UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

LUCAS HENNEMANN PERIN

Machine learning approaches for predicting diabetes and determining risk factors from epidemiological data

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Engeneering

Advisor: Prof[a]. Dra. Mariana Recamonde Mendoza

Porto Alegre
July 2018

"Probability is not a mere computation of odds on the dice or more complicated variants; it is the acceptance of the lack of certainty in our knowledge and the development of methods for dealing with our ignorance."

— Nassim Nicholas Taleb

ACKNOWLEDGMENT

I am particularly grateful for the assistance, guidance and support given by Profª. Dra. Mariana Recamonde Mendoza and her useful critiques during this research. I would also like to thank Prof. Dr. Raúl Andrés Mendoza Sassi for providing data and collaborating with this research.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

ABSTRACT

The medical field has an urgent need for new analytical methods that are able to process complex and voluminous data, improving diagnostic tools and the knowledge regarding disease risk factors. In this sense, machine learning (ML) algorithms have become increasingly popular in the analysis of clinical and epidemiological data. The aim of this work is twofold. First, we carry out a systematic literature review (SLR) to investigate recent efforts towards the use of ML algorithms in the study of chronic diseases and summarize, in a comparative way, the performance of distinct methods for training prediction models and detecting risk factors. Second, based on the knowledge derived from the SLR, we apply a ML methodology to analyze data from an epidemiological study that investigated the impact of socioeconomic factors on the occurrence of chronic diseases, including diabetes. We apply multiple ML algorithms and assess their performance for training accurate prediction models and identifying important risk factors for the development of this disease. Our SLR results corroborate the notion that ML technology is growing exponentially in the medical research community, with several ML methods presenting promising results, which are extremely competitive in relation to traditional approaches such as clinical prediction scores derived by experts. Moreover, our experimental results with the diabetes dataset suggest that the Random Forest algorithm have the best predictive capability in the explored scenario, and that ML, in general, has great potential to elucidate new associations among socio-demographic variables and diabetes that may be useful for the development of new public health intervention programs to reduce the incidence of this disease.

Keywords: Machine learning. diabetes. decision tree. support vector machine. random forest. logistic regression. nested cross validation. disease prediction. risk factors.

Abordagens de aprendizado de máquina para predição de diabetes e determinação de fatores de risco a partir de dados epidemiológicos

RESUMO

A área médica tem uma necessidade urgente de novos métodos analíticos capazes de processar dados complexos e volumosos, melhorando as ferramentas de diagnóstico e o conhecimento sobre os fatores de risco de doenças. Nesse sentido, algoritmos de aprendizado de máquina têm se tornado cada vez mais populares na análise de dados clínicos e epidemiológicos. O objetivo deste trabalho é duplo. Primeiro, realizamos uma revisão sistemática da literatura para investigar os esforços recentes para o uso de algoritmos de aprendizado de máquina no estudo de doenças crônicas e resumir, de forma comparativa, o desempenho de métodos distintos para treinar modelos de previsão e detectar fatores de risco. Em segundo lugar, com base no conhecimento derivado da revisão sistemática da literatura, aplicamos uma metodologia de aprendizado de máquina para analisar dados de um estudo epidemiológico que investigou o impacto de fatores socioeconômicos na ocorrência de doenças crônicas, incluindo diabetes. Aplicamos vários algoritmos de aprendizado de máquina e avaliamos seu desempenho para o treinamento de modelos precisos de previsão e identificação de fatores de risco importantes para o desenvolvimento desta doença. Nossos resultados da revisão sistemática corroboram a noção de que a tecnologia de aprendizado de máquina está crescendo exponencialmente na comunidade de pesquisa médica, com vários métodos de aprendizado de máquina apresentando resultados promissores, extremamente competitivos em relação às abordagens tradicionais, como os escores de previsão clínica obtidos por especialistas. Além disso, nossos resultados experimentais com o conjunto de dados do diabetes sugerem que o algoritmo Random Forest tem a melhor capacidade preditiva no cenário explorado, e que o aprendizado de máquina, em geral, tem grande potencial para elucidar novas associações entre variáveis sociodemográficas e diabetes que podem ser úteis para o desenvolvimento de novos programas de intervenção em saúde pública para reduzir a incidência desta doença.

Palavras-chave: aprendizado de máquina, diabetes, árvore de decisão, regressão logística, floresta aleatória, máquina de vetores de suporte, validação cruzada aninhada, predição de doenças, fatores de risco.

LIST OF ABBREVIATIONS AND ACRONYMS

ML	Machine Learning
SLR	Systematic Literature Review
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
NCV	Nested Cross Validation

LIST OF FIGURES

Figure 2.1 Example of a simple decision tree trying to predict if a customer will buy a computer 16
Figure 2.2 An example of a SVM hyperplane maximizing the margins in relation to the closest points in each class (i.e., the support vectors). 18
Figure 2.3 An example of the iterations and data separation for a 10-fold cross validation 20
Figure 2.4 Diagram showing the processing steps of the Nested K-Fold Cross Validation 22

Figure 3.1 SLR algorithm steps, containing the initial search phase, which gather the initial list of articles for processing up until the very last phase, where the whole articles are considered 30
Figure 3.2 The PRISMA flowchart summarizing the phases of this SLR and the number of articles that passed or were excluded during each phase. 30
Figure 3.3 The yearly distribution of papers resulting from the SLR. 31
Figure 3.4 A graph illustrating the number of papers containing each type of study: diagnostic prediction (PD), risk factors (FR), prognostic prediction (FP) 32
Figure 3.5 The number of papers according to the ICD-10 chapter 33
Figure 3.6 A graph illustrating the number of articles that included the respective algorithms in their research 34
Figure 3.7 A candle graph illustrating the ACCURACY score of each algorithm, according to the measurements of papers resulting from the SLR. 35
Figure 3.8 A candle graph illustrating the AUC_ROC score of each algorithm, according to the measurements of papers resulting from the SLR. 35

Figure 5.1 LR scores computed from 10-fold cross-validation. The rightmost columns depics the average over all iterations. 46
Figure 5.2 Variables selected for at least five iterations of the LR outerloop training. 46
Figure 5.3 DT scores computed from 10-fold cross-validation. The rightmost columns depics the average over all iterations. 49
Figure 5.4 Variables selected for at least five iterations of the DT outerloop training. 49
Figure 5.5 RF scores computed from 10-fold cross-validation. The rightmost columns depics the average over all iterations. 50
Figure 5.6 Variables selected for at least five iterations of the RF outerloop training. 51
Figure 5.7 SVM scores computed from 10-fold cross-validation. The rightmostcolumns depics the average over all iterations. 53
Figure 5.8 Variables selected for at least five iterations of the SVM outerloop training. 54

Figure B.1 LR - inner loop best variables ranking 77
Figure B.2 DT - inner loop best variables ranking 77
Figure B.3 RF - inner loop best variables ranking 78
Figure B.4 SVM - inner loop vest variables ranking 78

LIST OF TABLES

Table 3.1 Complete SLR Protocol 27

Table 4.1 Logistic Regression Parameters 42
Table 4.2 SVM Parameters 42
Table 4.3 Decision Tree Parameters 43
Table 4.4 Random Forest Parameters 43

Table 5.1 LR average scores and their standard deviation 45
Table 5.2 Description of the most important variables meanings, according to the results of the LR algorithm. 47
Table 5.3 LR - C parameter iteration use count 48
Table 5.4 LR - penalty parameter use count 48
Table 5.5 DT average scores and their standard deviation. 48
Table 5.6 Description of the most important variables meanings, according to the results of the DT algorithm. 49
Table 5.7 DT - criterion parameter iteration use count 50
Table 5.8 DT - splitter parameter iteration use count 50
Table 5.9 RF average scores and their standard deviation. 50
Table 5.10 Description of the most important variables meanings, according to the results of the RF algorithm. 52
Table 5.11 RF - criterion parameter iteration use count 53
Table 5.12 RF - N. estimators parameter iteration use count 53
Table 5.13 SVM average scores and their standard deviation. 53
Table 5.14 Description of the most important variables meanings, according to the results of the SVM algorithm. 54
Table 5.15 SVM - C parameter iteration use count. 55
Table 5.16 Most important variables - summary 60

CONTENTS

1 INTRODUCTION 12
2 THEORETICAL REFERENCE 14
2.1 Supervised Learning 14
2.2 Algorithms 15
2.2.1 Logistic Regression 15
2.2.2 Decision Tree 16
2.2.3 Random Forest 17
2.2.4 SVM 18
2.3 K-Fold Cross Validation 19
2.4 Parameter Grid Search 20
2.5 Variable Selection 21
2.6 Nested Cross Validation 21
2.7 Score Functions 23
3 SYSTEMATIC LITERATURE REVIEW 25
3.1 Protocol Definition 25
3.2 Search Preparation 27
3.3 Execution 29
3.4 Results Overview 30
4 METHODOLOGY 36
4.1 Dataset 36
4.2 Framework 37
4.3 Pre-processing 37
4.4 Global Pre-processing 38
4.4.1 Quiz Fields 38
4.4.2 Transformation of Field Values 38
4.4.3 Ignored Fields 39
4.4.4 Mandatory Variables 39
4.4.5 Balanced Class Size 39
4.4.6 Invalid Values 39
4.5 Local Pre-processing 40
4.6 Algorithms 40
4.7 Training 41
4.7.1 Feature Selection 41
4.7.2 Model Tuning 41
4.7.3 Logistic Regression Model Tuning 42
4.7.4 SVM Model Tuning 42
4.7.5 Decision Tree Model Tuning 42
4.7.6 Random Forest Model Tuning 43
4.7.7 Model Training 43
4.8 Output Reporting 43
5 RESULTS AND COMPARISON 45
5.1 Logistic Regression 45
5.2 Decision Tree 48
5.3 Random Forest 50
5.4 SVM 51
5.5 Execution 55
5.6 Limitations 56

5.7 Comparison 56
5.7.1 SLR Results Comparison 57
5.7.2 The Optimal Algorithm 57
5.7.3 Risk Factors 59
6 CONCLUSION 63
REFERENCES 65
ANNEX A — SLR RESULTING ARTICLES 72
ANNEX B — BEST VARIABLES RANKING ACCORDING TO THE INNER LOOP 77

1 INTRODUCTION

According to Mendis (2014), "Noncommunicable diseases (NCDs) are one of the major health and development challenges of the 21st century, in terms of both the human suffering they cause and the harm they inflict on the socioeconomic fabric of countries". More than 14 million people die each year from noncommunicable diseases, and most of them are from developing countries (MENDIS, 2014). In particular, diabetes is a significant contributor to increased mortality rates due to chronic diseases, being responsible for 1.6 million deaths around the world in 2015, and projected to be the seventh leading cause of worlwide deaths in 2030 (International Diabetes Federation.,). Hence, there is an urgent need for solutions that may help improve early diagnosis of diabetes and other chronic diseases, as well as the discovery of risk factors in a cost-effective and efficient way.

Numerous hospitals around the globe are, on a daily basis, collecting and saving data from their patients in databases, which can be a valuable resource for developing new methods to assist in clinical decision making. Substantial efforts are, however, required to integrate and make sense of health care data in a big data scale (BEAM; KOHANE, 2018). Due to data volume and complexity, new approaches are necessary to effectively deal with large number of variables and detect the complex relationships in the data. To this end, machine learning (ML) algorithms have proven to be effective solutions for analyzing sizeable amounts of data simultaneously, and have helped to train predictive models with competitive performance in comparison to traditional statistical methods, such as logistic regression.

Machine learning is a technology that can be observed throughout contemporary society and is able to learn a task with little human instruction or prior assumptions (BEAM; KOHANE, 2018). ML algorithms have undergone continuous development and have been applied with significant levels of success in almost all areas of knowledge. Among these areas, medical sciences have profited from applications of ML in several distinct practical problems, including the procedures of disease diagnosis and prognosis, prediction of disease recurrence, and identification of preventable risk factors (KONONENKO, 2001; GUI; CHAN, 2017). It is already recognized that ML is an increasingly necessary tool for the modern health care system (BEAM; KOHANE, 2018). Nonetheless, the depth, power, and effectivenness of these approaches within the study of the epidemiology of chronic diseases, as well

as how they compare to traditional statistical predictive models, are still not well characterized.

Therefore, the goal of this work is to investigate and explore the potential of ML algorithms in the analysis of epidemiological data with the purpose of training predictive models for NCDs and identifying preventable risk factors. In this context, we first perform a systematic literature review (SLR) to identify the the state-of-the-art on ML approaches in NCDs research and advance our understanding on which ML methods have been applied and what is their potential in this specific domain. Second, based on the solid review of the related literature and the practical references provided by the SLR, we explore and evaluate the incorporation of ML algorithms in epidemiological studies. Specifically, this research trains and analyzes the results from four ML algorithms with the intent to advance diagnosis prediction and risk factor analysis for diabetes disease. The four algorithms trained are the logistic regression (LR), the decision tree (DT), the random forest (RF), and the support vector machine (SVM). Our results corroborate the idea that ML has become an increasingly prominent technology in medical research, especially in epidemiological studies related to chronic diseases, and that it is extremely competitive in relation to classical predictive models provided by logistic regression and clinical medical scores. In addition, our experiments with diabetes-related data evidenced the suitability of ML methods for pursuing good predictive performance for disease diagnosis and also for highlithing important contributing factors for increased risk of disease, which could be useful for the development of new public health intervention programs to reduce the incidence of diabetes.

This thesis is divided into three major sections, a systematic literature review, the methodology adopted for our ML experiments, and the results obtained from the analysis of diabetes-related data using the aforementioned ML algorithms. Firstly, Chapter 2 summarizes concepts related to ML and its algorithms that are related to this research. Secondly, Chapter 3 details the SLR protocol, its execution, and the statistics about the results. Thirdly, Chapter 4 describes the dataset and the methodology adopted in our ML approach for training predictive models, including pre-processing strategies. Lastly, Chapter 5 presents the results, the difficulties that arose during model training, and, importantly, a detailed analysis of the resulting statistics.

2 THEORETICAL REFERENCE

This chapter acts as a theoretical basis for the concepts and algorithms utilized and discussed in this thesis. Michalski, Carbonell e Mitchell (2013) provide a strong conceptual definition of ML: "The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."

Machine learning can operate as a set of algorithms that try to imitate the human mind. ML algorithms achieve this imitation by "learning" a specific subject without a human having to provide strict rules that define how a given subject works. There are two major forms of ML algorithms, unsupervised and supervised learning. Unsupervised learning occurs when there is no output variable. Supervised learning occurs when both input and output variables are given, and the aim is then to explain the dependent variable, which is the output, in terms of the independent variables (ALGHAMDI et al., 2017). The performance of supervised learning is measured by comparing the value of the known output and the predicted one. Since the objective of this research is diagnosis prediction, which has a predefined output, the focus here will be only on supervised algorithms.

2.1 Supervised Learning

Supervised Learning is a form of learning by experience, in which the algorithms try to learn by examining a previously verified set of data, usually called the 'training data'. This data contains a verified set of input and output combinations, which allows the algorithm, firstly, to learn the connections between the input and output values and, secondly, to try to develop a valid strategy to achieve the desired output from an arbitrary input. After the algorithm has developed a way to predict the output from the input based on the known data, new data can be fed into the algorithm so as to predict or evaluate its efficiency.

In short, supervised learning aims to find a relationship between the input data and the desired output, which is valid beyond specific case analyses. Thus, the the extracted relationship should function for an arbitrary input, and this makes it useful in predicting new cases where the same set of independent variables (i.e. features) is available and the same dependent variable (i.e. class or output value) is

necessary.

2.2 Algorithms

The main focus of this study is supervised ML algorithms for the purposes of classification. Classification is the process of assigning groups or categories for a determined set of inputs (ALGHAMDI et al., 2017). The supervised ML algorithm maps each input onto one option among a predefined and limited set of possible outputs.

In the following section, we provide a brief summary of the particular algorithms that have been selected to be applied in the scope of this work. Since this thesis does not aim to implement these algorithms, the particulars of this implementation are not discussed. Instead, this section focus on giving a basic understanding of the algorithms and how they function.

2.2.1 Logistic Regression

Logistic regression is a linear statistical classifier that provides the probability for each output class as an equation from the input values (ALGHAMDI et al., 2017). The algorithm is based in mathematical theorems and can be more or less sophisticated, depending on the intent, but the basic feature of LR is that the algorithm, given an input, can predict the probability of a defined output class.

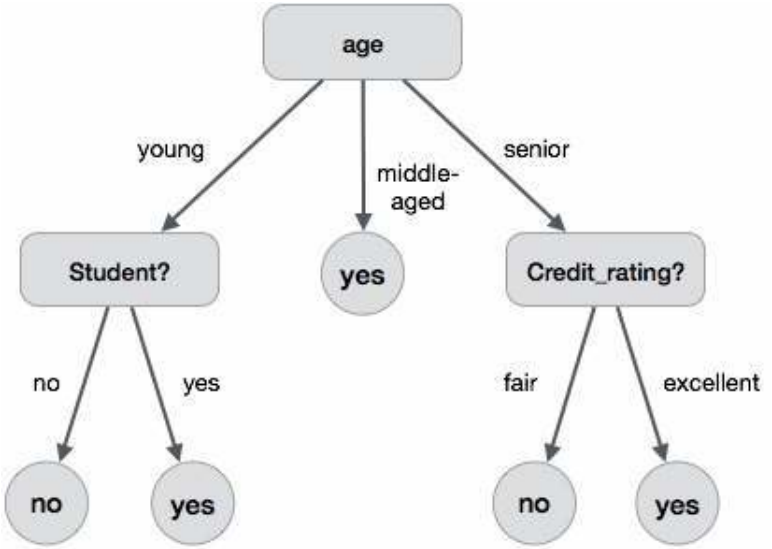
The training of a LR model involves minimizing an equation to a curve, in a model that has extremely high probabilities in one class (y = 1), and extremely low probabilities in the other (y = 0). The desired curve model is usually achieved by a sigmoid function, which is obtained through the combination of all variable values. There are certain mathematical models which explain how this function is constructed and how it is minimized to fit the dataset; these explanations are highly technical, and we refer to Géron (2017) for further details.

The LR algorithm contains a parameter for configuring how quickly the minimization function should approach the result. Apart from this, it is an extremely simple algorithm and it is often considered the “classical” or mathematical form of ML.

2.2.2 Decision Tree

Decision trees are versatile ML algorithms, which can perform both classification and regression tasks. They are powerful algorithms, capable of fitting complex datasets (GÉRON, 2017). A DT consists of nodes in the form of a tree; starting from the root, these nodes evaluate a predefined rule, for example, a condition, or a mathematical operation. The leaves of the tree represent the predicted class or value. One of the decision tree's advantageous features is its transparency, this means the resulting tree can be read and analyzed by a human (RAMEZANKHANI et al., 2014). It is, furthermore, interesting to note that important features usually appear closer to the root of the tree, whereas unimportant ones appear closer to the leaves; thus the depth of the variables within the DT could be used as a metric of importance (GÉRON, 2017). The employment of DT algorithms is widespread in medicine and because of its interpretability feature, it is widely used for evaluating risk factors (KARAOLIS et al., 2010).

Figure 2.1: Example of a simple decision tree trying to predict if a customer will buy a computer



Source: Decision Tree page on TutorialsPoint [1]

Figure 2.1 gives an example of a DT, where three variables are evaluated, and the tree tries to predict wether or not a given customer will purchase a computer. Each variable can have multiple values and might or might not branch out, depending on how relevant a given variable is to the classification. In Figure 2.1,

[1] Available in <https://www.tutorialspoint.com/data_mining/dm_dti.htm>, accessed in july 2018

age is the first variable. It branches out into three options, depending on whether the user is young, middle aged, or senior. It can also function by branching out into the actual age, by, for example, verifying if the age is higher or lower than 20. As can be discerned from the figure, middle-aged people will probably buy a computer; other variables are not relevant in this specific case, and the result can be reached before evaluating them. In addition, each branch can have different variables that are under consideration; this allows for different combinations of variables to be considered independently.

2.2.3 Random Forest

A RF algorithm, is, in fact, an evolved form of the DT algorithm. It follows the strategy of the ensemble method; the mentioned strategy is the combination of more than one tree-structured classifiers for improved prediction capabilities. The RF algorithm can be interpreted as a collection of combined DT, wherein each DT is trained over a slightly different part of the dataset, causing the DTs to be slightly different from one another. Once the trees have been trained, they are given a weight depending on their prediction performance, and the resulting class is calculated by combining all results and multiplying by their weight. The class with highest weighted probability is decided as the output value for the RF. This method provides a higher capacity for complexity and thus allows for higher prediction accuracy (GÉRON, 2017).

When the nodes are split in a RF algorithm, the internal DT, unlike the standard decision tree, does not try to find the optimum split among all variables; instead it performs the splitting action by utilizing a subset of randomly chosen variables (LIAW; WIENER et al., 2002). This counter intuitive way of splitting nodes contributes to the process by which each tree inside the RF maintains a different view of the dataset and thus provides more overall information for the output prediction.

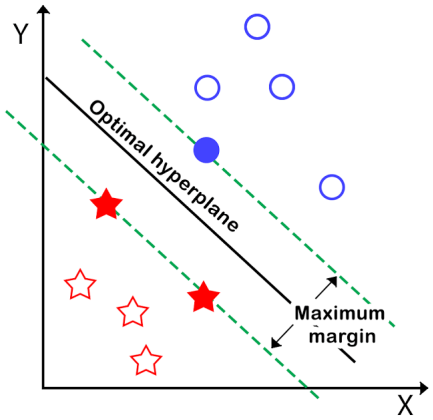
The RF algorithm can also be used as a tool for understanding which features are the most important during the training. Since RFs are built from a group of DTs, a statistical analysis on the depth of variables on each internal DT gives an extremely useful and, usually, accurate depiction of which variables are important (GÉRON, 2017).

The two main configurable parameters of a RF are the number of classifiers in the ensemble and the splitting method. The number of classifiers is the number of DTs to be trained within the Random Forest, which increases the complexity and the prediction ability of the algorithm. The splitting method is how the node splitting is applied, both in terms of how many random variables will be selected for evaluation in each node split and the splitting criterion used to determine the best variable among this subset (GÉRON, 2017).

2.2.4 SVM

Support vector machines are the state-of-the-art margin classifiers (SCHULDT; LAPTEV; CAPUTO, 2004). The core SVM algorithm was originally a binary classification method that was developed by Vapnik and colleagues at Bell Laboratories, although there were significant improvements and developments to the algorithm by others (MADZAROV; GJORGJEVIKJ; CHORBEV, 2009).

Figure 2.2: An example of a SVM hyperplane maximizing the margins in relation to the closest points in each class (i.e., the support vectors).



Source: Support Vector Machine page on Quantra [2]

SVM functions by using the inputs as points in space and trying to minimize the output errors by dividing the classes with hyperplanes, which are called margins (SCHULDT; LAPTEV; CAPUTO, 2004). This margin can be conceptualized as a "street" that separates the data in two groups, and the algorithm aims to find the widest possible "street" between the classes. This means that all new data that lie outside the boundaries of the street do not affect the training, only points on the

[2] Available in <https://quantra.quantinsti.com/glossary/Support-Vector-Machine>, accessed in july 2018

edge of the margin can change the results. Those points are considered support vectors (GÉRON, 2017). The classification of a new data point depends on where this point is located relative to this hyperplane. The success of SVM algorithms depends greatly on the training data and how the classes are divided in the given dataset. When the data are not linearly separable and a single hyperplane can not properly divide the classes, the algorithm adopts a kernel function to transform the data into a new space where it is separable.

In summary, SVMs locates hyperplanes that aim at dividing the input data into groups according to their output classes in the best possible way. Two exemples of SVM applications can be observed in image pattern recognition (SCHULDT; LAPTEV; CAPUTO, 2004), and disease diagnostic (BERIKOL; YILDIZ; OZCAN, 2016).

2.3 K-Fold Cross Validation

All algorithms base themselves in the training data and this can cause an effect known as 'overfitting', which means that a model has been fitted very well to the given dataset, but does not generalizes properly to unknown data. Thus, when an algorithm tries to predict a case that lies outside the previously trained dataset, the predictions are either poor or are not as efficient as the training scores may suggest.

To estimate the performance and generalization power of ML models, a frequently employed solution is the holdout strategy, which divides the input data into two disjoint sets, a training and a testing set. The training set is used during the training of the algorithm and then the testing set would be used to generate a standalone prediction and compare the results, generating a score. This means that the trained algorithm needs to be able to work with unknown data as well as the training data. In an ideal scenario, the dataset should be large enough so that the training and testing sets can be sufficiently diverse and large so as to be representative of the whole process (KRSTAJIC et al., 2014). Since the available real-world datasets usually are not sufficiently diverse nor large enough, most cases do not contain sufficient data for a single split. This is why certain solutions had to be developed.

[3]Available in <https://www.researchgate.net/The-K-fold-cross-validation-scheme-133-Each-of-

Figure 2.3: An example of the iterations and data separation for a 10-fold cross validation

Source: Generalized Methods for User-Centered Brain-Computer Interfacing - Scientific Figure on ResearchGate. [3]

In k-fold cross validation the dataset is randomly divided into k smaller datasets known as folds, then the desired ML model is retrained k times, such that each iteration selects one of the subdivisions in the dataset as the testing data, and all the remaining folds are merged into a training set (KRSTAJIC et al., 2014). Figure 2.3 shows a 10-fold example of this process. At the end of this process, an average score for the k folds can be calculated. This evaluates the capacity of the algorithm to work with unknown data in a reliable way, without requiring an oversized dataset. The disadvantage to this process is the requirement of approximately k times more processing, only to evaluate the real-world score of the given algorithm, which can be an issue in certain scenarios.

2.4 Parameter Grid Search

Another feature of machine learning algorithms is that they usually contain several parameters that can be configured for the specific data or domain that is being trained. This, however, introduces a new problem when training; the optimal parameter values have to be employed, if not, the algorithm could have an underused potential.

There are a few options for this problem. The parameters could be manually

the-K-partitions-is-used-as-a-test_fig10_323969239>, accessed in july 2018

changed to try to find the best score, although this is ineffective and time-consuming. Another solution is called parameter grid search, in which a grid of the parameter and the possible values are designed, and the model is then configured so as to train for every combination of parameter value that is present in the grid (KRSTAJIC et al., 2014). Similar to the k-fold cross validation, the parameter grid search is a computationally expensive solution; although for the purpose of finding the optimal parameter values, it is the most satisfactory method.

2.5 Variable Selection

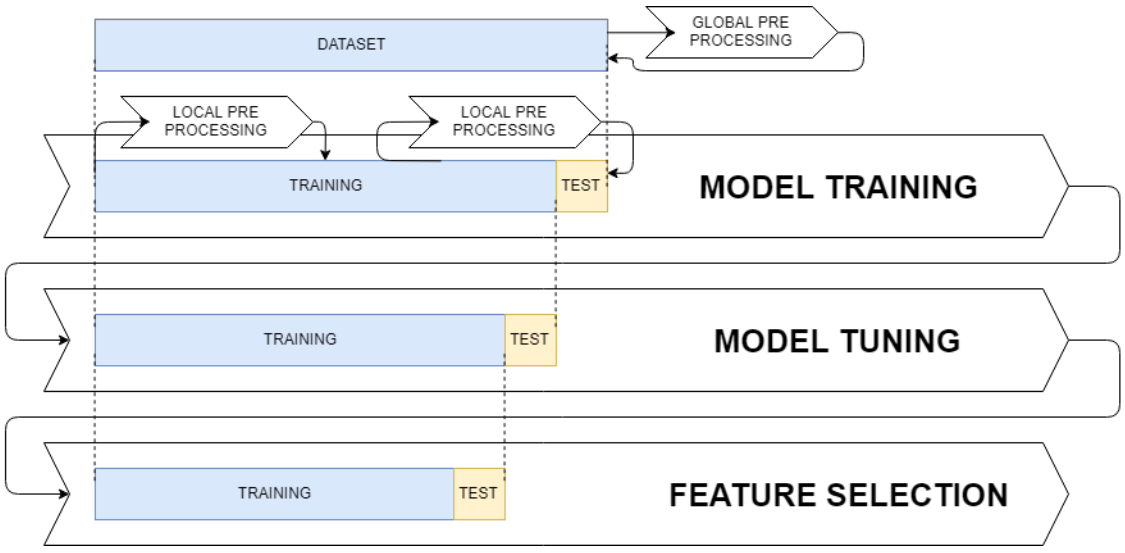
Datasets can contain a substantial number of variables for training, but not all of these may be useful for predicting the output. Indeed, since there can be imperfections in the ML algorithms, those extra variables can actually cause negative impact in the prediction capabilities. The process of selecting which variables should be included in the training is somewhat difficult; before training one does not know to what extent a given variable will affect the result. The simplest way to achieve this is to sort the variables based on a predefined score function and then select the k top variables. This is a currently employed solution, but it is not the optimal selection procedure, because there is no flexibility with respect to the top variables that are presented to the algorithm. Another choice, which is well suited for instances where more than one algorithm needs to be trained, is called recursive feature elimination with cross validation (RFECV). This process identifies the relevant features by repeatedly removing features that possess small impact on the trained model and evaluating the prediction capability of the model trained with the remaining features (JO; AHN; EGGER, 2017). It is an extremely time-consuming strategy for selecting variables, but still widely applied in several domains.

2.6 Nested Cross Validation

The goal when training ML algorithms is to select the optimal combination of variables and parameter values. The most effective way to accomplish this, and to allow the variables and parameter values to converge, is called a nested cross validation (NCV). An NCV is a combination of three already discussed steps: a

k-fold cross validation, a parameter selection, and a variable selection. These steps run in nested loops, comparing every combination of every algorithm possible and ensuring that models will be evaluated using independent test sets. When compared to running a feature selection and parameter selection independently, a NCV considerably reduces the bias and features the best optimization possible (KRSTAJIC et al., 2014). Figure 2.4 and the following bullet points summarize how each loop in the NCV algorithm functions, according to Krstajic et al. (2014):

Figure 2.4: Diagram showing the processing steps of the Nested K-Fold Cross Validation



- In the first step, the original dataset is splitted into k-folds. Global data pre-processing techniques that do not take data distribution or class information into account may be applied previous to this division.
- The outermost loop can be compared to a k-fold cross validation, except it executes the training with the best variables and parameters calculated by the inner loops. The outer loop is where the score, best variables, and best parameters are also recorded for later comparison and analysis. At this stage, local pre-processing techniques that in some way use information on data distribution and class values may be applied separately for training and testing sets.
- The middle loop executes the parameter selection; on each iteration of this loop, the algorithm being trained employs the best variables calculated by the innermost loop, and then runs the desired parameter selection algorithm. This ensures that the best variables for each combination of parameters is

used. This step returns the best parameter and variable combination for the dataset given.

- The innermost loop is for variable selection, it receives the current parameters being tested, and runs the variable selection algorithm, which returns the best variable combination for that specific algorithm and parameters.

Note that both in model tuning and in model training, the trained models are evaluated for generalization and performance assessment with completely independent test sets.

2.7 Score Functions

As a way of comparing the success of different training models, score functions are employed. There is a substantial number of different performance measures available, to simplify the understanding, only the three measures employed during this research are defined: the accuracy, the ROC and AUC score, and the F-measure.

The most straigthforward form of measuring the performance of an algorithm model is by computing its accuracy. The accuracy is calculated by dividing the number of correctly predicted entries by the total number of entries (GÉRON, 2017). This provides an useful, yet, very simple scoring function, which may be largely biased by unbalanced classes.

The F-measure is the harmonic mean of precision and recall. Precision evaluates when an algorithm correctly predicts the positive values. Recall is a metric indicating the fraction of positive values that the algorithm could correctly identify. As the harmonic mean of these metrics, the F-measure evaluates how balanced in terms of positive predictive power and true positive rate an algorithm is (GÉRON, 2017). Here, we adopt the F1-measure, in which recall and precision are evenly weighted.

Receiver operating characteristic (ROC) is a very common tool used with binary classifiers. ROC curve plots the fraction of correctly predicted positive examples (i.e., the true positive rate) in relation to the fraction of negative examples incorrectly classified as positive (i.e., the false positive rate) for various threshold settings. The area under the curve (AUC score), which ranges from 0 to 1, provides a measure of predictive performance for the algorithm and can be understood as

follows: in a perfect algorithm the AUC score is 1, whereas in a completely random classifier algorithm the score is equal to 0.5 (GÉRON, 2017).

3 SYSTEMATIC LITERATURE REVIEW

ML is a large and ever-expanding subject. Within the topic, there is substantial collection of algorithms that are employed. Indeed, ML has numerous sub-areas of research, and has recently gained the interest of diverse researchers. To better understand the potential of this data analysis technique and to harness its power for this research, an systematic literature review (SLR) was executed. An SLR is an algorithm that is followed when new areas of research are approached. It entails a well-defined set of instructions on how to identify, evaluate, and summarize the state-of-the-art within a specific area or theme. It provides a fixed algorithm as a way of finding the current optimal literature on the specific area of this dissertation, and it allowed, moreover, to calculate certain statistics about what researchers have been employing and what results have been reached so far. Altogether, this information can help provide the basis for some of the decisions that were made during the experiments carried out in this work. (MARIANO et al., 2017).

The protocol of SLR has three main phases for the reviewing and filtering of papers. All the phases are executed independently by at least 2 persons to avoid personal bias in the results. After each phase, the results from the participants are compared and only articles that were approved by the majority of the participants are kept.

The protocol is iterative. This means that at every phase an analysis should be made over the resulting articles for previously unconsidered cases, or changes should be made in the protocol for results that better suit the expected knowledge to be acquired (MARIANO et al., 2017).

3.1 Protocol Definition

This systematic review was designed and performed according to the principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and the guideline recently proposed by Mariano et al. (2017) for conducting SLR in the interdisciplinary field of Bioinformatics. It was chosen primarily for being a guide based exactly on the bioinformatics field, which is the focus of this research.

The SLR algorithm requires the definition of its protocol as the first step.

This protocol should contain information regarding the scope of the algorithms being searched, the criteria used to include or exclude an article, and certain questions to be used in ranking the articles according to the information they provide (MARIANO et al., 2017). In the following bullet points, more specific definition of the protocol's composition is provided, which is based on the description given by (MARIANO et al., 2017):

- The main question. This defines the scope of the review in a quite loose or open-ended manner. In the case of this research, the aim was to find articles about predictive models and analysis of risk factors that were associated with diseases, but which also involved experiments so that a comparison with this research would be possible.
- Objective. This contains the main objective when undertaking the SLR; this is similar to the main question but affirms what the research wants to achieve. It functions as a reference point for the participants so that they understand the final goal of the SLR.
- Inclusion criteria. This category sets certain criteria to be met by those papers added to the review; usually not all of these criteria have to be met, but preference is given to those which meet a greater number. The criteria functions as a guidance system as to what is expected from the accepted articles, thus removing somewhat the burden of selection from the participants.
- Exclusion criteria. These criteria include certain rules that the selected papers **have** to follow. If a paper contains one exclusion criterion, it must be removed. As the second most important item of the SLR table, the exclusion criteria defines a few "no exception" rules that the article has to meet.
- Specific questions. This is the most important item in the SLR table. It entails several questions that define more specifically what is wanted from the review. These questions regarding the expected aspects of the article are used during the final phase of the review. Papers are given a score based on whether or not they answer these questions. Each question is given a value of 0, 1 or 2, depending on whether the article does not fulfill any of the question requirements, partially fulfill its requirements, or completely fulfill its requirements, respectively, and all values summed for each participant. This produces a score between 0 to 10, considering the recommendation of defining

five specific questions. After all the participants have finished, the average score is calculated and only articles that reach an average score above 6 are retained.

Table 3.1 contains the details of the protocol that guided the SLR undertaken during this research.

Table 3.1: Complete SLR Protocol

Main Question	How machine learning has been applied in epidemiological studies to identify predictive or risk factors associated to chronic diseases, and to build clinical prediction models?
Objective	This SLR aims at evaluating the potential of machine learning methods, applied to, epidemiological studies, with the goal of elucidating predictive or risk factors associated to chronic diseases and/or building prediction models.
Inclusion Criteria	Studies that mention Machine Learning or related terms (Data Mining, Supervised Learning)
	Studies that aim to identify factors (risk or predictive) associated to chronic diseases or build prediction models
	Papers that clearly specify the Machine Learning algorithm(s) adopted
	Papers that perform experiments based on data derived from epidemiologic studies (e.g.„clinical, demographic, laboratory, behavior, socioeconomic, symptoms...)
	Papers that report performance metrics for the developed approaches, based on well-estalished statistics for evaluation of ML models (specificity, sensitivity, ROC curve, accuracy, precision, recal..)
Exclusion Criteria	Studies that aren't about Human
	Studies with no experiments (e.g., reviews)
	Studies that aim at evaluating economical costs in Health
	Studies that aim at evaluating or comparing drug efficacy
	Studies that are based on text mining, solely.
	Studies that are related to genetic or genomic data, solely.
	Studies that do not clearly specify the machine learning algorithms adopted and evaluation approaches
	Studies that fail in properly evaluating the models or reporting these results
Specific Questions	Is the paper main subject about machine learning applications in the study of human diseases?
	Is the paper concerned with building prediction models or identifying risk/predictive factors associated to chronic diseases using ML algorithms?
	Does the paper methologogy involves data from epidemiologic studies, such as clinical, demographic, laboratory, behavior, socieconomic, among others?
	Is the methodology related to the training and evaluation of the ML algorithms clear?
	Does the paper perform experiments with ML algorithms and clearly reports results regarding either predictive power or disease-associated factors identified?

3.2 Search Preparation

Prior to the first phase of the SLR, several preparatory actions were required: the protocol table definition (for a description, see section 3.1); the search terms, based on the protocol and the selected databases; the specific databases to undertake the search; and the participants of the SLR.

The databases selected for this research were PubMed and Computer Science Bibliography (DBLP); they were chosen for containing a large set of articles in the subject of medicine (PubMed) and computer science (DBLP). Additionally, certain

articles were obtained from references cited by the selected papers.

Regarding the participants, given the interdisciplinary topic of this SLR, three reviewers with different backgrounds were responsible for article's analysis: a researcher with a focus on machine learning, a researcher with a focus on medicine and an undergraduate student.

The keywords were selected based on an iterative process, adjusting the query terms after initial searches. The objective was to obtain a reasonable number of articles that were as close as possible to the research area of "machine learning applied to the medical field". Various synonyms and logical combinations of the following words were considered: "machine learning", "data mining", "supervised", "risk factors" and "prediction model". After they were tested, the optimal search term was selected for PubMed as follows:

```
(
    "Machine Learning"[MeSH Terms] OR
    "Machine Learning"[Other term] OR
    "Data mining"[MeSH Terms] OR
    "Supervised Machine Learning"[MeSH Terms]
) AND
(
    "causality"[MeSH Terms] OR
    "risk factors"[MeSH Terms] OR
    "risk factor"[Other term] OR
    "protective factors"[MeSH Terms] OR
    "prediction model"[Other term] OR
    "Feasibility Studies"[MeSH Terms]
)
```

For DBLP, search was limited to medicine terms given that all indexed articles are within computer science bibliography, and the query was adjusted according to the database's syntax.

3.3 Execution

After the search was performed against the databases and all participants were provided with the list of articles, the SLR was able to be executed. The first phase entailed evaluating the title of all articles and considering whether a given article matched or was otherwise related to the main question of the SLR table. When faced with doubt on accepting or not a specific article, acceptance is recommended to the participant, because the article can be better scrutinized later in the process (MARIANO et al., 2017). Even though some degree of personal bias might have been present, it should be noted that only articles accepted by at least two, out of the three, participants were considered. The first phase of this research started with 1429 articles. After the results from individual participants were combined, 1087 articles were eliminated, thus leaving 342 articles accepted for the second phase.

The main objective during the second step was abstract evaluation, and, just as was the case with the first phase, the main question of the SLR protocol acted as a guideline. During this stage, however, exclusion criteria should also be examined, so as to remove articles not following the requirements of the protocol (MARIANO et al., 2017). The second phase started with the aforementioned 342 articles; subsequently 256 articles were eliminated, while 86 articles remained for the last phase. In this phase it is also important to note that a few points changed in the protocol definition, because of a few articles that only contained experiments and information about communicable diseases were being included.

The final phase is the most rigorous stage. It comprises a full textual reading of all the remaining articles. Each participant must read each article and assign it a score. In the case of this research, the score was based on the five specific questions defined in the SLR protocol. Each question was answered with a score ranging from 0 to 2: being 2 assigned when an article completely fulfilled the question; 0 assigned when the requirements were not met, and 1 assigned in cases of a partial fulfillment. After gathering the results from all participants, all articles with a score higher than 60% were considered accepted (MARIANO et al., 2017). During the execution of the protocol, this phase started with 86 articles. Sixteen of those were eliminated because of a low score or because of a previously unidentified exclusion criteria. Ultimately, 70 articles remained and were included in this SLR.

Figure 3.1: SLR algorithm steps, containing the initial search phase, which gather the initial list of articles for processing up until the very last phase, where the whole articles are considered

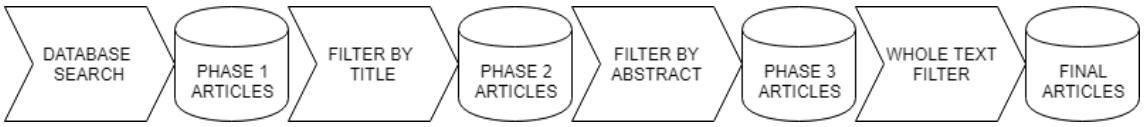
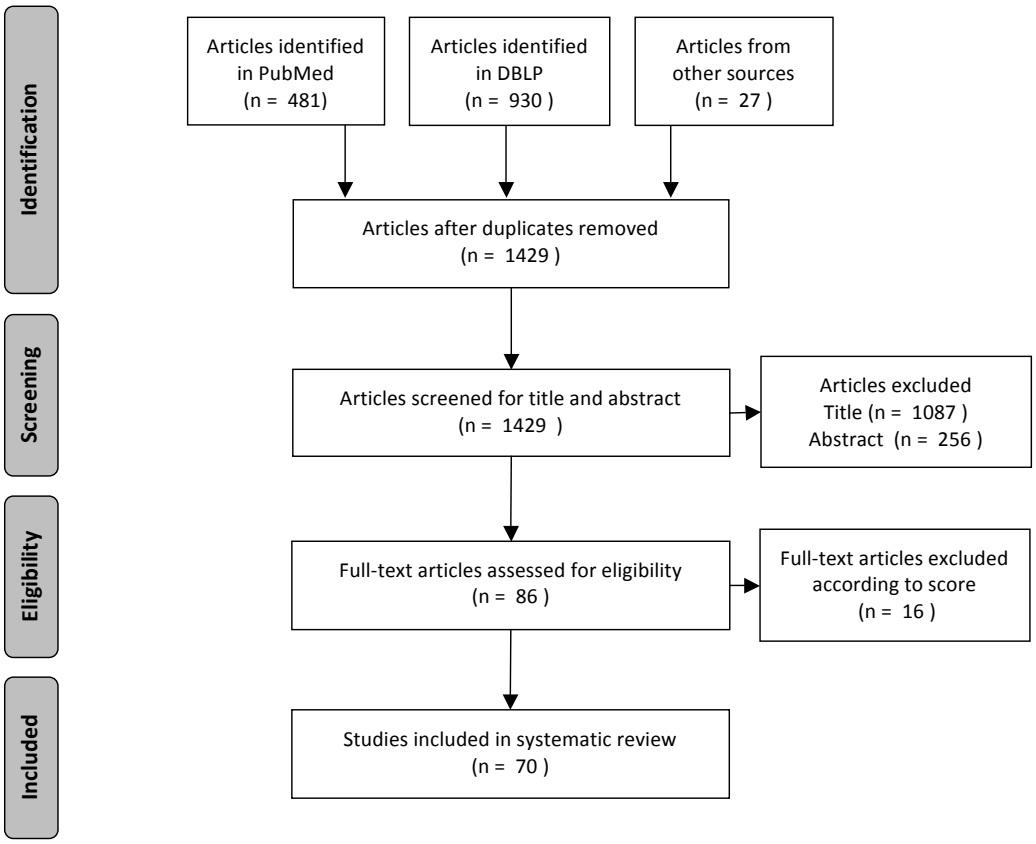


Figure 3.2: The PRISMA flowchart summarizing the phases of this SLR and the number of articles that passed or were excluded during each phase.



3.4 Results Overview

As a result of the SLR review, 70 articles were selected as compatible for studying machine learning in the context of the medical field. All the articles are referenced in appendix A. In the scope of this work, data extraction from selected studies was oriented by the research questions and performed using a standardized form, focused on the analysis of the following aspects: (a) year of publication, (b) type of analysis (i.e., prediction model, analysis of risk factors), (c) chronic disease investigated (ICD-10 code), (d) ML algorithm(s) used, and (e) algorithms performance.

Regarding the year of publication, as can be seen in Figure 3.3, the interest in ML is growing at an appreciable rate since 2010. This is probably due to the vast number of possibilities that ML algorithms provide for parsing and analyzing huge amounts of data, certainly with much less research effort and human intervention than was previously needed. Indeed, it coincides with the huge spike observed in 2010 in the area of big data in terms of popularity and development.

Figure 3.3: The yearly distribution of papers resulting from the SLR.

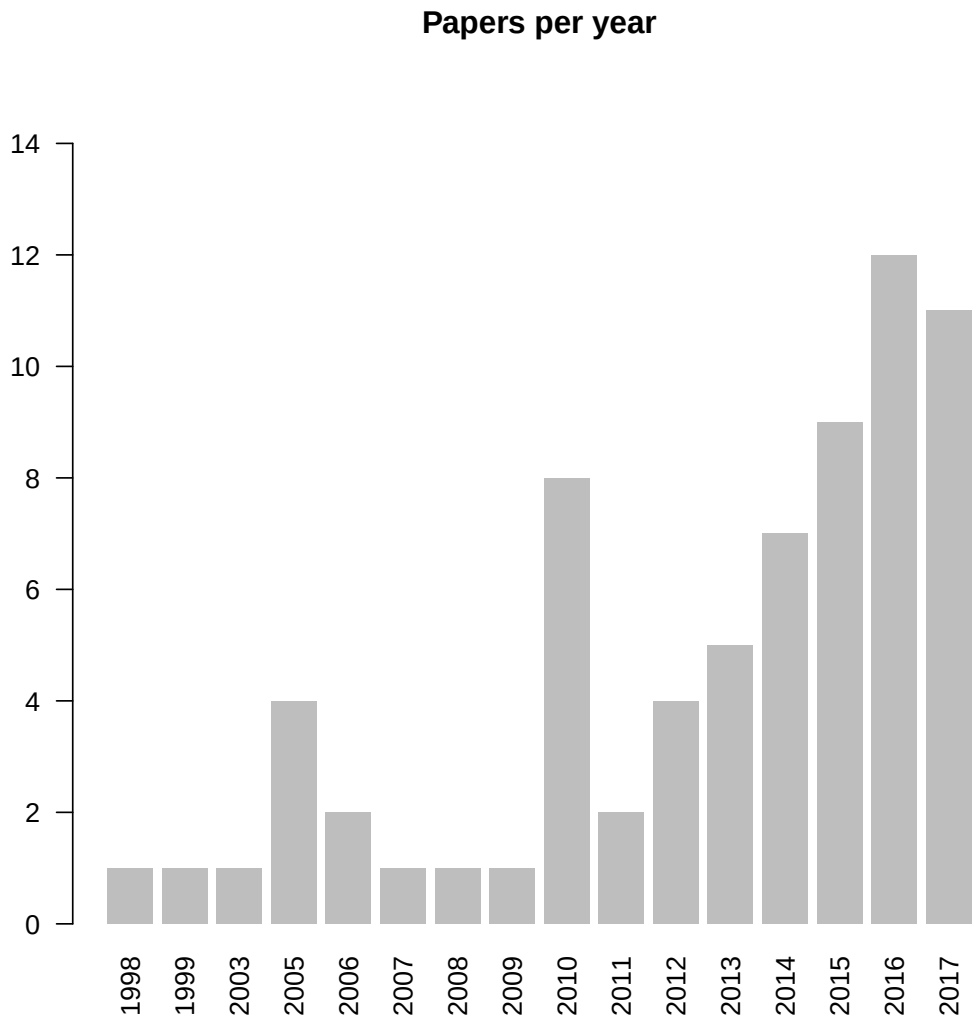


Figure 3.4 illustrates the number of articles containing each type of ML study: PD - diagnostic prediction (prediction of the development of a disease), FR - risk factors (which factors contribute to the development of a disease), and FP - prognostic prediction (prediction of death or complication after treatment). We may observe that the great majority of studies was focused on the development of diagnostic prediction models using supervised ML algorithms.

Figure 3.4: A graph illustrating the number of papers containing each type of study: diagnostic prediction (PD), risk factors (FR), prognostic prediction (FP)

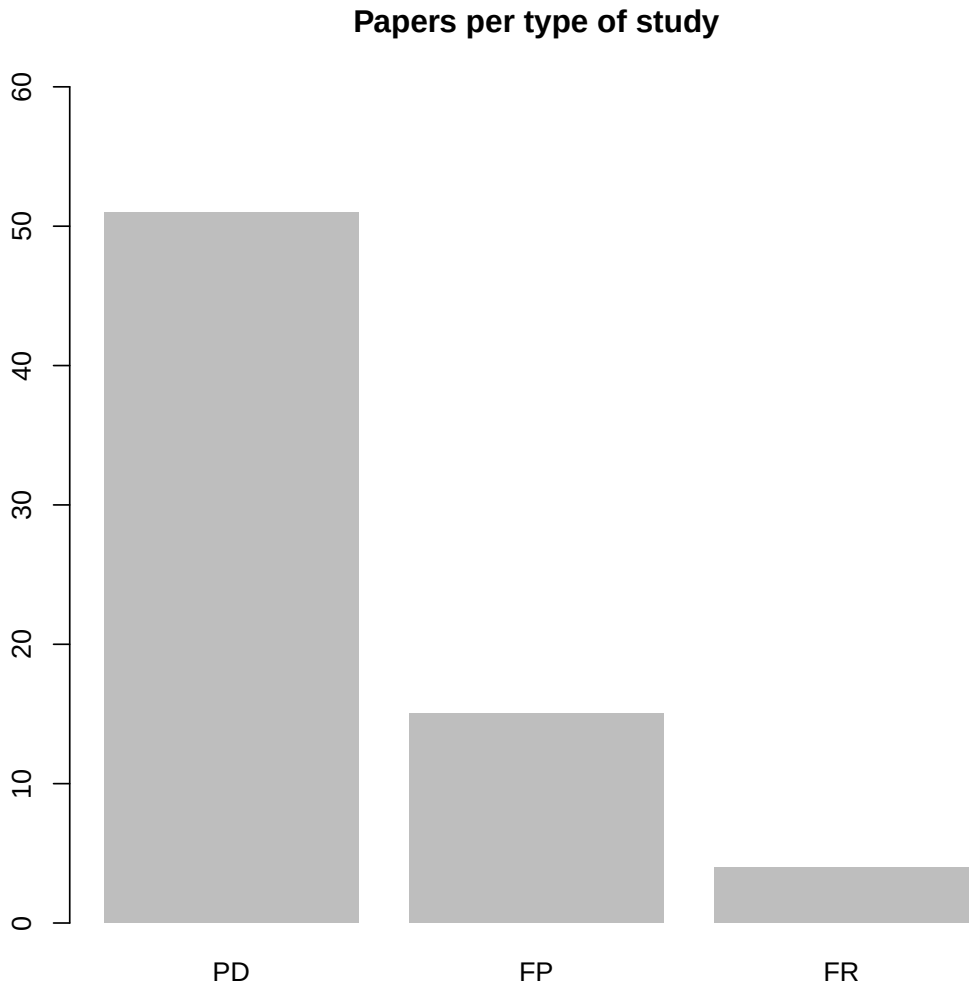


Figure 3.5 illustrates the number of articles referencing each ICD-10 chapter. ICD-10 is the 10th version of the International Classification of Diseases (ICD), which defines the universe of diseases, disorders, injuries, and other related health condition, and is the international standard for reporting diseases and health conditions for all clinical and research purposes. As a reference for the understanding of the graph, the illustrated codes are described below:

- **II** - malignant neoplasms
- **IV** - endocrine, nutritional and metabolic diseases
- **V** - mental and behavioural disorders
- **VI** - diseases of the nervous system

- **IX** - diseases of the circulatory system
- **X** - diseases of the respiratory system
- **XI** - diseases of the digestive system
- **XIII** - diseases of the musculoskeletal system and connective tissue
- **XIV** - diseases of the genitourinary system
- **XXI** - factors influencing health status and contact with health services

Figure 3.5: The number of papers according to the ICD-10 chapter

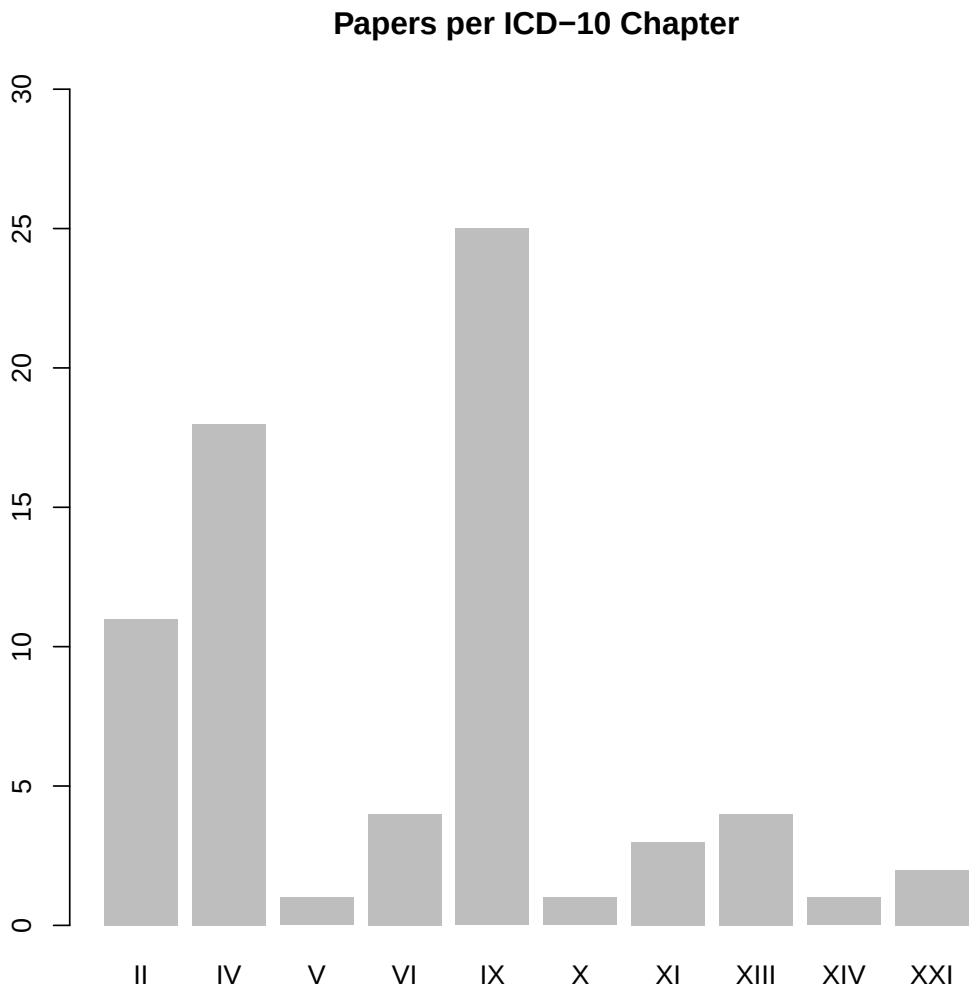
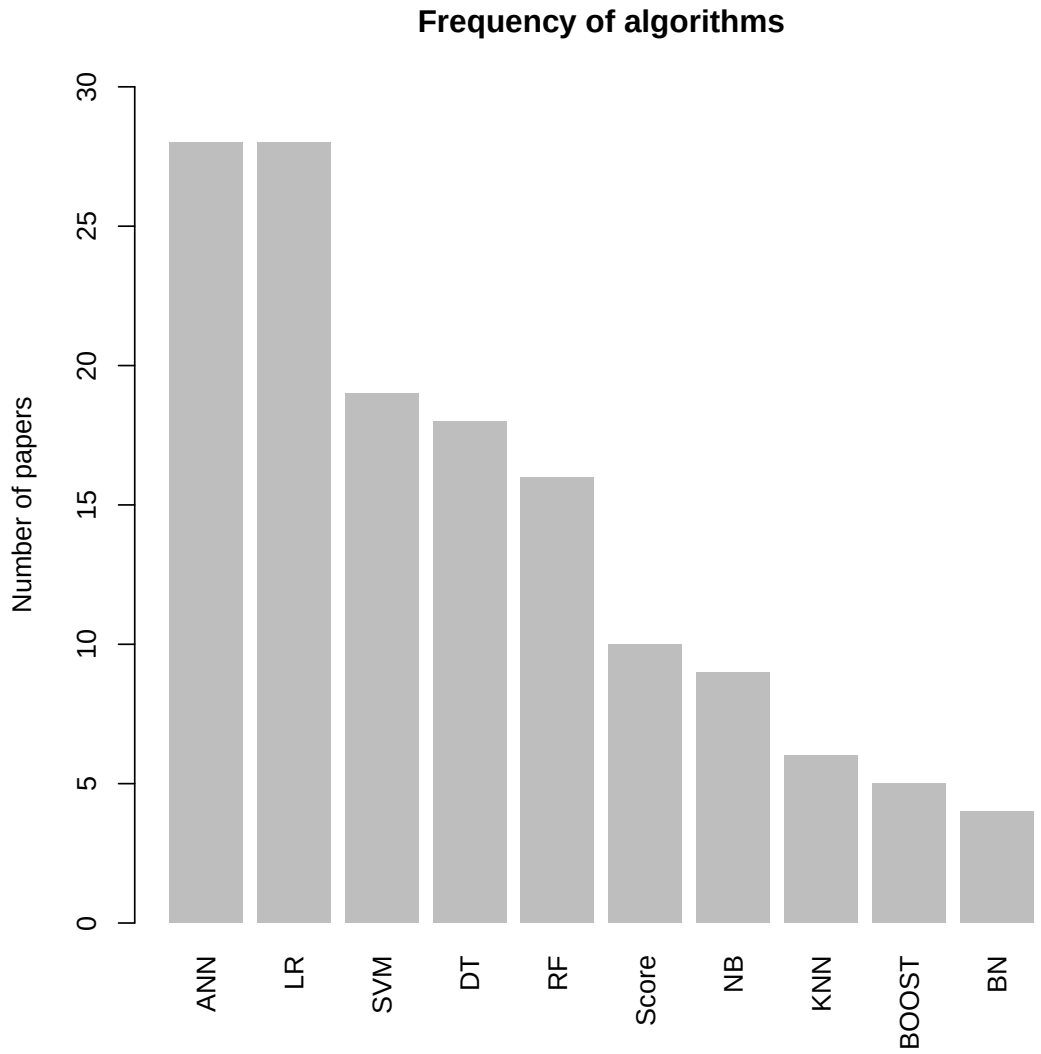


Figure 3.6 shows the top 10 algorithms according to their frequency in the included papers. As we can observe, the most commonly used algorithms among the reviewed articles are are DTs, RF, neural networks (ANN), LRs, and SVMs.

To better compare these methods, we analyzed the distribution of their accuracy and AUC scores across the 70 papers included in this SLR. Figures 3.7 and 3.8

Figure 3.6: A graph illustrating the number of articles that included the respective algorithms in their research



show these results in the form of a candle graph. As we may observe, none of the algorithms had an outstanding performance compared to others - both the accuracy and the AUC scores were largely comparable among the top 10 algorithms. For the Boosting algorithm, only one paper reports the accuracy metric for this method, justifying the lack of a proper distribution. While the medians for accuracy demonstrate great variability among the algorithms, we can see that the AUC score is quite comparable among the top 10 methods (i.e., around 75%), except for DT and NB algorithms.

Based on the results of this SLR, the most widespread ML algorithms were selected for our data analysis methodology if they could prove useful to this research, specifically with regard to disease prediction models and analysis of risk factors. As

Figure 3.7: A candle graph illustrating the ACCURACY score of each algorithm, according to the measurements of papers resulting from the SLR.

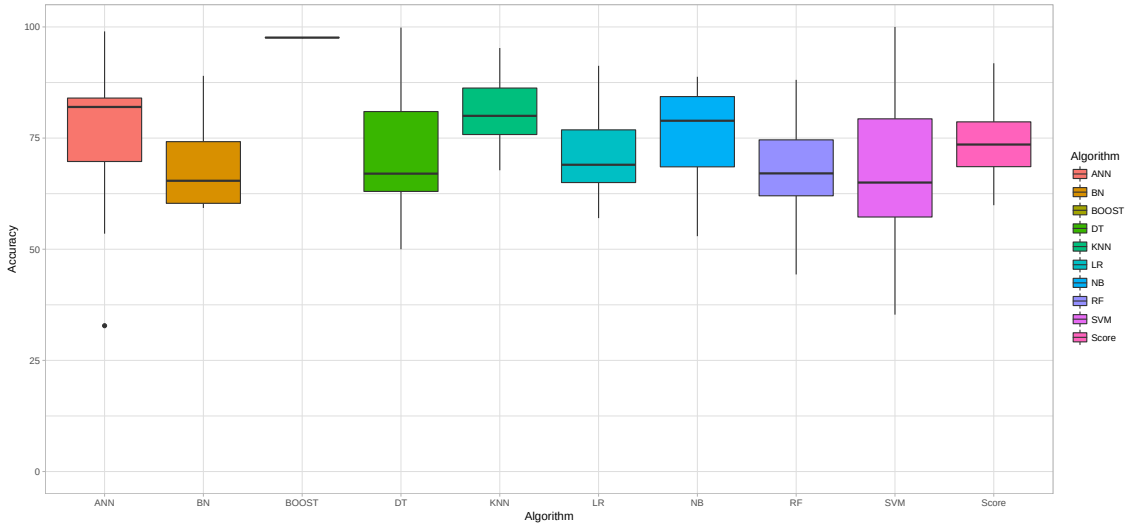
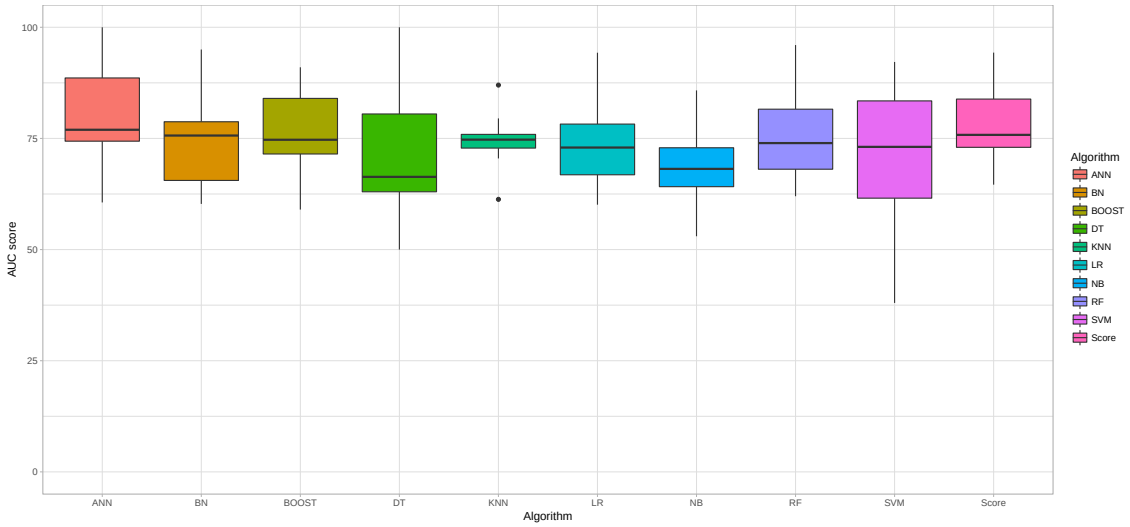


Figure 3.8: A candle graph illustrating the AUC_ROC score of each algorithm, according to the measurements of papers resulting from the SLR.



previously mentioned, the five most common algorithms according to the reviewed articles are ANN, LR, SVM, DT, and RF. Given their satisfactory and competitive performance, these methods were chosen for the ML analyses carried out in this work. The only exception was neural network, which had to be excluded from the research due to the substantial number of variables to be configured, and the computational cost for training the algorithm following the developed methodology (more details will be given on Chapter 4).

4 METHODOLOGY

The SLR results described in the previous Chapter served as a foundation for the second specific aim of this work, which concerns the development of a ML methodology for analyzing a diabetes-related dataset. Diabetes is a disease that affects millions of people annually, and despite the current technology and the potential for information to reach a wide demographic, it is still affecting an increasing number of people every year (NCD Risk Factor Collaboration et al., 2016). Therefore, the main experimental objective in this thesis was to evaluate the possibility of predicting diabetes from a predetermined set of variables collected in an epidemiological study. In addition, from the most significant variables, we aim at assessing the potential to determine which variables have the greatest impact on the result and, on account of this, to determine risk factors for diabetes. To perform these analyses, we have selected the algorithms that were identified as the most popular and most efficient during the SLR. In what follows we describe the experimental methodology adopted in the current work.

4.1 Dataset

The dataset used for this research is from the project named "Education, knowledge on risk factors, and the use of health services for women residents in a city in the south of Brazil; a population-based study" (GONÇALVES et al., 2017). The project consisted of a quiz that was given to a random selection of women (18 years old or above) in the urban municipality area of Rio Grande. The university responsible for the research was the "Universidade Federal do Rio Grande (FURG)", and the study was focused on investigating how education affects women's knowledge about key risk factors and preventive services, and in turn, how it affects the use of public health services. The set of specific diseases investigated by the study included diabetes, cancer, and cardiovascular diseases (GONÇALVES et al., 2017).

The quiz had 136 questions related to several factors including social-economic factors, access to information, access to internet and usage of technology, lifestyle habits, general knowledge regarding certain chronic diseases and risk factors, and health status. Some of these questions, however, were divided in a number of variables for facilitating digital processing of answers prior to pre-processing (see section

4.3). A total of 434 variables were collected. This dataset was primarily chosen for its potential in linking the development of diabetes in an individual with her knowledge about the disease. As a secondary reason, their preemptive thinking on digital processing also facilitated its choosing: the quiz variables were described in a way that facilitates parsing and processing.

4.2 Framework

The purpose of this research is to assess and compare the performance of ML algorithms with regard to their ability to predict a diagnosis and identify risk factors related to diabetes. For this reason, a ML framework was the optimal option for model training and evaluation. The python language with the scikit-learn framework was decided as the best option, given its popularity and the solid and comprehensive set of tools it provides for ML.

4.3 Pre-processing

Pre-processing is an important component of a ML methodology, in which the input data needs to be processed and standardized so that it is able to ensure accurate and meaningful analysis by ML algorithms. In the context of this work, data pre-processing was a time-consuming task, in which a large part of the practical development of this research was concentrated.

This pre-processing was initially divided into two steps, a global and a local step. The global pre-processing was executed against the whole dataset, without considering the division of data into training and testing sets. At this stage, pre-processing included everything that could be globally executed without leaking information from the testing dataset to the training one, a phenomenon that could introduce bias in the model's results. After the dataset was divided into training and testing sets, a local pre-processing was applied separately to the training and testing data; this pre-processing consisted of transformations and standardizations that were, in some respect, based on information about data distribution, and thus, could cause the data leakage phenomenon.

Data leakage is when data from the testing set can leak into the training set,

causing the algorithms to indirectly receive information from the testing set during the training phase. For example, if an average function is applied over the complete dataset, the calculated average will include data from the testing set as well, and this may cause an improper positive impact on the training process of the model.

4.4 Global Pre-processing

The global pre-processing entailed six steps. Because these steps may cause the removal of unexpected fields or entries, they were applied in a predefined order to avoid conflicts in the data. In what follows, these six steps are presented and explained.

4.4.1 Quiz Fields

The dataset and the quiz contained certain diverging fields. There were fields present in the dataset that were not present in the quiz, which probably derived from a combination of other fields. However, since it is not possible to decipher exactly how these fields were related to the actual data, they were removed from the dataset.

4.4.2 Transformation of Field Values

The dataset contained a small number of invalid variable values, which were determined by the quiz. For example, a question could be answered with “Yes”, “No”, or “I don’t know”. Those values were originally converted into a numerical representation, such as 0, 1, or 88. The value of “I don’t know” is far greater than the other two. Since a normalization was later applied to the data, this could have created problems in the training process, because the “I don’t know” response would have represented a much higher weighting than the others. For that reason, those invalid values were removed and they were set as null for later recalculation (for more on the recalculation, see section 4.5), allowing the values to remain in the dataset without affecting the result too much.

4.4.3 Ignored Fields

Some fields were completely ignored during the training process, because they were completely unrelated to the research, were text inputs, or contained invalid values that did not have a clear meaning in the context of this work. These invalid values could affect the training performance or even invalidate results, their removal was required.

4.4.4 Mandatory Variables

The current research was based on an output variable that contained the answer to the question: "Do you have diabetes?". Thus, the variable containing the answer to this question was mandatory, and those responses that did not contain an answer or that represented an invalid answer were removed.

4.4.5 Balanced Class Size

The dataset used during this research presented a significant imbalance in the sizes of output classes for the diabetes answer. The ratio of classes was 1:20; this means that for every “Yes”, there were 20 answers indicating “No”. As Alghamdi et al. (2017) discuss, “In practice, several studies have shown that better prediction performance can be achieved by having balanced data”. For that reason, during the preprocessing phase we applied an upsampling strategy such that for every “Yes” entry found, an arbitrary “No” from the dataset was included, therefore, maintaining a 1:1 ratio. Since the dataset used by this research followed no pattern as to the order of the entries, the applied upsampling function selected the “No” entries based on their position, the closest entry after the “Yes” was the one selected.

4.4.6 Invalid Values

After all the previous pre-processing steps, there were still multiple fields with empty values and there were participants who answered a negligible number of questions in the quiz. As a way to keep the results sane and not include invalid or

duplicated data, all columns and rows of the dataset were analyzed and those that contained more than 5% of null values, were completely removed from the training data. This 5% value was chosen so as not to remove too great a number of entries.

4.5 Local Pre-processing

Local pre-processing was applied separately to the training data and testing data, so that information from the testing data would not leak into the training data. After data partitioning, the training data is pre-processed considering only data distribution and characteristics of this subset of instances. Next, testing data is pre-processed considering characteristics from the complete dataset to guarantee that data distributions and other aspects are comparable among training and testing examples. The reason for this difference is that, in the real world, we do not have the future values to influence our predictions, but we do have the past data when testing the predictions.

The following two methods of pre-processing were applied locally onto the dataset (in the order in which they were executed): firstly, all invalid and missing values were filled with the average value of the respective attributes; secondly, the data was normalized so that all numeric attributes had their values within the interval [0, 1].

4.6 Algorithms

The selection of the ML algorithms trained during this research was based on the results from the SLR. Except for the ANN, the four most popular algorithms were chosen: DT, RF, SVM, and LR. They represent a variety of induction biases within the field of ML and provide the opportunity for comparison with the previous work in this research area.

The ANN algorithm has a substantial number of variables to be configured, and the computational cost for training the algorithm is significantly higher; this ensured that a full-scale training, which followed the same protocol as the other algorithms, was not practical on account of available resources and time.

4.7 Training

The training was executed by means of a NCV, wherein three main loops are nested into each other; it is a highly efficient way of dealing with parameter and feature selection. The strategy functions by separating the training process into three steps: feature selection, model tuning, and model training (for more details, see 2.6).

4.7.1 Feature Selection

Feature selection is the step that selects the optimal variables for the training process. Normally, the best solution when working with a feature selection is to test all combinations of all variables; this was not possible in this research because of the large number of variables (more than 200) after pre-processing. This step entailed choosing the optimal number of variables by an incremental analysis, the process was executed by means of a pre-implemented function in Python's machine learning framework, recursive feature elimination and cross-validated (RFECV). This function iteractively removes the least relevant variables for the given algorithm or model, and for each attempt it performs a k-fold cross-validation training on the algorithm. In the end, the variables associated to the best score are selected.

4.7.2 Model Tuning

Model tuning selects the optimal set of parameters for the defined model. This model tuning was carried out by using the GridSearchCV from Python for each algorithm selected, which receives a list of parameters and possible values according to the specific algorithm considered and, thereafter, exhausts every possible combination so as to find the one with the optimal score. For each combination tried, a k-fold cross-validation training is executed with the feature selection embedded; this, ultimately, determines the optimal model's parameters and features.

The adjustable parameters available for each algorithm may be implementation specific, and thus their detailed description is not in the scope of this research. In the following subsections, a brief summary of their definition is given. The defini-

tion is based on the framework documentation (for more specific details, see Géron (2017)).

4.7.3 Logistic Regression Model Tuning

Table 4.1: Logistic Regression Parameters

Parameter	**Values**
Penalty	L1, L2
C	0.001, 0.112, 0.223, 0.334, 0.445, 0.556, 0.667, 0.778, 0.889, 1.0

In Table 4.1, the term **Penalty** is used to specify the norm used during the penalization, during which both supported L1 and L2 functions are tested. **C** denotes the inverse of regularization strength, for which smaller values mean stronger regularization, on this research 10 values ranging from 0.001 to 1.0 and linearly spaced were tested.

4.7.4 SVM Model Tuning

Table 4.2: SVM Parameters

Parameter	**Values**
C	0.001, 0.112, 0.223, 0.334, 0.445, 0.556, 0.667, 0.778, 0.889, 1.0

In Table 4.2, **C** is the penalty parameter for the error term and tells the SVM optimization how much it should avoid misclassifying each training example by choosing the margins width. Large values of C will cause a smaller-margin hyperplane to be fitted, whereas a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane. Here, 10 values, ranging from 0.001 to 1.0 and linearly spaced were tested. The predefined kernel used during this research was the linear SVM kernel.

4.7.5 Decision Tree Model Tuning

In Table 4.3, **Criterion** denotes the function used to measure the quality of a node split, for which both supported values, “gini” and “entropy” were tested. **Splitter** is the strategy used to choose the split at each node; both supported values

Table 4.3: Decision Tree Parameters

Parameter	Values
Criterion	gini, entropy
Splitter	best, random

"best" and "random" were tested.

4.7.6 Random Forest Model Tuning

Table 4.4: Random Forest Parameters

Parameter	Values
N. Estimators	3, 6, 9, 12, 15, 18, 21, 24, 27, 30
Criterion	gini, entropy

In Table 4.4, **N. Estimators** denotes the number of trees in the forest, for which 10 values between 3 and 30 were tested. **Criterion** is, again, the function that determines strategy to evaluate the quality of a node split, for which both supported values "gini" and "entropy" were tested.

4.7.7 Model Training

At the outermost loop, a normal k-fold cross-validation was executed, for making sure that our algorithm has not been overfitted for the training set. In this step, the local pre-processing is executed for each k-fold and then the inner loops for model tuning and feature selection are run as previously explained. The result from each of the k-folds is kept and their average is calculated as the general performance of the algorithm.

4.8 Output Reporting

For later evaluations of the results and algorithms comparison, a standard representation of results were adopted for all ML algorithms run. For each iteration from the outermost loop (model training) the optimal variables, optimal parameters, and three scores were saved. The saved scores were selected based on the popularity on the resulting SLR articles and their availability on the applied framework: ROC_AUC, F1, and ACCURACY. The optimal variables and parameters

were selected based on their score in the inner loops. For comparing variables and parameters, when sorting was required, a ROC_AUC scoring function was applied. In this way, the algorithm with the top score was selected as the most effective, and the optimal variables and parameters were extracted from it.

For each iteration of the inner loop model tuning, the optimal variables were saved and later exported to a CSV file. Each iteration is saved as one line in the output CSV file; each column of the file is a variable; and for each [column, line] pair, a value of 0 or 1 is written, where 0 denotes that the variable was not used, and 1 denotes that it was used. This process allows the researcher to obtain a detailed analysis of the distribution of variable's inclusion in the trained models, the average number of variables selected, and, if necessary, the frequency of selection of a particular variable for the algorithms used.

5 RESULTS AND COMPARISON

As a result of the extensive training (described in the previous chapter), the accuracy scores, variable ranking and the average number of variables that constitute a good prediction were produced. In the following analysis, these scores are discussed and, ultimately, compared to one another and, moreover, to the articles resulting from the SLR.

The source code, configuration files and detailed results of the whole experiment are available at <https://github.com/HPerin/tcc-machine-learning.git>. For each algorithm, a table with those variables considered most important and relevant is presented. The importance of a variable is a rather difficult concept to define. The variable might be important during only a single stage of training, or it might be important only for one particular algorithm. For this research, the objective was to define the most important variables in a perspective of risk factor analysis, which, in turn, could be used in practice (even without the algorithms). Thus, the method adopted for choosing the important variables entailed that the variable should be contained in all outer loop iterations. Nonetheless, in the results described below, we present all variables selected in at least five out of the 10 iterations of the k-fold cross-validation procedure as possible risk factors for diabetes.

5.1 Logistic Regression

Figure 5.1 illustrates all the scores calculated for each iteration, with the average of all executions in the last column. Table 5.1, shows the average scores obtained for a LR model.

Table 5.1: LR average scores and their standard deviation

Score	Average	Standard Deviation
AUC_ROC	0.739	0.133
F1	0.733	0.143
ACCURACY	0.740	0.132

Appendix B shows which variables were most often used during the inner loop training, it is a representation of the 30 most used variables. The average size of variables subset used by the LR algorithm was 51.02.

Table 5.2 lists the variables that were present for a minimum of five iterations during training, and their meanings are subsequently described. Among these,

Figure 5.1: LR scores computed from 10-fold cross-validation. The rightmost columns depics the average over all iterations.

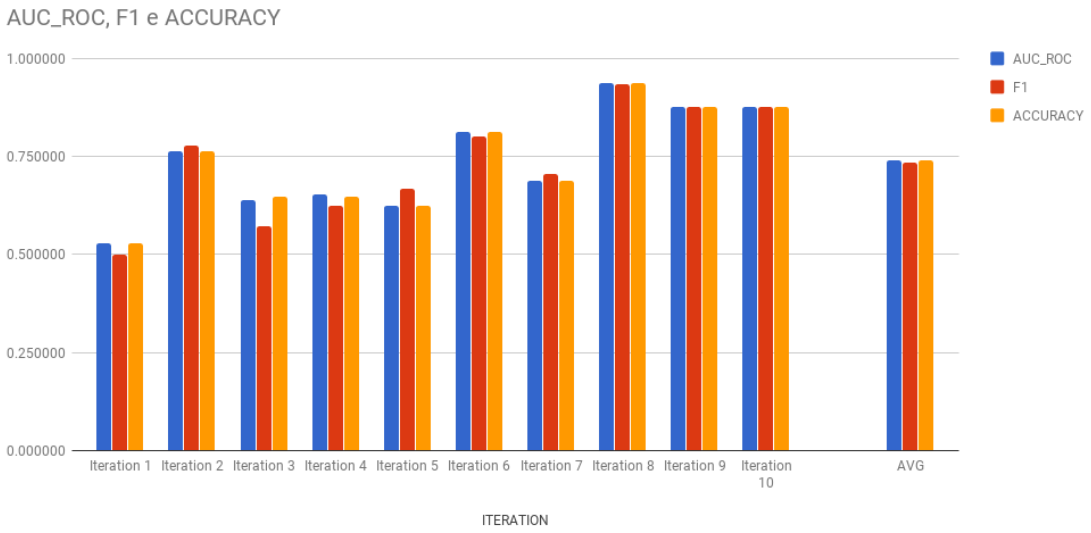
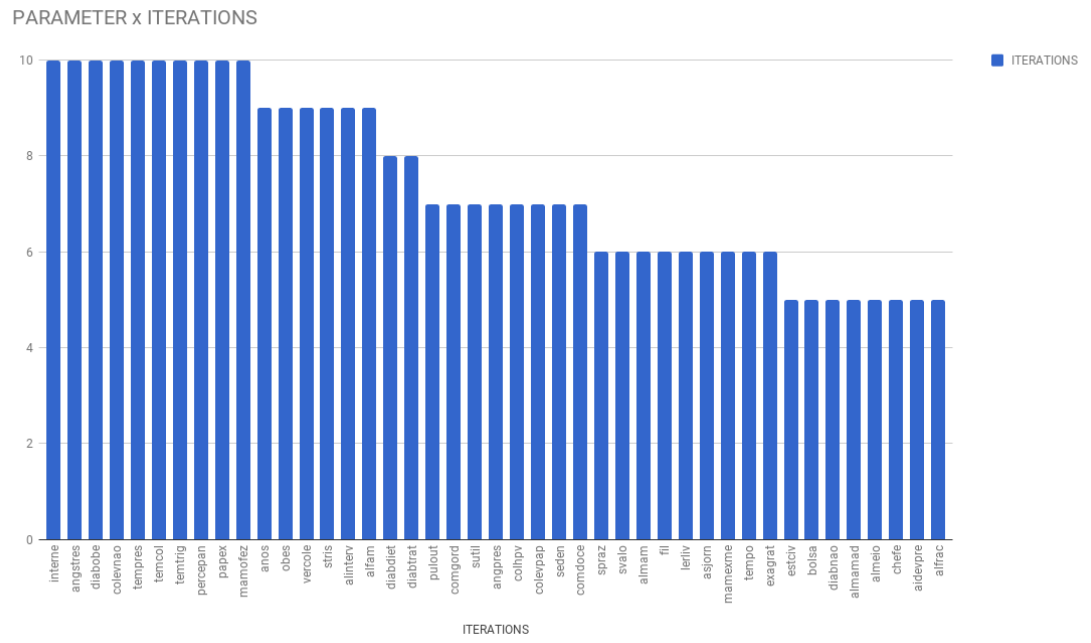


Figure 5.2: Variables selected for at least five iterations of the LR outerloop training.



six variables were selected in all iterations of the outer CV, thus being the most important variables for the prediction of diabetes within the current work.

Table 5.2: Description of the most important variables meanings, according to the results of the LR algorithm.

Iterations	**Variable**	**Meaning**
10	interne	Internet usage frequency
10	angstres	Knowledge question: does stress lead to heart attack?
10	diabobe	Knowledge question: what makes obesity worse?
10	colevnao	Avoid or prevent uterine cervix cancer.
10	tempres	Has high blood pressure
10	temcol	Has high cholesterol
10	temtrig	Has high triglycerides
10	percepan	How well have been your health on the last 12 months
10	papex	Has done preventive exams for uterine cancer
10	mamofez	Has done mammography before.
9	anos	Age
9	obes	Is Obese
9	vercole	Has done cholesterol exam or has cholesterol
9	stris	Has been feeling sad lately
9	alinterv	Knowledge question: how often should a baby by breastfed.
9	alfam	Family taught about breastfeeding
8	diabdiet	Poor diet makes diabetes worse
8	diabtrat	Knowledge questions: not treating can make diabetes worse
7	pulout	Knowledge questions: what leads to lung cancer
7	comgord	Eat greasy foods
7	sutil	Do you have a useful role in your life
7	angpres	Knowledge question: high blood pressure leads to heart attack
7	colhpv	Knowledge question: virus leads to uterine cancer
7	colevpap	Knowledge question: preventive exams help avoid uterine cancer
7	seden	Knowledge question: being sedentary leads to heart attack
7	comdoce	Eat a lot of candy
6	spraz	Do you feel pleasure on your daily activities
6	svalo	Do you feel worthless
6	almam	Problems using the baby bottle whilst breastfeeding
6	fil	Has kids
6	lerliv	Reading books frequency
6	asjorn	Watch news on television
6	mamexme	Does making the exam help find breast cancer earlier
6	tempo	Waiting time affects the decision of which clinic to visit
6	exagrat	Free exams affects the decision of which clinic to visit
5	estciv	Marital status
5	bolsa	Do you receive financial aid from the government
5	diabnao	Does not know what makes diabetes worse
5	almamad	Sees a problem on using the baby bottle when breastfeeding
5	almeio	Has learned about breastfeeding through media
5	chefe	Are you the householder
5	aidevpre	Condoms help avoid AIDS
5	alfrac	In your opinion, are mothers with weak milk suitable for breastfeeding

In Tables 5.3 and 5.4, the number of iterations that each parameter has been selected as the optimal value is shown. We may observe that L2 regularization yielded the best results in 8 out of the 10 folds of the NCV, while there was no clear

prevalence of a specific C value as the best performing parameter.

Table 5.3: LR - C parameter iteration use count

C	Use Count
0,334000	3
0,223000	2
0,556000	2
0,112000	2
0,778000	1

Table 5.4: LR - penalty parameter use count

Penalty	Use Count
L2	8
L1	2

5.2 Decision Tree

Figure 5.3 illustrates all the scores calculated for each iteration, with the average of all executions in the last column. Table 5.5, shows the average scores computed from the outerloop of the 10-fold NCV. We may observe that the average performance metrics are lower as compared to the LR models, and that the standard deviation for the F1 measure is higher in DT in contrast to LR.

Table 5.5: DT average scores and their standard deviation.

Score	Average	Standard Deviation
AUC_ROC	0.622	0.126
F1	0.588	0.152
ACCURACY	0.622	0.127

Appendix B shows the top 30 variables most often used during the inner loop training. The average size of variable subsets used by the algorithm was 52.44.

Table 5.6 lists the variables that were present for a minimum of five iterations during training, and their meanings are subsequently described. According to the defined criterion, a smaller number of relevant variables was identified in contrast to the results obtained by LR.

In Tables 5.7 and 5.8, the number of iterations that each parameter has been selected as the optimal value is shown. According to these results, DTs trained with the Gini index and with the random splitter tend to achieve the best performance.

Figure 5.3: DT scores computed from 10-fold cross-validation. The rightmost columns depics the average over all iterations.

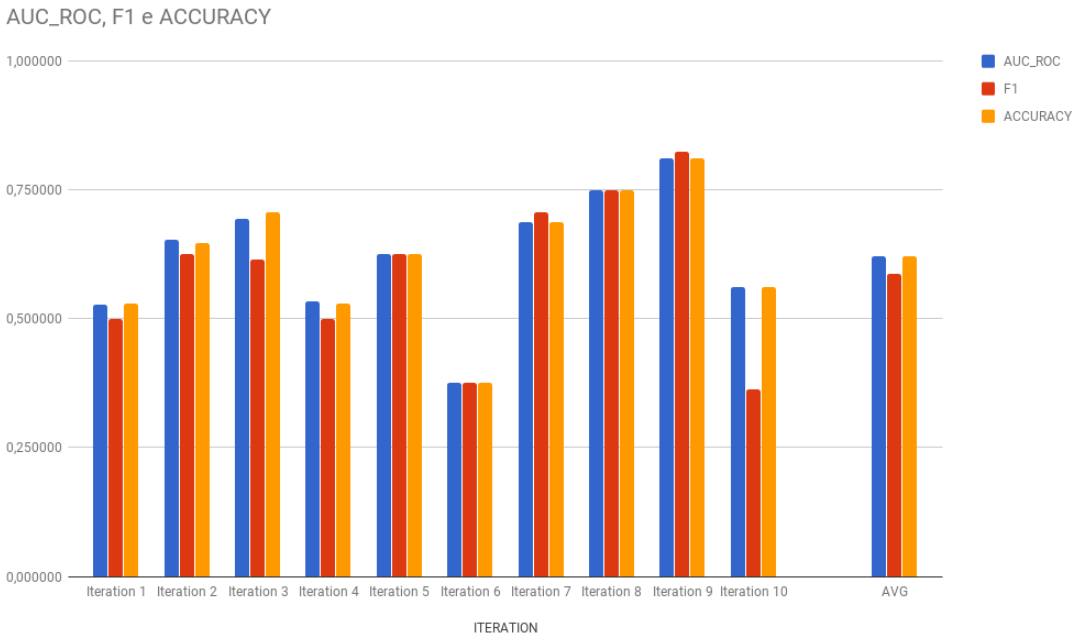


Figure 5.4: Variables selected for at least five iterations of the DT outerloop training.

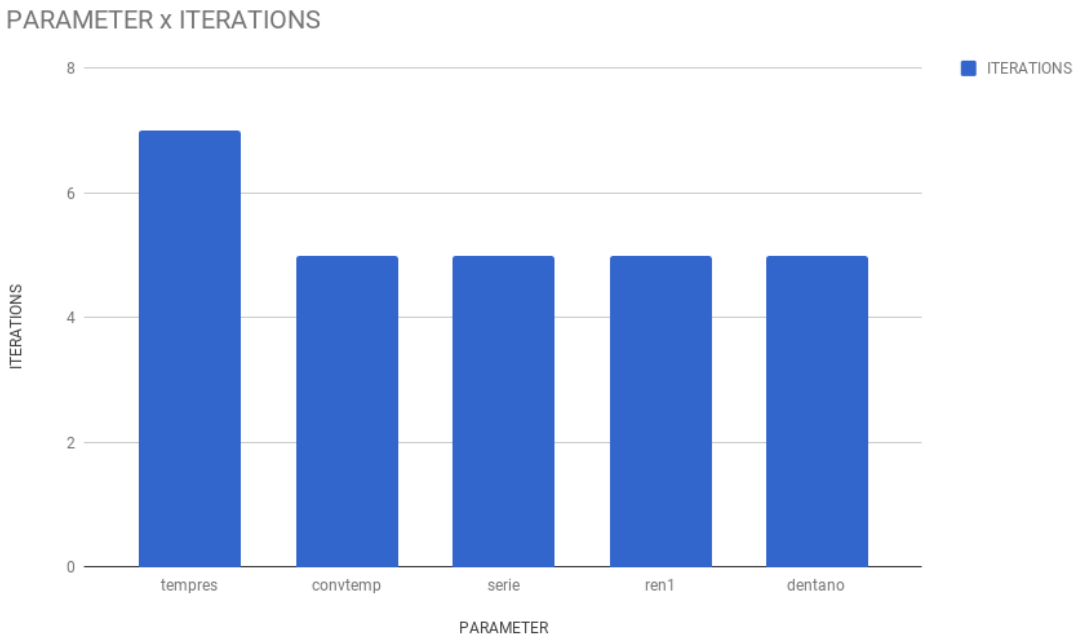


Table 5.6: Description of the most important variables meanings, according to the results of the DT algorithm.

Iterations	**Variable**	**Meaning**
7	tempres	Has high blood pressure
5	convtemp	For how long has had medical insurance
5	serie	Scholarity
5	ren1	Income
5	dentano	Has visited a dentist in the previous year

Table 5.7: DT - criterion parameter iteration use count

Criterion	Use Count
Gini	8
Entropy	2

Table 5.8: DT - splitter parameter iteration use count

Splitter	Use Count
Random	6
Best	4

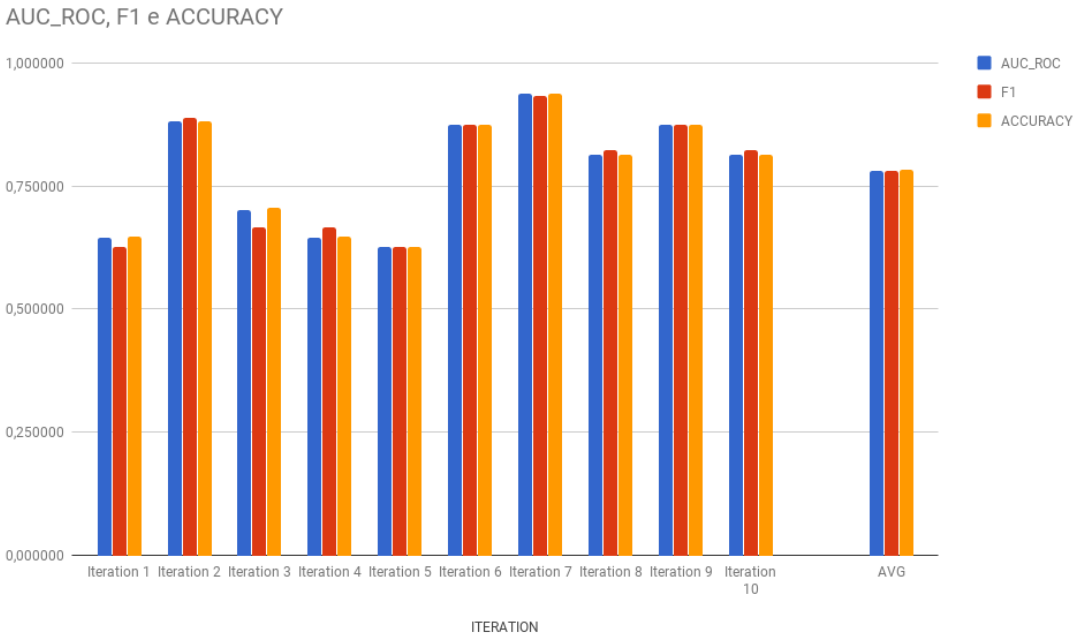
5.3 Random Forest

Figure 5.5 illustrates all the scores calculated for each iteration, with the average of all executions in the last column. Table 5.9, shows the average scores for the RF algorithms in the NCV.

Table 5.9: RF average scores and their standard deviation.

Score	Average	Standard Deviation
AUC_ROC	0.781	0.116
F1	0.780	0.120
ACCURACY	0.781	0.115

Figure 5.5: RF scores computed from 10-fold cross-validation. The rightmost columns depics the average over all iterations.



Appendix B shows which variables were most often used during the inner loop training, limited to the top 30 most frequent variables. The algorithm's average overall variable subset size was 98.02.

Figure 5.6: Variables selected for at least five iterations of the RF outerloop training.

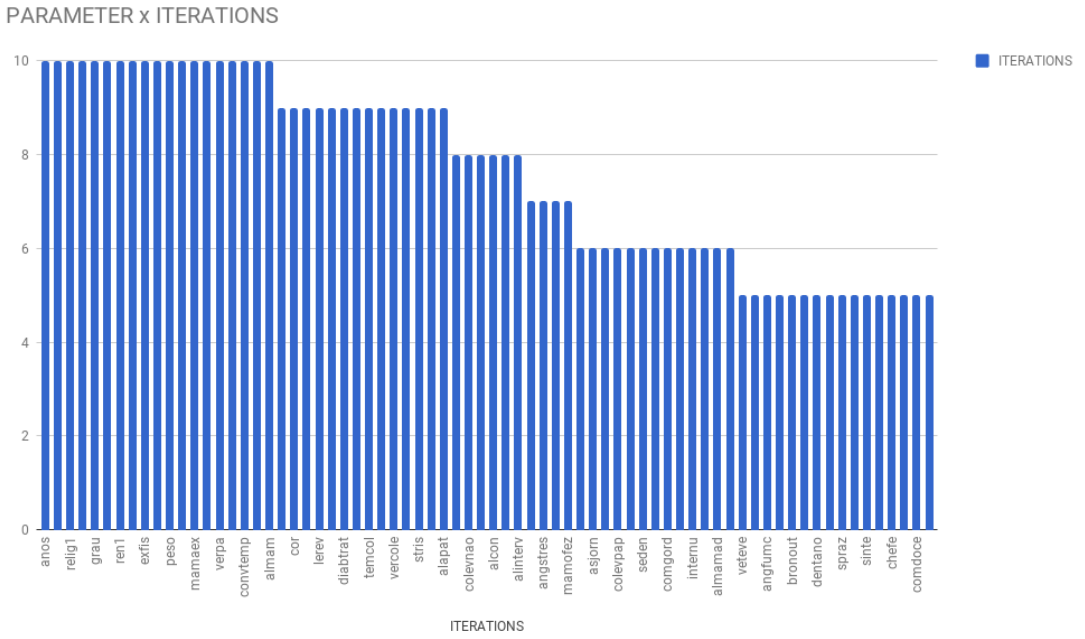


Table 5.10 lists the variables that were present, as best variables, for a minimum of five iterations during training, and their meanings are subsequently described. Among these, 18 variables were selected in all iterations of the NCV and are thus considered the possible risk factors for diabetes.

In Tables 5.11 and 5.12, the number of iterations that each parameter has been selected as the optimal value is shown. Again, we observed higher predictive performance associated with the Gini index as the node split method.

5.4 SVM

Figure 5.7 illustrates all the scores calculated for each iteration, with the average of all executions in the last column. Table 5.13, shows the average scores for the SVM models.

Appendix B shows the top 30 variables most often used during the inner loop training. On average, the overall variable subset size used by the algorithm was 49.11.

Table 5.14 lists the variables that were present for a minimum of five iterations during training, and their meanings are subsequently described. Among these, two variables were selected in all iterations of the NCV and are thus considered the

Table 5.10: Description of the most important variables meanings, according to the results of the RF algorithm.

Iterations	Variable	Meaning
10	anos	Age
10	pesmora	How many people live in your house
10	relig1	Religion
10	serie	Scholarity
10	grau	Scholarity
10	empreg	Working state
10	ren1	Income
10	ouvrad	Listen to the radio
10	tempres	Has high blood pressure
10	peso	Weight
10	percepan	Health perception in the last 12 months
10	mamaex	Last mamma exam
10	mamonde	Location of the last mamma exam
10	verpa	Ever measured blood pressure
10	locons	Last doctor visit
10	convtemp	For how long have had medical insurance
10	baitemp	For how long do you have doctor visits in the local clinic
10	almam	Problems using the baby bottle whilst breastfeeding
9	estciv	Marital status
9	cor	Skin color
9	lerjor	Reading newspaper frequency
9	lerev	Reading magazines frequency
9	interne	Internet usage frequency
9	diabtrat	Knowledge questions: not treating can make diabetes worse
9	supamig	Social events attendance frequency
9	temcol	Has high cholesterol
9	papmes	Time since last uterine cancer exam
9	vercole	Has done cholesterol exam or has cholesterol
9	medtemp	For how long have you had the same doctor
9	stris	Has been feeling sad lately
9	alleite	For how long do you think a baby should only receive breastfeeding
9	alapat	How do you feel about the grandmother (from the father family) help during breastfeeding
8	lerliv	Reading books frequency
8	colevnao	Knows how to avoid Uterine Cervix Cancer
8	temtrig	Has high triglycerides
8	alcon	How much do you know about breastfeeding
8	altemp	For how long should a baby be breastfed
8	alinterv	Knowledge question: How often should a baby by breastfed.
7	fumo	Do you smoke
7	angstres	Knowledge question: does stress lead to heart attack
7	pulout	Knowledge questions: what leads to lung cancer
7	mamofez	Has done mammography before
6	fil	Has kids
6	asjorn	Watches news on television
6	angpres	Knowledge question: high blood pressure leads to heart attack
6	colevpap	Knowledge question: preventive exams help avoid uterine cancer
6	supsoc	How often someone has helped you when in need (for example, when sick)
6	seden	Knowledge question: being sedentary leads to heart attack
6	obes	Is obese
6	comgord	Eats greasy foods
6	medef	Has a specific doctor for when you have a medical issues
6	internu	How many times has been admitted to the hospital in the last 12 months
6	ssono	Sleeps poorly
6	almamad	Sees a problem on using the baby bottle during breastfeeding
6	alfam	Family taught about breastfeeding
5	veteve	Television watching frequency
5	infsaud	Receives good health tips
5	angfumc	Knowledge question: smoking leads to heart attacks
5	bronao	Does not know what leads to bronchitis
5	bronout	What leads to bronchitis
5	diabobe	Knowledge question: what makes obesity worse
5	dentano	Has visited a dentist in the past year
5	mednome	Knows your doctor's name
5	spraz	Feels pleasure during daily activities
5	sdeci	Has a hard time making decisions
5	sinte	Has been losing interest in things
5	aldoen	Believe breastfed children catch fewer diseases
5	chefe	Is the head of the household
5	colhpv	Knowledge question: virus leads to uterine cancer
5	comdoce	Eats a lot of candy
5	sdoca	Has frequent headaches

Table 5.11: RF - criterion parameter iteration use count

Criterion	Use Count
Gini	8
Entropy	2

Table 5.12: RF - N. estimators parameter iteration use count

N. Estimators	Use Count
21	3
30	3
24	2
12	2

Table 5.13: SVM average scores and their standard deviation.

Score	Average	Standard Deviation
AUC_ROC	0.709	0.112
F1	0.700	0.115
ACCURACY	0.709	0.112

Figure 5.7: SVM scores computed from 10-fold cross-validation. The rightmost-columns depics the average over all iterations.

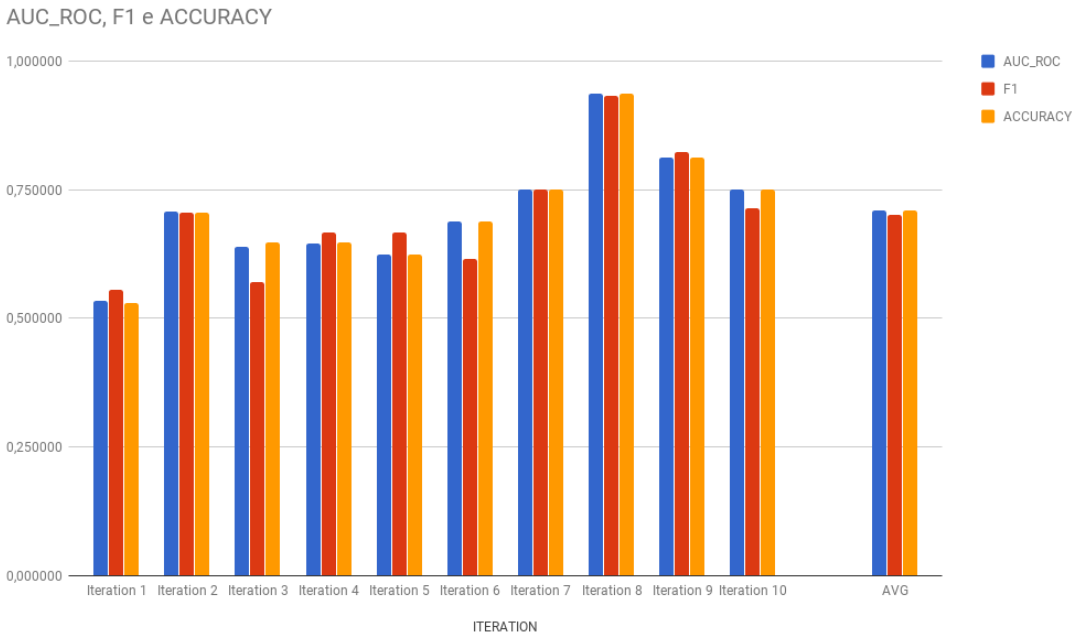
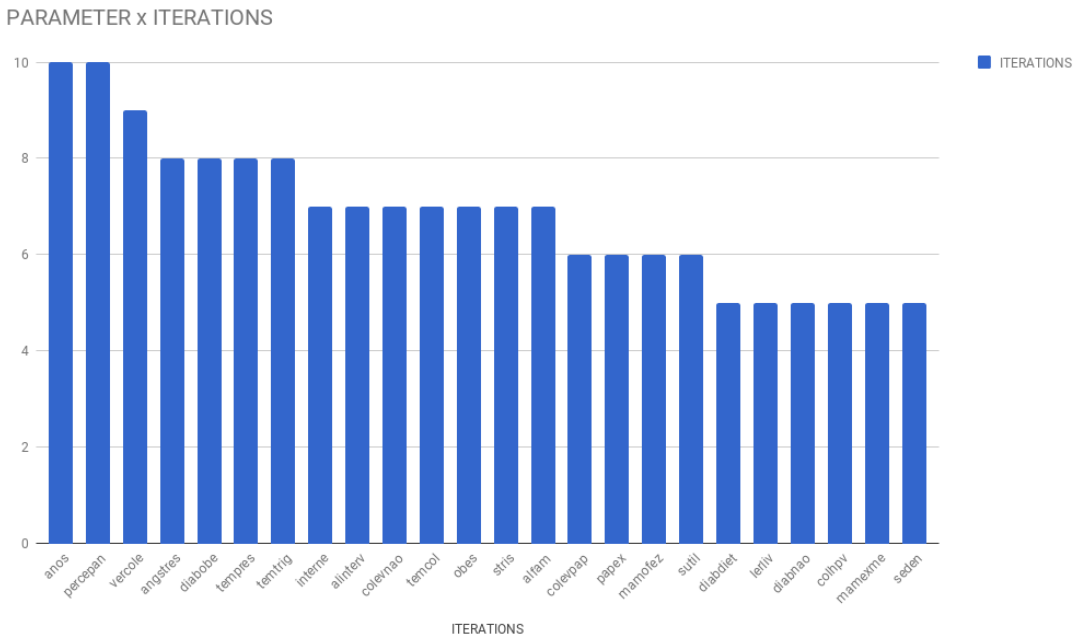


Figure 5.8: Variables selected for at least five iterations of the SVM outerloop training.



possible risk factors for diabetes.

Table 5.14: Description of the most important variables meanings, according to the results of the SVM algorithm.

Iterations	**Variable**	**Meaning**
10	anos	Age
10	percepan	How well has felt about your health in the last 12 months
9	vercole	Has done cholesterol exam or has high cholesterol
8	angstres	Knowledge question: stress leads to heart attack
8	diabobe	Knowledge question: what makes obesity worse
8	tempres	Has high blood pressure
8	temtrig	Has high triglycerides
7	interne	Internet usage frequency
7	alinterv	Knowledge question: how often should a baby by breastfed.
7	colevnao	Know how to avoid uterine cervix cancer
7	temcol	Has high cholesterol
7	obes	Is obese
7	stris	Has been feeling sad lately
7	alfam	Family taught about breastfeeding
6	colevpap	Knowledge question: preventive exams help avoid uterine cancer
6	papex	Has done preventive exams for uterine cancer
6	mamofez	Has done mammography before
6	sutil	Has a useful role in life
5	diabdiet	Poor diet makes diabetes worse
5	lerliv	Books reading frequency
5	diabnao	Does not know what makes diabetes worse
5	colhpv	Knowledge question: virus leads to uterine cancer
5	mamexme	Does making the exam help find breast cancer earlier
5	seden	Knowledge question: being sedentary leads to heart attack

In Table 5.15, the number of iterations that each parameter has been selected as the optimal is shown.

Table 5.15: SVM - C parameter iteration use count.

C	Use Count
0.112	4
0.778	3
0.001	2
0.889	1

5.5 Execution

During the execution of the training, a few unexpected problems arose. On the first run, several algorithms received a perfect score for predicting diabetes. This is a clear sign of overfitting. The cause behind these problems and how these issues were solved are discussed below.

The first consideration was whether there might have been a leakage of data from the testing data to the training during the pre-processing. On account of this the developed code had to be reviewed and the algorithms retrained with a different set of pre-processing steps. The removed pre-processing step that provided a helpful observation on the issue, was the class equalization. This step should not have leaked data from training to testing dataset, but was removed because it could, depending on the dataset, have made the number of entries for training too small and thus "easy" for the algorithms to predict.

As thought, the removal of the class equalization step did not solve the issue, but it did show where the problem lay. When training without class equalization, most iterations of the outer loop resulted in a single variable being used for prediction. When the variable meaning was looked up by referring to the research definition, the following question arose: "Do you have high blood sugar?". This is obviously directly connected to diabetes. This was a problem, because with that single variable the algorithms could predict perfectly whether or not a given person had diabetes. Since we already knew that high blood sugar was part of the diabetes disease, and it was interfering with the ability of the algorithms to correctly use other variables for risk factors discovery, it was removed from the training variables and all algorithms were retrained again.

When executing the training again without the high blood sugar variable,

the same problem arose again. By repeating the already mentioned steps, another variable arose as the only factor associated to diabetes. This variable pertained to whether a given person had been tested for diabetes in the last few months. For the same reason as the previous one, this variable was removed from the training dataset. Once those two variables were removed, everything was trained again and a more useful result surfaced.

5.6 Limitations

Two scaling limitations were found during the training process. They were not breaking limitations in the sense that they did not stop the progress of the whole research, but they limited the pool of used algorithms, not allowing the inclusion of the neural network algorithm. Depending on the circumstances, they could be a problem for other researches following the same methodology or framework, so it is important to mention them.

The methodology used in this research is very efficient in finding the best combination of variables and parameters, but it comes at a cost of computing power. The required computing power may be a limitation for a few applications of ML, which is why this methodology would probably not be recommended in the case of really large datasets or complex algorithms. In those cases, a slightly faster and more computationally efficient technique may be required.

The "sklearn" framework is also a limiting factor when the training is computationally expensive. It is indeed very easy to use and very friendly for those that do not have a very extensive knowledge in the ML field, but it does not scale very well across cores, and cannot be scaled at all across different machines.

5.7 Comparison

In possession of all the information from the training results, we can start comparing the algorithm values with one another and with the SLR results, evaluate their efficiency, and analyze the risk factors identified by the algorithms.

5.7.1 SLR Results Comparison

For a direct comparison of values with the SLR results, only the AUC score is going to be considered. The statistics from the SLR results are taken from the candle graph in Figure 3.8 on page 35.

- **Logistic Regression**. The LR algorithm scored 0.739 in this research, which is located in the upper bounds of the candle graph. This means that the algorithm developed during this research is highly competitive compared to the the current state-of-the-art.
- **Decision Tree**. The DT algorithm scored 0.622 during this work, which is below the lower bounds of the candle graph. This means that the algorithm is not effective.
- **Random Forest**. The RF algorithm scored 0.781, the highest score in this research, and above the SLR average. This means that the algorithm can be considered highly successful for this research and for the medicine field.
- **Support Vector Machine**. The SVM algorithm scored 0.709, which is located inside the bounds of the candle graph, although quite close to the bottom. This means that the algorithm is not very effective.

In summary, two algorithms received an excellent score when compared to the SLR. Both can be used by researchers in the medical field, and perhaps even doctors, for screening patients at risk.

5.7.2 The Optimal Algorithm

Selecting the optimal algorithm is more challenging than simply identifying which one has the highest score. In practice, a particular algorithm might be better suited to a specific scenario. Four scenarios that could change what is perceived as the optimal algorithm are presented below:

- For a doctor performing a priority, it might be more effective to use the algorithm that has a combination of the highest score with the least amount of variables; since a given doctor would personally assess the situation afterward, a number of mistakes for the sake of an improved time could be beneficial.

- When one is developing an online quiz for diagnosing possible diseases, it is definitely better to have a greater accuracy at the cost of a several questions more, especially since, in this case, there may be no doctors involved in the analysis.
- In a case where the variables have already been gathered, or it is hard to change which variables are recorded, the diagnosis should be given by means of the available variables. This situation can happen when trying to use the hospitals current database for this purpose, where the data is already there, and it is difficult to make a change in the system to gather different variables. One could then decide which algorithm to use based on the most commonly used variables of the algorithms because this would provide the highest accuracy in the results - even though in a more general case, the given algorithm might perform less effectively than others.
- Depending on how much computing power or time is available, a question about the complexity of the algorithms is also important to define the best one. RF is an algorithm that is very effective and has only two important parameters to configure. On the other hand, neural networks could be better than RF, but it has a multitude of parameters, which could lead to an expensive and time consuming process of training and evaluation.

In the first scenario, the optimal algorithm, based on the findings of this research, is the LR. It has a lower score than RF, but only uses 51.02 variables on average, compared to 98.02 from RF. Furthermore, in LR the importance of variables declines at a significantly faster pace then in RF; this happens because RF is an algorithm that by definition is better when dealing with an increased number of variables, opposite to how LR functions.

In the second scenario, RF is clearly the optimal algorithm: it possesses the highest accuracy, although this is at the cost of a significantly high variable requirement. In addition, it has a remarkably steady use of the available variables; this means that it can consider variables that may not play so important a role in the prediction of diabetes but can statistically help other, more correlated, variables with the diagnosis.

The third scenario, presents a significant challenge, and the problem depends upon what the input data is. If one examines the most important variables from each algorithm, there might be some overlap, but usually not a complete fit. This

means that in these cases, it is recommended to train the algorithms again against the new database. And with those results in mind, decide the optimum algorithm.

5.7.3 Risk Factors

A striking difference between the various algorithms involves the selection of the top variables. Each algorithm produced a different set of variables that were selected as the most important risk factors to be considered. The reasons for this phenomenon go beyond the scope of this research; generally speaking, however, it involves a difference in how each algorithm sees and interprets the input data, and how they induce the model through their learning process.

To be characterized as a risk factor, the variable must have a connection with the output. These connections are about finding a way to correlate directly a variable with the outcome, usually on a direct relation. With this study, a more complex form of correlation is available, since the ML algorithms are able to detect non-linear correlations.

By means of the graphs previously depicted (Figures 5.2 on page 46, 5.4 on page 49, 5.6 on page 51, and 5.8 on page 54), which showed the variables that were considered the most relevant according to each iteration, we can produce an analysis of the most important risk factors discovered. Firstly, one should note that not all variables considered important by the algorithms should be considered risk factors. There are other factors to consider: for example, the variable could actually be a **consequence** of developing diabetes, and, as such not technically a risk factor, although a correlation may exist. The algorithms may not be able to differentiate between the variables that are the cause of diabetes and those that are caused by it. In this research, those points are not discussed in depth. It is, however, important to keep in mind that the variables discussed and compared in the following analysis could be considered consequences and not risk factors, such that a further detailed analysis by specialists is necessary to confirm the validity of results.

Table 5.16 summarizes the most important variables and lists the different algorithms that referenced them as risk factors. By analyzing the number of algorithms that point each specific variable as a risk factor, it can be observed that the first three variables in the table are the ones with higher chance of having an impact. Following, these variables are described in more detail.

Table 5.16: Most important variables - summary

Variable	Meaning	Algorithms
percepan	Health perception in the last 12 months	LR, RF, SVM
tempres	Has high blood pressure	LR, RF
anos	Age	RF, SVM
pesmora	How many people live in his/her house	RF
relig1	Religion	RF
serie	Scholarity	RF
grau	Scholarity	RF
empreg	Working state	RF
ren1	Income	RF
ouvrad	Listen to the radio	RF
peso	Weight	RF
mamaex	Last mamma exam	RF
mamonde	Location of the last mamma exam	RF
verpa	Has ever measured blood pressure	RF
locons	Where was his/her last doctor visit	RF
convtemp	For how long has he/she had medical insurance	RF
baitemp	For how long does he/she visit doctors in the local clinic	RF
almam	Problems using the baby bottle while breastfeeding	RF
interne	Internet usage frequency	LR
angstres	Knowledge question: does stress lead to heart attack	LR
diabobe	Knowledge question: what makes obesity worse	LR
colevnao	Know how to avoid uterine cervix cancer	LR
temcol	Has high cholesterol	LR
temtrig	Has high triglycerides	LR
papex	Has done preventive exams for uterine Cancer	LR
mamofez	Has done mammography before	LR

The first variable, “Health perception in the last 12 months”, was indicated by three algorithms. This makes it the most relevant variable and risk factor for this research. The threefold indication of this variable suggests that a given person’s perception of his/her own health could be an indication that he/she has diabetes. It means, moreover, that developing diabetes makes the particular person feel considerably worse with regards to her health. This statement might appear obvious, but it should be emphasized that not all diseases cause that level of perception on an individual - sometimes a disease simply functions as something bothering a given person, but does not really appear terribly important. As indicated above, there is a the possibility that this results could also be an effect of a given person realizing he/she has diabetes and thus feeling negative because of it. Having this perception may also motivate an individual to look for medical assistance, which may aid in the diagnosis of diabetes.

The second variable, “High blood pressure”, was indicated by two algorithms, the LR and RF. The variable might appear it is an obvious - and one that does not need the help of a machine learning algorithm since most patients with diabetes,

especially those with type 2 diabetes, develop high blood pressure at some stage. Nonetheless, this result confirms that the algorithms were trained correctly, since it was pointed as a risk factor, and indeed, it may lead to several complications of diabetes.

The third variable, "Age", was indicated by two algorithms, the RF and SVM. Although middle-aged and older adults are still at the highest risk for developing type 2 diabetes, a recent work has discussed the greater negative effect of diabetes on mortality and morbidity for patients diagnosed at young age and their elevated risk for renal and nerve complications (AL-SAEED et al., 2016). This emphasizes the importance of age in the time of diagnosis and the need to develop strategies for management of high-risk patient groups to control mortality.

Besides these three variables, several others have been indicated by a single algorithm during every iteration of the training; this means that they are worth examining to a greater extent. However, since these variables are only used by a single algorithm, they might or might not be useful; thus we shall only cite a selected number that appeared interesting and worthy of mention in this research. Furthermore, various theories as to why they appeared are discussed: it is important to note, however, this reason is not the objective of this research, and, accordingly, should not be considered as definitive, only as guidelines for future research.

- "Internet usage frequency" is the first variable that stands out as an interesting associated variable. Usually, with a higher frequency of internet usage, the person tends to be more sedentary, and this might explain why internet usage may be associated to the development of diabetes. Certainly, it is not common knowledge that internet usage and diabetes are correlated.
- "The number of people living in the same house" is also correlated with the development of diabetes. Usually when more people live in the same house, it is due to the adults deciding to have kids. The children may be able to cause changes in the habits of the parents: for example, adults eating less healthy, or neglecting care related to their health due to the attention required by their children.
- The variable "Working state" also deserves some discussion. It is not clear if people that work have a higher chance of developing diabetes or it is the opposite that is true. Either way, it is an interesting variable for further research in the subject.

- A few “Knowledge variables” also appear important indicators in predicting diabetes, such as the “Scholarity” of a given person or how much they understand about obesity and heart attacks. This is an indication that the knowledge about the disease can actually help in its prevention (these risk factors were also the original object of research of the dataset in which this work was based).
- “Income” is also an important variable in this research. This is an interesting finding because there might be a correlation between an individual’s income and his/her knowledge and, most importantly, interest in his/her health.

6 CONCLUSION

Noncommunicable diseases are a major public health issue faced worldwide, specially in developing countries. To help decrease the global burden caused by chronic diseases such as diabetes, a more efficient system of diagnosis is necessary. The data that is saved on a daily basis from patients in hospitals and clinics can be used for that purpose when analyzed by modern solutions as those provided by ML. ML algorithms allow the analysis of more complex data than traditional statistical approaches, extracting non-linear relationships and representing the gathered knowledge as applicable predictive models.

In the context of ML applications in the medical field, the SLR performed in the scope of this work has shown that ML is indeed an increasingly popular analytical tool in the study of chronic diseases. Numerous articles have adopted a variety of ML algorithms, among which SVM, RF, and DT stand out, to train predictive models of diagnoses and to identify potential risk factors for chronic diseases. In general, these models were highly competitive in relation to LR, which is by far the most traditional approach for data analysis in epidemiology, as well as in contrast to clinical prediction scores derived from specialists.

By performing experimental research with SVM, RF, DT, and LR to analyze a diabetes-related database with the goal of training a predictive model and determining potential risk factors, we observed that the best overall performance was achieved by the RF algorithm. The scores given by the algorithms trained during this research are not the highest in the field, as can be seen by the SLR comparison. In spite of that, they could be used in multiple scenarios successfully. It is also important to mention that the notion of best algorithm depends greatly on the specific scenario for which the ML model would be applied, and should consider the tradeoff between number of variables, variables available, and predictive performance. We have also noted that the optimal parameters for the algorithms also suffer large variation among iterations of the training process, suggesting that parameters configuration is not a straightforward process and that an interaction among training set, variable selection, and model tuning is important.

Finally, the ML training results provided the means to investigate the most relevant variables, which may be interpreted as important variables associated with the occurrence of diabetes. We have highlighted a collection of potential risk factors

that present some plausibility regarding their association with diabetes. Some of these were identified by more than one algorithm, increasing their relevance. Further research involving specialists is necessary to evaluate their actual relevance and expand the knowledge about their impact on diabetes onset. Overall, our results corroborate previous studies on the suitability and power of ML approaches for helping in data analysis to improve medical diagnosis, and may motivate the development of further works within this interdisciplinary field.

REFERENCES

AL-SAEED, A. H. et al. An inverse relationship between age of type 2 diabetes onset and complication risk and mortality: the impact of youth-onset type 2 diabetes. **Diabetes care**, Am Diabetes Assoc, v. 39, n. 5, p. 823–829, 2016.

ALGHAMDI, M. et al. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. **PLoS ONE**, v. 12, n. 7, p. e0179805, 2017.

ANDERSON, A. E. et al. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. **J Biomed Inform**, v. 60, p. 162–168, Apr 2016.

ANDERSON, J. P. et al. Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. **J Diabetes Sci Technol**, v. 10, n. 1, p. 6–18, Dec 2015.

ASLAN, K. et al. Can neural network able to estimate the prognosis of epilepsy patients accorrding to risk factors? **J. Medical Systems**, v. 34, n. 4, p. 541–550, 2010.

AUSSEM, A.; MORAIS, S. R. de; CORBEX, M. Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks. **Artificial Intelligence in Medicine**, v. 54, n. 1, p. 53–62, 2012.

AYER, T. et al. Breast cancer risk estimation with artificial neural networks revisited. **Cancer**, Wiley Online Library, v. 116, n. 14, p. 3310–3321, 2010.

BEAM, A. L.; KOHANE, I. S. Big data and machine learning in health care. **Jama**, American Medical Association, v. 319, n. 13, p. 1317–1318, 2018.

BERIKOL, G. B.; YILDIZ, O.; OZCAN, I. T. Diagnosis of Acute Coronary Syndrome with a Support Vector Machine. **J Med Syst**, v. 40, n. 4, p. 84, Apr 2016.

CAFFO, B. et al. A novel approach to prediction of mild obstructive sleep disordered breathing in a population-based sample: the Sleep Heart Health Study. **Sleep**, v. 33, n. 12, p. 1641–1648, Dec 2010.

CHOI, S. B. et al. Screening for prediabetes using machine learning models. **Comput Math Methods Med**, v. 2014, p. 618976, 2014.

CHONG, C. et al. Stratification of adverse outcomes by preoperative risk factors in coronary artery bypass graft patients: An artificial neural network prediction model. In: **AMIA 2003, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 8-12, 2003**. [S.l.: s.n.], 2003.

COLAK, C.; KARAMAN, E.; TURTAY, M. G. Application of knowledge discovery process on the prediction of stroke. **Comput Methods Programs Biomed**, v. 119, n. 3, p. 181–185, May 2015.

CONFORTI, D.; GUIDO, R. Kernel-based support vector machine classifiers for early detection of myocardial infarction. **Optimization Methods and Software**, Taylor & Francis, v. 20, n. 2-3, p. 401–413, 2005.

DELEN, D.; WALKER, G.; KADAM, A. Predicting breast cancer survivability: a comparison of three data mining methods. **Artificial intelligence in Medicine**, Elsevier, v. 34, n. 2, p. 113–127, 2005.

DOMBI, G. W.; ROSBOLT, J. P.; SEVERSON, R. K. Neural network analysis of employment history as a risk factor for prostate cancer. **Comp. in Bio. and Med.**, v. 40, n. 9, p. 751–757, 2010.

EASTON, J. F.; STEPHENS, C. R.; ANGELOVA, M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. **Comput. Biol. Med.**, v. 54, p. 199–210, Nov 2014.

FARRAN, B. et al. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. **BMJ open**, British Medical Journal Publishing Group, v. 3, n. 5, p. e002457, 2013.

GÉRON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems**. [S.l.]: " O'Reilly Media, Inc.", 2017.

GOLDSTEIN, B. A. et al. Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records. **Clin J Am Soc Nephrol**, v. 9, n. 1, p. 82–91, Jan 2014.

GONÇALVES, C. V. et al. Women's knowledge of methods for secondary prevention of breast cancer. **Ciencia & Saude Coletiva**, SciELO Brasil, v. 22, n. 12, p. 4073–4082, 2017.

GREEN, M. et al. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. **Artificial intelligence in Medicine**, Elsevier, v. 38, n. 3, p. 305–318, 2006.

GUI, C.; CHAN, V. Machine learning in Medicine. **University of Western Ontario Medical Journal**, v. 86, n. 2, p. 76–78, Dec. 2017.

HABIBI, S.; AHMADI, M.; ALIZADEH, S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. **Glob J Health Sci**, v. 7, n. 5, p. 304–310, Mar 2015.

HUANG, Y. et al. Feature selection and classification model construction on type 2 diabetic patients' data. **Artificial intelligence in Medicine**, Elsevier, v. 41, n. 3, p. 251–262, 2007.

International Diabetes Federation. <http://www.diabetesatlas.org>. Accessed June 13, 2018.

ISLAM, F. et al. Potential risk factor analysis and risk prediction system for stroke using fuzzy logic. In: **Artificial Intelligence Trends in Intelligent Systems - Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 1**. [S.l.: s.n.], 2017. p. 262–272.

JAJROUDI, M. et al. Prediction of survival in thyroid cancer using data mining technique. **Technol. Cancer Res. Treat.**, v. 13, n. 4, p. 353–359, Aug 2014.

JO, C.; AHN, C.; EGGER, B. A machine learning-based approach to live migration modeling. **Proceedings of the 2017 Symposium on Cloud Computing**, p. 351–364, 2017.

KANERVA, N. et al. Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design. **Scandinavian journal of public health**, SAGE Publications Sage UK: London, England, p. 1403494817736944, 2017.

KARAOLIS, M. A. et al. Assessment of the risk factors of coronary heart events based on data mining with decision trees. **IEEE Trans Inf Technol Biomed**, v. 14, n. 3, p. 559–566, May 2010.

KINAR, Y. et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 23, n. 5, p. 879–890, 2016.

KONERMAN, M. A. et al. Assessing risk of fibrosis progression and liver-related clinical outcomes among patients with both early stage and advanced chronic hepatitis c. **PloS one**, Public Library of Science, v. 12, n. 11, p. e0187344, 2017.

KONERMAN, M. A. et al. Improvement of predictive models of risk of disease progression in chronic hepatitis c by incorporating longitudinal data. **Hepatology**, Wiley Online Library, v. 61, n. 6, p. 1832–1841, 2015.

KONONENKO, I. Machine learning for medical diagnosis: history, state of the art and perspective. **Artificial Intelligence in Medicine**, Elsevier, v. 23, n. 1, p. 89–109, 2001.

KRSTAJIC, D. et al. Cross-validation pitfalls when selecting and assessing regression and classification models. **Journal of cheminformatics**, Nature Publishing Group, v. 6, n. 1, p. 10, 2014.

KUROSAKI, M. et al. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C. **J. Hepatol.**, v. 56, n. 3, p. 602–608, Mar 2012.

KUSIAK, A.; DIXON, B.; SHAH, S. Predicting survival time for kidney dialysis patients: a data mining approach. **Computers in biology and medicine**, Elsevier, v. 35, n. 4, p. 311–327, 2005.

LEE, B. J.; KIM, J. Y. Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning. **IEEE J Biomed Health Inform**, v. 20, n. 1, p. 39–46, Jan 2016.

LEE, C. et al. Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only. **J. Korean Med. Sci.**, v. 30, n. 8, p. 1025–1034, Aug 2015.

LEE, C. H.; CHEN, J. C.; TSENG, V. S. A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. **Comput Methods Programs Biomed**, v. 101, n. 1, p. 44–61, Jan 2011.

LI, H. et al. Prediction and Informative Risk Factor Selection of Bone Diseases. **IEEE/ACM Trans Comput Biol Bioinform**, v. 12, n. 1, p. 79–91, 2015.

LI, X. et al. Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. **AMIA Annual Symposium Proceedings**. [S.l.], 2016. v. 2016, p. 799.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. **R news**, v. 2, n. 3, p. 18–22, 2002.

LUANGRUANGRONG, W.; RODTOOK, A.; CHIMMANEE, S. Study of type 2 diabetes risk factors using neural network for thai people and tuning neural network parameters. In: **Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC 2012, Seoul, Korea (South), October 14-17, 2012**. [S.l.: s.n.], 2012. p. 991–996.

LUNDIN, M. et al. Artificial neural networks applied to survival prediction in breast cancer. **Oncology**, Karger Publishers, v. 57, n. 4, p. 281–286, 1999.

MADZAROV, G.; GJORGJEVIKJ, D.; CHORBEV, I. A multi-class svm classifier utilizing binary decision tree. **Informatica**, v. 33, n. 2, 2009.

MANTZARIS, D. H. et al. A soft computing approach for osteoporosis risk factor estimation. In: **Artificial Intelligence Applications and Innovations - 6th IFIP WG 12.5 International Conference, AIAI 2010, Larnaca, Cyprus, October 6-7, 2010. Proceedings**. [S.l.: s.n.], 2010. p. 120–127.

MARIANO, D. C. et al. A guide to performing systematic literature reviews in bioinformatics. **arXiv preprint arXiv:1707.05813**, 2017.

MENDIS, S. **Global status report on noncommunicable diseases 2014**. [S.l.]: World health organization, 2014.

MENG, X.-H. et al. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. **The Kaohsiung Journal of Medical Sciences**, v. 29, n. 2, p. 93 – 99, 2013. ISSN 1607-551X.

MENTI, E. et al. Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients. **AMIA Annu Symp Proc**, v. 2016, p. 884–893, 2016.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning: An artificial intelligence approach**. [S.l.]: Springer Science & Business Media, 2013.

NCD Risk Factor Collaboration et al. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with $4 \cdot 4$ million participants. **The Lancet**, Elsevier, v. 387, n. 10027, p. 1513–1530, 2016.

NG, K. et al. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. **Circulation: Cardiovascular Quality and Outcomes**, Am Heart Assoc, v. 9, n. 6, p. 649–658, 2016.

OLIVERA, A. R. et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. **Sao Paulo Med J**, v. 135, n. 3, p. 234–246, 2017.

PANAHIAZAR, M. et al. Using EHRs and Machine Learning for Heart Failure Survival Analysis. **Stud Health Technol Inform**, v. 216, p. 40–44, 2015.

PEREIRA, N. R. P. et al. Development of a Prognostic Survival Algorithm for Patients with Metastatic Spine Disease. **J Bone Joint Surg Am**, v. 98, n. 21, p. 1767–1776, Nov 2016.

RAMEZANKHANI, A. et al. Classification-based data mining for identification of risk patterns associated with hypertension in Middle Eastern population: A 12-year longitudinal study. **Medicine (Baltimore)**, v. 95, n. 35, p. e4143, Aug 2016.

RAMEZANKHANI, A. et al. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. **Diabetes Res. Clin. Pract.**, v. 105, n. 3, p. 391–398, Sep 2014.

RAO, V. S. H.; KUMAR, M. N. Novel approaches for predicting risk factors of atherosclerosis. **CoRR**, abs/1501.07093, 2015.

SCHULDT, C.; LAPTEV, I.; CAPUTO, B. Recognizing human actions: a local svm approach. In: IEEE. **Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on**. [S.l.], 2004. v. 3, p. 32–36.

SENGUPTA, P. P. et al. Cognitive Machine-Learning Algorithm for Cardiac Imaging: A Pilot Study for Differentiating Constrictive Pericarditis From Restrictive Cardiomyopathy. **Circ Cardiovasc Imaging**, v. 9, n. 6, Jun 2016.

SHANKER, M. S. Using neural networks to predict the onset of diabetes mellitus. **Journal of chemical information and computer sciences**, ACS Publications, v. 36, n. 1, p. 35–41, 1996.

SHOUVAL, R. et al. Prediction of Allogeneic Hematopoietic Stem-Cell Transplantation Mortality 100 Days After Transplantation Using a Machine Learning Algorithm: A European Group for Blood and Marrow Transplantation Acute Leukemia Working Party Retrospective Data Mining Study. **J. Clin. Oncol.**, v. 33, n. 28, p. 3144–3151, Oct 2015.

SLADOJEVI?, M. et al. DATA MINING APPROACH FOR IN-HOSPITAL TREATMENT OUTCOME IN PATIENTS WITH ACUTE CORONARY SYNDROME. **Med. Pregl.**, v. 68, n. 5-6, p. 157–161, 2015.

SONG, T. et al. Usefulness of the heart-rate variability complex for predicting cardiac mortality after acute myocardial infarction. **BMC Cardiovascular Disorders**, v. 14, n. 1, p. 59, May 2014. ISSN 1471-2261.

SUNG, S. F. et al. Developing a stroke severity index based on administrative data was feasible using data mining techniques. **J Clin Epidemiol**, v. 68, n. 11, p. 1292–1300, Nov 2015.

TAATI, B. et al. Data mining in bone marrow transplant records to identify patients with high odds of survival. **IEEE journal of biomedical and health informatics**, IEEE, v. 18, n. 1, p. 21–27, 2014.

TEMURTAS, H.; YUMUSAK, N.; TEMURTAS, F. A comparative study on diabetes disease diagnosis using neural networks. **Expert Systems with applications**, Elsevier, v. 36, n. 4, p. 8610–8615, 2009.

TSENG, C.-J. et al. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. **Artificial intelligence in Medicine**, Elsevier, v. 78, p. 47–54, 2017.

TSIEN, C. L. et al. Using classification tree and logistic regression methods to diagnose myocardial infarction. **Medinfo**, v. 98, 1998.

VOLZKE, H. et al. A new, accurate predictive model for incident hypertension. **J. Hypertens.**, v. 31, n. 11, p. 2142–2150, Nov 2013.

WANG, C. et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach. **Diabetes research and clinical practice**, Elsevier, v. 100, n. 1, p. 111–118, 2013.

WANG, W.; RICHARDS, G.; REA, S. Hybrid data mining ensemble for predicting osteoporosis risk. In: IEEE. **Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the**. [S.l.], 2006. p. 886–889.

WEI-JIA, L.; LIANG, M.; HAO, C. Particle swarm optimisation-support vector machine optimised by association rules for detecting factors inducing heart diseases. **J. Intelligent Systems**, v. 26, n. 3, p. 573, 2017.

WENG, S. F. et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? **PLoS ONE**, v. 12, n. 4, p. e0174944, 2017.

WORACHARTCHEEWAN, A. et al. Identification of metabolic syndrome using decision tree analysis. **Diabetes Res. Clin. Pract.**, v. 90, n. 1, p. e15–18, Oct 2010.

XIE, J. et al. A novel hybrid subset-learning method for predicting risk factors of atherosclerosis. In: **2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, Kansas City, MO, USA, November 13-16, 2017**. [S.l.: s.n.], 2017. p. 2124–2131.

YARGHOLI, E.; PARVANEH, S. Novel cardiac risk factor stratification using neuro-fuzzy tool. In: **2008 International Conferences on Computational Intelligence for Modelling, Control and Automation (CIMCA 2008), Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2008), Innovation in Software Engineering (ISE 2008), 10-12 December 2008, Vienna, Austria**. [S.l.: s.n.], 2008. p. 1199–1204.

YOO, T. K. et al. Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. **Yonsei medical journal**, v. 54, n. 6, p. 1321–1330, 2013.

YU, W. et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. **BMC medical informatics and decision making**, BioMed Central, v. 10, n. 1, p. 16, 2010.

ZAROGIANNI, E. et al. Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. **Schizophrenia research**, Elsevier, v. 181, p. 6–12, 2017.

ZHANG, H. et al. Risk factors of heart failure for patients classification with extreme learning machine. In: **International Conference on Machine Learning and Cybernetics, ICMLC 2016, Jeju Island, South Korea, July 10-13, 2016**. [S.l.: s.n.], 2016. p. 814–819.

ZHAO, Y. et al. Exploration of machine learning techniques in predicting multiple sclerosis disease course. **PloS one**, Public Library of Science, v. 12, n. 4, p. e0174866, 2017.

ANNEX A — SLR RESULTING ARTICLES

1. (ALGHAMDI et al., 2017) - Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project.
2. (ANDERSON et al., 2015) - Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records.
3. (ANDERSON et al., 2016) - Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study.
4. (BERIKOL; YILDIZ; OZCAN, 2016) - Diagnosis of Acute Coronary Syndrome with a Support Vector Machine.
5. (CAFFO et al., 2010) - A novel approach to prediction of mild obstructive sleep disordered breathing in a population-based sample: the Sleep Heart Health Study.
6. (CHOI et al., 2014) - Screening for prediabetes using machine learning models.
7. (COLAK; KARAMAN; TURTAY, 2015) - Application of knowledge discovery process on the prediction of stroke.
8. (EASTON; STEPHENS; ANGELOVA, 2014) - Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach.
9. (GOLDSTEIN et al., 2014) - Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records.
10. (HABIBI; AHMADI; ALIZADEH, 2015) - Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining.
11. (JAJROUDI et al., 2014) - Prediction of survival in thyroid cancer using data mining technique.
12. (KARAOLIS et al., 2010) - Assessment of the risk factors of coronary heart events based on data mining with decision trees.
13. (KUROSAKI et al., 2012) - Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C.

14. (LEE; CHEN; TSENG, 2011) - A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring.
15. (LEE et al., 2015) - Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only.
16. (LEE; KIM, 2016) - Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning.
17. (LI et al., 2015) - Prediction and Informative Risk Factor Selection of Bone Diseases.
18. (MENG et al., 2013) - Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.
19. (MENTI et al., 2016) - Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients.
20. (OLIVERA et al., 2017) - Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study.
21. (PANAHIAZAR et al., 2015) - Using EHRs and Machine Learning for Heart Failure Survival Analysis.
22. (PEREIRA et al., 2016) - Development of a Prognostic Survival Algorithm for Patients with Metastatic Spine Disease.
23. (RAMEZANKHANI et al., 2014) - Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study.
24. (RAMEZANKHANI et al., 2016) - Classification-based data mining for identification of risk patterns associated with hypertension in Middle Eastern population: A 12-year longitudinal study.
25. (SENGUPTA et al., 2016) - Cognitive Machine-Learning Algorithm for Cardiac Imaging: A Pilot Study for Differentiating Constrictive Pericarditis From Restrictive Cardiomyopathy.
26. (SHOUVAL et al., 2015) - Prediction of Allogeneic Hematopoietic Stem-Cell Transplantation Mortality 100 Days After Transplantation Using a Machine Learning Algorithm: A European Group for Blood and Marrow Transplantation Acute Leukemia Working Party Retrospective Data Mining Study.

27. (SLADOJEVI? et al., 2015) - DATA MINING APPROACH FOR IN-HOSPITAL TREATMENT OUTCOME IN PATIENTS WITH ACUTE CORONARY SYNDROME.
28. (SONG et al., 2014) - Usefulness of the heart-rate variability complex for predicting cardiac mortality after acute myocardial infarction.
29. (SUNG et al., 2015) - Developing a stroke severity index based on administrative data was feasible using data mining techniques.
30. (VOLZKE et al., 2013) - A new, accurate predictive model for incident hypertension.
31. (WENG et al., 2017) - Can machine-learning improve cardiovascular risk prediction using routine clinical data?
32. (WORACHARTCHEEWAN et al., 2010) - Identification of metabolic syndrome using decision tree analysis.
33. (ASLAN et al., 2010) - Can Neural Network Able to Estimate the Prognosis of Epilepsy Patients Accorrding to Risk Factors?
34. (AUSSEM; MORAIS; CORBEX, 2012) - Stratification of Adverse Outcomes by Preoperative Risk Factors in Coronary Artery Bypass Graft Patients: An Artificial Neural Network Prediction Model
35. (CHONG et al., 2003) - Neural network analysis of employment history as a risk factor for prostate cancer
36. (DOMBI; ROSBOLT; SEVERSON, 2010) - Potential Risk Factor Analysis and Risk Prediction System for Stroke Using Fuzzy Logic
37. (ISLAM et al., 2017) - Study of Type 2 diabetes risk factors using neural network for Thai people and tuning neural network parameters
38. (LUANGRUANGRONG; RODTOOK; CHIMMANEE, 2012) - A Soft Computing Approach for Osteoporosis Risk Factor Estimation
39. (MANTZARIS et al., 2010) - Novel Approaches for Predicting Risk Factors of Atherosclerosis
40. (RAO; KUMAR, 2015) - Particle Swarm Optimisation-Support Vector Machine Optimised by Association Rules for Detecting Factors Inducing Heart Diseases
41. (WEI-JIA; LIANG; HAO, 2017) - A novel hybrid subset-learning method for predicting risk factors of atherosclerosis

42. (XIE et al., 2017) - Novel Cardiac Risk Factor Stratification Using Neuro-fuzzy Tool
43. (YARGHOLI; PARVANEH, 2008) - Risk factors of heart failure for patients classification with extreme learning machine
44. (ZHANG et al., 2016) - Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study
45. (KINAR et al., 2016) - Improvement of Predictive Models of Risk of Disease-Progression in Chronic Hepatitis C by IncorporatingLongitudinal Data
46. (KONERMAN et al., 2015) - Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room
47. (GREEN et al., 2006) - Kernel-based Support Vector Machine classifiers for early detection of myocardial infarction
48. (CONFORTI; GUIDO, 2005) - Osteoporosis Risk Prediction for Bone Mineral Density Assessment of Postmenopausal Women Using Machine Learning
49. (YOO et al., 2013) - Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes
50. (YU et al., 2010) - Feature selection and classification model construction on type 2 diabetic patients' data
51. (HUANG et al., 2007) - A comparative study on diabetes disease diagnosis using neural networks
52. (TEMURTAS; YUMUSAK; TEMURTAS, 2009) - Predicting breast cancer survivability: a comparison of three data mining methods
53. (DELEN; WALKER; KADAM, 2005) - Artificial Neural Networks Applied to Survival Prediction in Breast Cancer
54. (LUNDIN et al., 1999) - Predicting survival time for kidney dialysis patients: a data mining approach
55. (KUSIAK; DIXON; SHAH, 2005) - "Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach "
56. (WANG et al., 2013) - Breast Cancer Risk Estimation With Artificial Neural Networks Revisited
57. (AYER et al., 2010) - Hybrid Data Mining Ensemble for Predicting Osteo-

porosis Risk

58. (WANG; RICHARDS; REA, 2006) - Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study
59. (FARRAN et al., 2013) - Using Neural Networks to Predict the Onset of Diabetes Mellitus
60. (SHANKER, 1996) - Using Classification Tree and Logistic Regression Methods to Diagnose Myocardial Infarction
61. (TSIEN et al., 1998) - Exploration of machine learning techniques in predicting multiple sclerosis disease course
62. (ZHAO et al., 2017) - Computational classifiers for predicting the short-term course of Multiple sclerosis
63. (TAATI et al., 2014) - Data mining in bone marrow transplant records to identify patients with high odds of survival.
64. (KANERVA et al., 2017) - Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design.
65. (KONERMAN et al., 2017) - Assessing risk of fibrosis progression and liver-related clinical outcomes among patients with both early stage and advanced chronic hepatitis C.
66. (NG et al., 2016) - Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time Before Diagnosis, Data Diversity, Data Quantity, and Data Density.
67. (TSENG et al., 2017) - Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence.
68. (ZAROGIANNI et al., 2017) - Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features.
69. (LI et al., 2016) - Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation.
70. (AUSSEM; MORAIS; CORBEX, 2012) - Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks.

ANNEX B — BEST VARIABLES RANKING ACCORDING TO THE INNER LOOP

These are the top 30 variables from each algorithm in the inner loop of the training, which corresponds to the variable selection.

Figure B.1: LR - inner loop best variables ranking

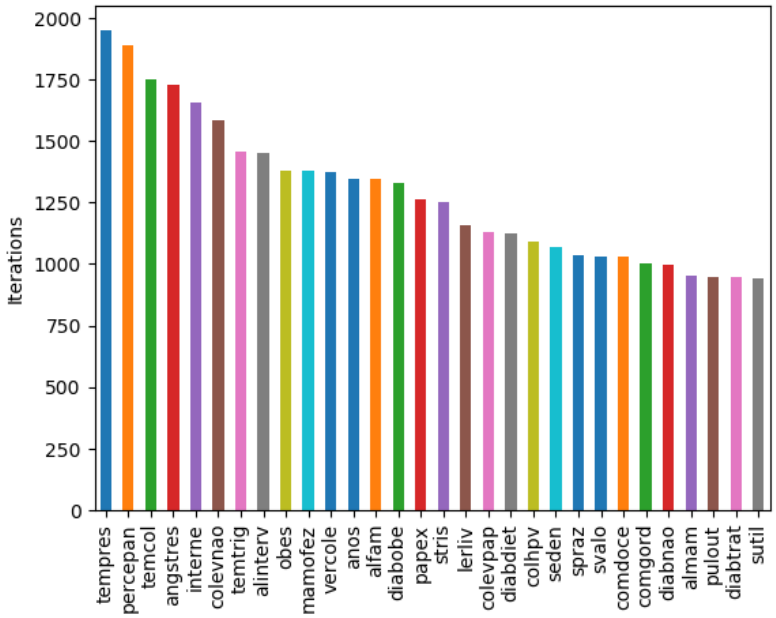


Figure B.2: DT - inner loop best variables ranking

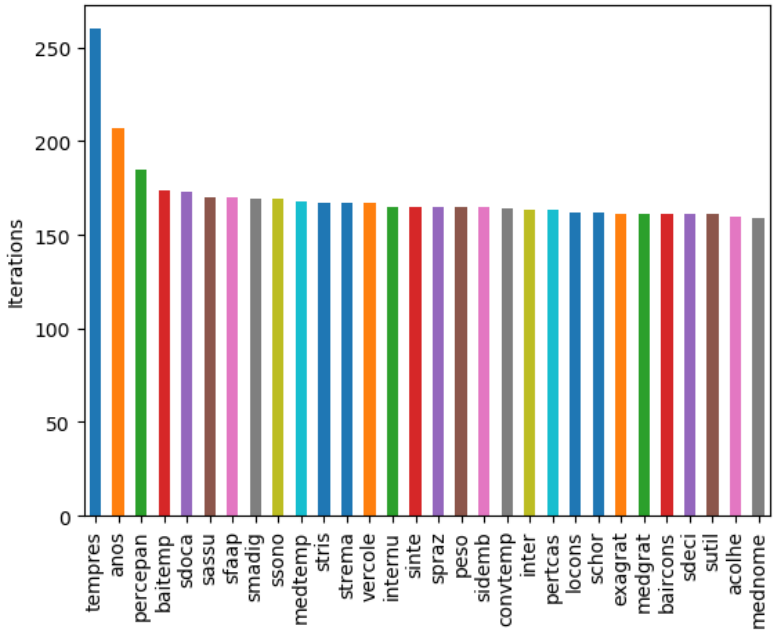


Figure B.3: RF - inner loop best variables ranking

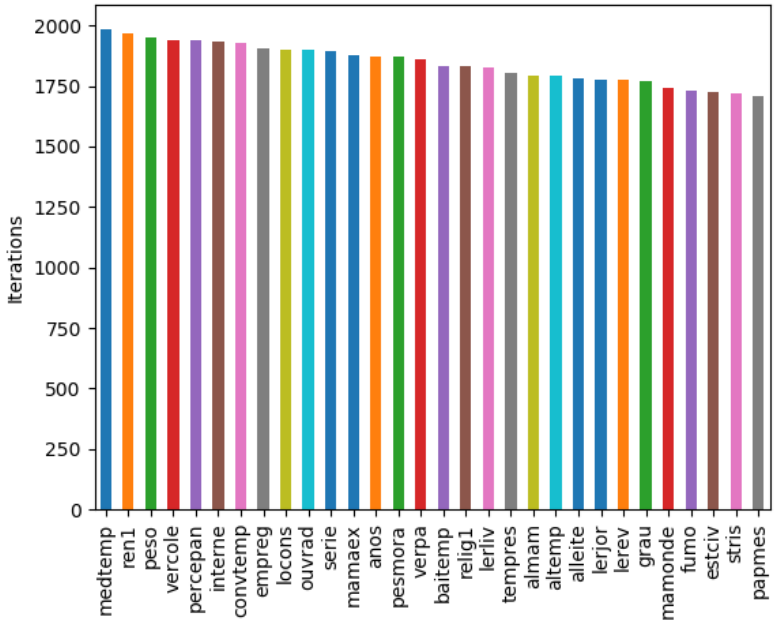


Figure B.4: SVM - inner loop vest variables ranking