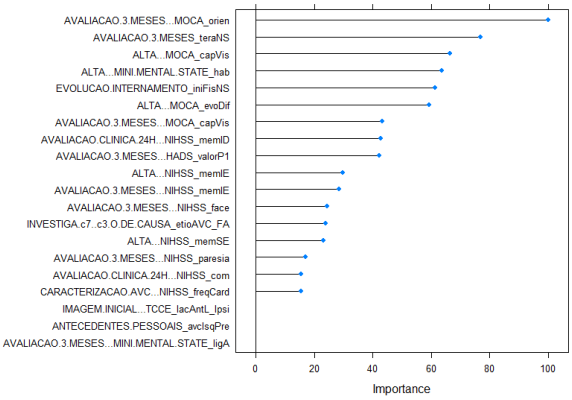
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

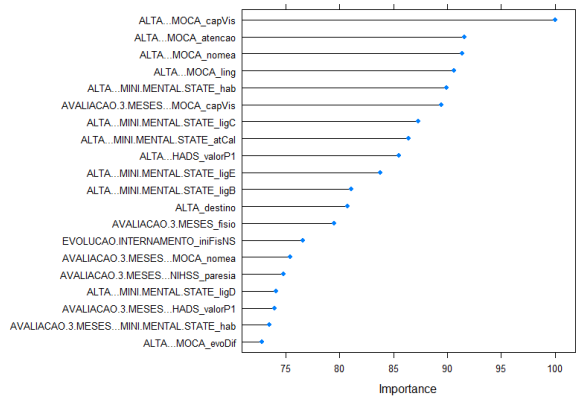


(e) Decision Trees Imputation

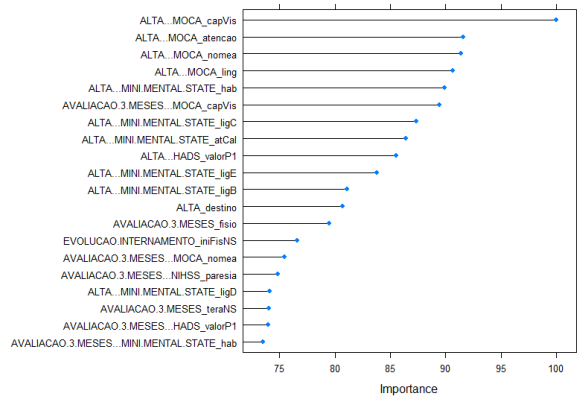


(f) Multiple Imputation

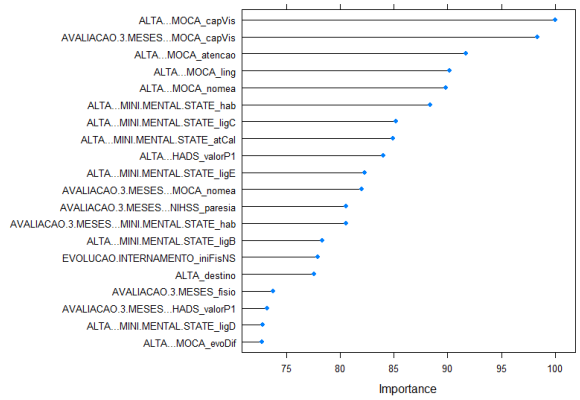
Figure B.5: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model CART predicting the modified Rankin Scale at one year.



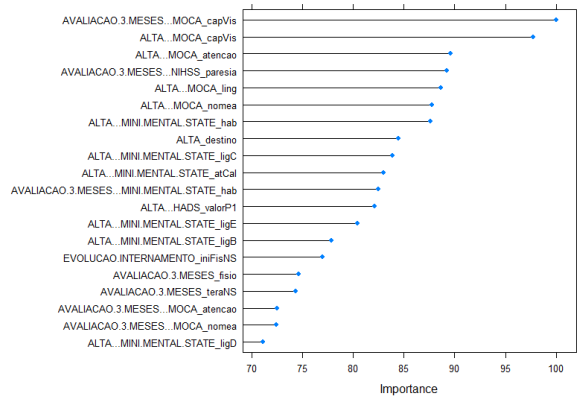
(a) Mode/Median Imputation



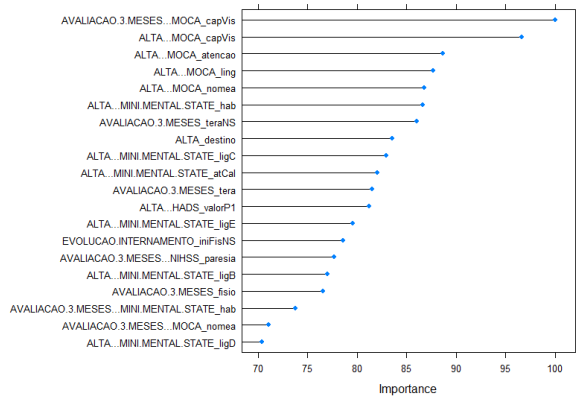
(b) Mode/Median Imputation taking into account the dependence of a few variables



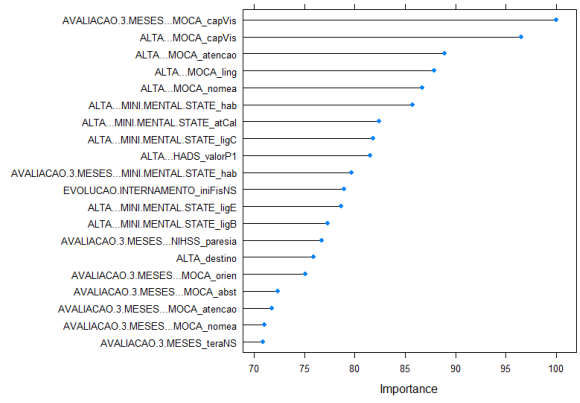
(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

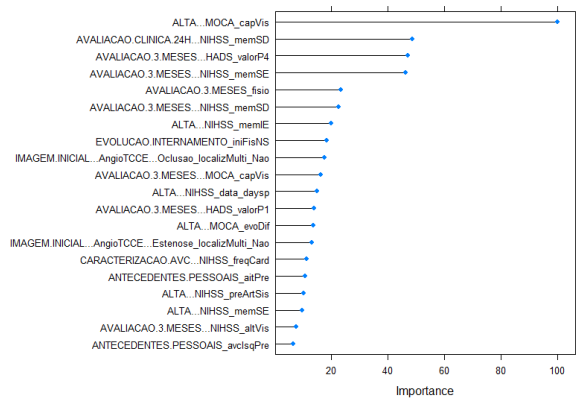


(e) Decision Trees Imputation

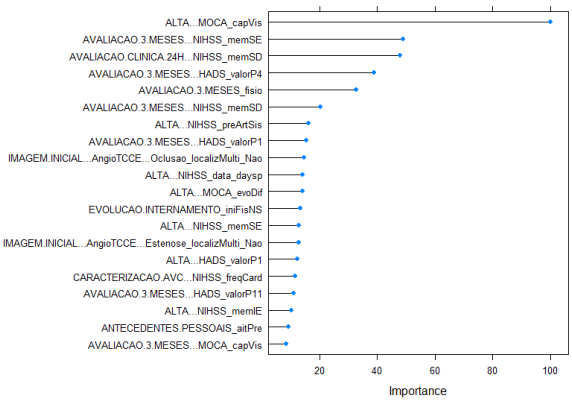


(f) Multiple Imputation

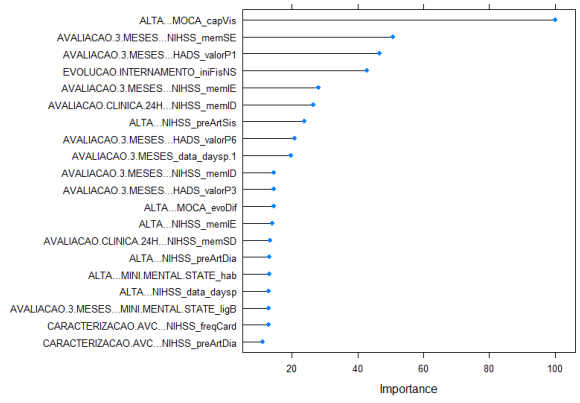
Figure B.6: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Support Vector Machines predicting the modified Rankin Scale at one year.



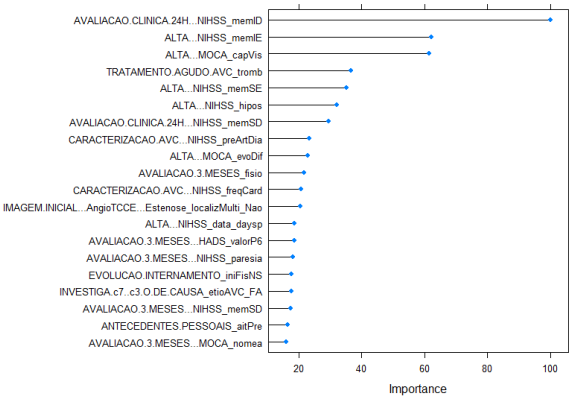
(a) Mode/Median Imputation



(b) Mode/Median Imputation taking into account the dependence of a few variables



(c) Hotdeck Imputation



(d) K-Nearest Neighbours Imputation

Figure B.7: The twenty most important variables and its relative importance (scale of 100%) for each imputation method for the model Extreme Gradient Boosting predicting the modified Rankin Scale at one year.