 doi:10.20944/preprints202012.0318.v1

Article

Minimum Relevant features to Obtain Explainable Systems for Predicting Cardiovascular Disease Using the Statlog Dataset

Roberto Porto [1], **José M. Molina** [2], **Antonio Berlanga**[2] **and Miguel A. Patricio** [2]

[1] Corporación Universitaria Americana, Colombia; rporto@coruniamericana.edu.co
[2] Applied Artificial Intelligence Group, Universidad Carlos III de Madrid; molina@ia.uc3m.es, aberlan@ia.uc3m.es, mpatrici@inf.uc3m.es

Abstract: Learning systems have been very focused on creating models that are capable of obtaining the best results in error metrics. Recently, the focus has shifted to improvement in order to interpret and explain their results. The need for interpretation is greater when these models are used to support decision making. In some areas this becomes an indispensable requirement, such as in medicine. This paper focuses on the prediction of cardiovascular disease by analyzing the well-known Statlog (Heart) Data Set from the UCI's Automated Learning Repository. This study will analyze the cost of making predictions easier to interpret by reducing the number of features that explain the classification of health status versus the cost in accuracy. It will be analyzed on a large set of classification techniques and performance metrics. Demonstrating that it is possible to make explainable and reliable models that have a good commitment to predictive performance.

Keywords: Interpretable Artificial Intelligence; Cardiovascular disease prediction; Machine Learning in Healthcare

1. Introduction

In recent years there has been an increasing application of Machine Learning (ML) techniques in problem solving in a wide range of fields. One of them is medicine, where it is proved that it can increase in the help to medical diagnostics. The field of medicine incorporates a unique feature as it requires user confidence in the predictions and classifications of ML techniques. While in other fields of application of ML techniques it is sufficient to assess certain metrics such as accuracy or AUC, in recent years greater emphasis is being given to explanations of AI techniques [1]. In the application of these techniques in medicine it is necessary to know the reason behind the prediction or classification of a problem. In the present work we have focused on a specific problem in medicine as it is the prediction of Cardiovascular diseases (CVD). CVD are a group of conditions in the heart and blood vessels of the human body, standing out for their frequency in patients the following: arterial hypertension, coronary heart disease, peripheral vascular and vascular brain diseases, heart failure, rheumatic heart disease and congenital, and cardiomyopathies. Cardiovascular diseases top the list of causes of death worldwide, in fact, in 2015 alone a total of 17.7 million people was counted, representing 31% of deaths worldwide [2]. Symptoms and behaviors in people are the first signal to identify diseases and prevent them. 90% of the heart attacks that occur are associated with known classic risk factors, such as hypertension, high cholesterol levels, smoking, diabetes or obesity. Also, in this reality most of these factors are modifiable and therefore preventable. Predictions of various cardiovascular diseases vary depending on the population [3]. Many investigations that work on Data Science (e.g, [4–7]) approximate the work done in this study, since they use the relatively small set of Cardiovascular disease data composed of 14 features and 303 patient records and seek to recognize

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

patterns or diagnoses based on clinical test results established for each population (ie, data), to explain symptoms presented in patients. It should be noted that it is possible to diagnose risk factors based on the analysis of data from patients who may or may not suffer from diseases (eg, liver, diabetes mellitus, heart attacks, cancer, dengue among other infectious diseases) that can be treated even cured, from early detection through machine learning techniques [8–11]. These data are consolidated as a whole, which are used to create models that allow classifying or predicting diseases.

Another approach would be the research carried out in [12], in which they propose a model to predict heart disease from a set of private data, reducing the amount of features from 14 to 6, by using a genetic algorithm that allows the selection of categorical features, to subsequently use traditional classifiers for the prediction and diagnosis of heart disease, obtaining a classification percentage of 99.2% using the Decision Tree technique, out of 96.5% classification using the Naive Bayes technique. On the other hand, in [13] a classification model of heart disease is presented using the Statlog data set. This is composed of 2 systems, the first system uses the RelieF algorithm to extract the superior characteristics, discarding the features that offer less information, to later use the RS reduction heuristic for feature elimination, then the new data set is trained with the reduction of features, to perform tests with the different traditional classifiers, obtaining better results using the C4.5 technique with a classification percentage of 92.59% with the 7 features that presented the strongest qualities in distinction and of greater value.

An effort has been made to comply with the regulatory framework proposed by the commission of the European parliament [14], as a strategy for the future, which is based on building confidence in artificial intelligence and how it focuses on the human being, so the Reliability of these systems will depend on 3 main components, which are: it must be in accordance with the law, it must respect the basic principles and it must be solid. In the same vein and complying with the requirements derived from the components that an AI must have, set out above (ie, Human intervention and supervision, Technical soundness and security, Privacy and data management, Transparency, Diversity, non-discrimination and equity, Social and environmental Welfare, Accountability), the proposed methodology pursues the proposal that occurs in the world, specifically in Europe, where they are concerned that the results of AI are not reliable. This proposal focuses on the strength, safety of techniques and transparency, which relate that their decisions must be correct and reflect the correct reproducibility of the results from the documentation of the decisions taken, in addition to allowing the interpretation of a system with few related features, as an explanatory mechanism in decision making.

In the field of Healthcare, the problem of interpretability plays a very important role, since it is not possible to make a decision if it cannot be directly described in understandable terms. On the one hand, the doctor could not trust a decision that he cannot explain, and the patient will not be able to trust an expert who bases his decisions on the result of a computational method. For this reason, some lines of research have sought simpler and more interpretable models such as semantic representations [15,16] or attention mechanisms [17–19]. However, it has been proved that medical reasoning is compatible with rule-based representations [20] and, in this sense, one of the most interpretable models is the decision tree [21,22].

This research arises from an interest in investigating the multi-objective nature of the problem of improving the accuracy of classification versus the "interpretability" of a model obtained. In this case, the work is focused on the estimation of cardiovascular diseases. The result of the research work will assist cardiologists who help to enrich the quality of patients who may or may not suffer from cardiovascular diseases by examining and verifying the results obtained in an agile and efficient manner, generating knowledge and expectations in doctors about how to increase the precision of diagnoses from prediction, in order to control or prevent the risk factors of CVD from an interpretable system. For this purpose, data from the free repository of the University of California (UCI) will be used [23], more specifically from the Statlog heart disease data set, since it is the most used and has the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

most articles about the application of Machine Learning (ML) techniques to be able to classify any type of cardiac anomaly from the analysis of this data set.

The interpretation of the results after applying ML techniques to a data set with many features is often complicated, although the techniques used theoretically are interpretable, as is the case with decision trees, in particular the use of so many features and so much division of the data makes the interpretation of the results difficult, so a methodology is proposed that allows reducing the characteristics (i.e. the variables) to find information that can be learned. On the other hand, the relationships that can be understood or interpreted based on the features, allow the prediction of the proposed methodology to be improved or at least allow results similar to be obtained, in terms of classification tests, precision, sensitivity, specificity, with a smaller number of features of the proposed methodology. In addition, this study is oriented so that its application can be used directly on other medical data sets, so that the knowledge extracted from the data set of the applicability of a new proposal, which can be used in any data set regardless of its area or that the procedure can be applied to other medical data sets that allow a better approach to disease prediction and at some point, give another focus to traditional ML techniques to build information on the structure of these types of medical data sets.

2. Related Works

Currently, in the field of ML, predictions about various types of diseases are addressed using different methods. One of the most widely used methods is that based on decision trees. In [24] the performance obtained in the precision, sensitivity, specificity and accuracy of the tests is compared to the free Heart Disease dataset of the University of California composed of 13 features. For this study the data mining algorithms C5.0, II.2. Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Neural Networks (NN) were used in order to build a model that can predict heart disease. C5.0 was the algorithm that manages to build a model with the highest accuracy 93.02% while KNN, SVM, NN have been 88.37%, 86.05% and 80.23% respectively. In [25] authors apply an integration of the results of the machine learning analysis applied in different data sets aimed at CAD disease. This will avoid missing, incorrect and inconsistent data problems that may appear in the data collection. The fast decision tree and the pruned C4.5 tree is applied where the resulting trees are extracted from different data sets and compared. The common characteristics among these data sets are extracted and used in the subsequent analysis of the same disease in any data set. The results show that the classification accuracy of the collected data set is 78.06% higher than the average classification accuracy of all separate data sets, which is 75.48%. In [26] authors propose a model to predict coronary artery disease, applying Dimensionality Reduction (MDR) and recursive partition (RP). The EMLA model was tested on a set of 648 records consisting of 364 cases of CAD and 284 healthy records, showing outstanding performance compared to other models for cataloging the disease with a (89.3%) and stenosis prediction (82.5%). In [27] the comparison of prediction and classification techniques such as Naïve Bayes, Decision Trees J48 and the bagging algorithm is evidenced from the evaluation of the precision with which the model is constructed. Subsequently, a cross-validation of 10 iteration data must be performed in order to estimate the unbiased accuracy of the model. The model with the highest classification accuracy was Bagging with 85% compared to Naïve Bayes with 83% and J48 with 84%.

Another of the most widely used methods in this area are those based on Support Vector Machine (SVM). In [28] the prediction of liver diseases using classification algorithms was raised. The algorithms used in this work are Naïve Bayes and SVM. These classification algorithms are compared according to performance factors, that is, the accuracy of the classification and the execution time. From the experimental results it is observed that the SVM is high in terms of accuracy with 76.6% compared with 59% of Naïve Bayes, so it shows that it is superior for predicting liver diseases. Zhao et al in [29] use SVM model to construct the classifier and compare with logistic regression (LR) using demographic, clinical and magnetic resonance tomography (MRI) data obtained in years one and two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

to predict multiple sclerosis disease at five years of follow-up, that adhering clinical and brain data in the short term, corrective measures of class imbalance and classification costs to the SVM, it can be a promising mean to predict the course of MS disease and for the selection of suitable patients for more aggressive treatment regimens. The use of non-uniform misclassification costs in the SVM model increases sensitivity, and predictions up to 86%.

Related to models based on neural networks, we can highlight the work of [30]. Authors propose hybrid method between Artificial Neural Network (ANN) and Fuzzy_AHP for the prediction of the risk of heart failure, obtaining results in prediction tests, cross entropy and outstanding ROC graphs compared to the conventional ANN. In [31] authors propose an architecture based on Long short-term memory (LSTM) architecture which is effective and robust for predicting heart failure. Its main contribution of this document is to predict heart failure through a neural network from electronic medical data of patients divided into 2 data sets: A with 5000 records diagnosed with heart disease and B with 15,000 undiagnosed records extracted for 4 years, using the basic principles of a long-term memory network model. This architecture was compared with popular methods such as Random Forest (RF) and AdaBoost, showing superior performance in predicting the diagnosis of heart failure.

Other works make use of fuzzy models for disease prediction. In [32] authors propose a system of diagnosis of heart diseases that uses a reduction of features based on approximate sets and a system of interval fuzzy logic type 2 (IT2FLS). IT2FLS uses a hybrid learning process that includes a fuzzy c-mean clustering algorithm and parameter tuning by firefly chaos and genetic hybrid algorithms. Attribute reduction based on approximate sets using the chaos firefly algorithm is investigated to find an optimal reduction that, therefore, reduces computational load and increases IT2FLS performance. The results of the experiment demonstrate a significant predominance of the proposed system with 4 features of 86% accuracy compared to other machine learning methods, such as Naïve Bayes 83.3%, SVM 75.9% and ANN 77.8%. It also obtained a superior result in sensitivity tests with 87.1%, in specificity with 90%.

As it has been verified with the investigations previously described, there are several works that have been looking for the ML methods that obtain greater values of accuracy in their predictions. However, in the field of Healthcare, it may be more interesting to obtain models that allow an explanation related to the reasoning that has led to a given classification, although the accuracy values may be slightly decreased. These methods are known as Explainable AI (XAI) methods [33].

In recent years, we can find works in healthcare and interpretable ML [34–36]. Obtaining explainable ML models are certainly a benefit to stakeholders interested in Healthcare. These models allow, on the one hand, to increase transparency by indicating the reasoning that has been used to reach a specific decision and, on the other hand, to know the relevant factors that affect the prediction of the results. Therefore, the aim of this work is to present a methodology that allows us to reach these models of explainable ML algorithms.

3. Interpretability analysis in the prediction of Cardiovascular diseases

The aim of this section is to find an interpretable system for the early detection of cardiovascular diseases, seeking the maximization of the percentage of successes with the minimum number of features for prediction.

The work follows a methodology based on two different concepts: attribute relevance and explain ability of the final system. These two concepts are far from a unique definition, and there is no way to define a mathematical formulation to assign a value to represent the problem. The general idea of this work is not to define these concepts but try to obtain some valuable information to take the final decision: how many attributes are needed to obtain a useful prediction system that could be explainable. Explainability is relevant in medical-biological problems, and from this problem a depth studied problem is the analysis of Statlog (Heart) Data Set from the UCI's Automated Learning Repository. In section 2, a review of the algorithms used for the prediction of different chronic noncommunicable diseases was carried out, taking into account that the percentage of accuracy of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

techniques and methodologies found in the literature serve as an assessment of the success rates of the techniques used allowing to select the best, considering the interpretability of the results.

In this sense, methodology defines the steps and the decisions that should be taken, but not how to evaluate mathematically these two concepts. The proposed method follows several procedures to analyse the importance of the attributes, the similarity results of the learned systems, and considers that trees are more explainable than other methods. No numerical evaluation is defined for "similarity results" or "explainable" but methodology considers that the developer analyses results and takes decisions using known metrics for classification and analysing the final trees.

First step is the estimation of how the attributes are related to the output. Several importance metrics are considered: Chi2, GainR, OneR, SU, IG, ReI, KT, PCF, Rimp, Rper, RFi, AUC and anova. In this paper, the average of these values is considered as the final importance value, and the attributes are ranking using this average.

Second step, based on the previous ranking of attributes, is a comparison of several classifier performance considering all attributes, 2/3 of attributes and 1/3 of attributes. For each classifier and each combination of attributes several performance metrics are evaluated: Accuracy, Precision,Specificity, Sensitivity, MCC, Kappa and AUC.

Third step, when results of previous experiments are similar, is selection of techniques that generate a result easy to understand. In this proposal trees are considered as the most easy way to understand the final solution of the machine learning technique.

Finally, trees obtained using RPART (Recursive Partitioning And Regression Trees) [37] and Ctree (Conditional inference trees) [38] algorithms, and different configurations of attributes are analysed to developer decides the more explainable tree that are able to obtain a performance similar to other methods.

As we will see in successive sections, if a measure of "explainable" can be defined, the final results will be a pareto front with two objectives: "attribute importance value" and "explainable index". In this case, we use the average of several performance measures as the "attribute importance value" and the tree depth as the "explainable index".

3.1. Heart Disease Dataset (Statlog)

This section presents a dataset used for the experiments, which belongs to the open repository of the University of California which has the name Heart Disease and is composed of 76 features, 270 instances and 2 classes (present and absent). It should be noted that this is one of the most used open data sets in machine learning publications applied to the medical field [5]. For this investigation we only take 14 features including the class attribute, whose distribution is as follows: the class absent is made up of 150 instances corresponding to 55.5% and the class present, which is made up of 120 instances that corresponds to 44.5% of the data set as shown in Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

Table 1. features of heart disease Dataset (Statlog).

#	*Name*	*Type*	*Description*	*features*
1	Age	Continuous	Age in years	Real
2	Sex	Discreet	1 = male 0 = female	Binary
3	Chest-Tdolor	Discreet	Type of chest pain: 1 = typical angina 2 = atypical angina 3 = without angina pa 4 = asymptomatic	Nominal
4	Trestbps-Pa	Continuous	Resting blood pressure (in mm Hg)	Real
5	Chol-Cs	Continuous	Serum cholesterol in mg / dl.	Real
6	Fbs-As	Discreet	Fasting glucose>120 mg / dl: 1 = true 0 = false	Binary
7	Restecg-Re	Discreet	Resting electrocardiographic results: 0 = normal 1 = it has an ST-T wave anomaly 2 = showing probable left ventricular hypertrophy or defined by the Estes criteria	Nominal
8	Thalach-Fcm	Continuous	Maximum heart rate reached	Real
9	Exang–Ei	Discreet	Exercise-induced angina: 1 = yes 0 = no	Binary
10	Oldpeak- Dir	Continuous	ST depression induced by exercise in relation to rest.	Real
11	Slope	Discreet	The slope of the maximum exercise segment: 1 = upward slope 2 = flat 3 = downward slope	Sorted
12	Ca-Nbp	Discreet	number of main vessels (0-3) colored by fluoroscopy	Real
13	Thal-Tt	Discreet	Thallium scan in the heart muscles: 3 = normal; 6 = irreversible defect; 7 = reversible defect	Nominal
14	Class	Discreet	Diagnostic classes: 0 = healthy 1 = possible heart disease	Binary Predictive

3.2. Features selection

An important point to consider is the quantity and quality of the features used. The current trend where machine learning techniques are applied is to incorporate all available features. This is due to the great computing capacity achieved and the fact that many techniques incorporate regularization mechanisms that avoid overfitting and simplify the models automatically. Reducing the independent features allows to reduce the dimensionality, suppress noise sources in the data that can produce biases and improve the interpretability of the models.

There are many approaches to determining the importance of features and thus making a selection of them. As the goal of this work is to show the balance between interpretability and predictive ability of the model, then we will show different metrics that measure the importance of the features in order to train the classifier with different amounts of them. The accuracy and area under roc curve (AUC) results will be compared. Thirteen importance metrics are used, based on statistical tests, anova, kruskal-wallis test (KT) and chi-square test (Chi2), based on entropy: information gain (IG), information gain ratio (GainR), symmetric uncertainty (SU), in measurements on models obtained with random forest: impurity (Rimp), permutation (Rper) and random forest importance (Rfi), based on decision tree models: one rule (OneR) and conditional variable importance (PCF) and the importance given by the relief algorithm (Rel). All these measures are obtained with the implementation done in the mlr library of R [39].

Table 2 shows the ranking value for each measure. The features are ordered by mean value (mean) of all their positions. The final value (rank) represents the order according to the mean. The proportional value between the features has not been considered, only the position in the ranking has been taken in order to be able to compare them easily.

Table 2 shows that the features for most of the metrics are in similar positions, only the Relief measure changes the positions of the features a little more. The rank value will be used to group the features into tertiles. That is, accuracy and AUC will be measured using decision trees with all features then with those in the first and second tertil and finally only with those in the first third. This will

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

Table 2. Importance metrics.

	Chi2	GainR	OneR	SU	IG	Rel	KT	PCF	Rimp	Rper	RFi	AUC	anova	mean	rank
cp	2	2	2	3	2	1	1	2	1	2	3	1	2	1,8	1
thal	1	1	1	1	1	4	6	1	2	3	2	5	7	2,7	2
ca	3	3	3	2	3	2	2	3	4	1	1	4	5	2,8	3
oldpeak	6	6	6	5	6	3	5	5	5	4	4	3	3	4,7	4
thalach	5	5	5	6	5	7	4	6	3	5	6	2	4	4,8	5
exang	4	4	4	4	4	12	3	4	9	8	7	6	1	5,4	6
slope	7	7	7	7	7	10	7	8	10	9	8	7	6	7,7	7
sex	9	9	9	8	9	5	8	7	11	6	5	9	8	7,9	8
age	8	8	8	9	8	9	9	9	6	7	9	8	9	8,2	9
restecg	10	10	10	10	10	11	10	10	12	11	10	10	11	10,4	10
trestbps	12	12	11	12	12	8	11	11	8	10	13	11	10	10,8	11
chol	12	12	11	12	12	6	12	12	7	12	12	12	12	11,1	12
fbs	11	11	11	11	11	5	13	13	13	13	11	13	13	11,5	13

allow to compare the loss of accuracy versus the gain in interpretability using different amount of features.

3.3. Performance Evaluation Methods

In order to measure the balance between performance and interpretability, the set of classification techniques will be compared using a different number of features. A cross-validation will be carried out to measure the performance of the classification task and to observe the average behavior of the algorithms when they have all the features. Then using 66% of the best features according to the ranking in Table 2 and finally with 33% better. In Figures 1 and 2 are shown boxplots, in blue and labeled as 100, the results when all features are used, in green for 66% and red with the third of the most important features. Accuracy and area under ROC curve (AUC) have been taken as quality measures.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

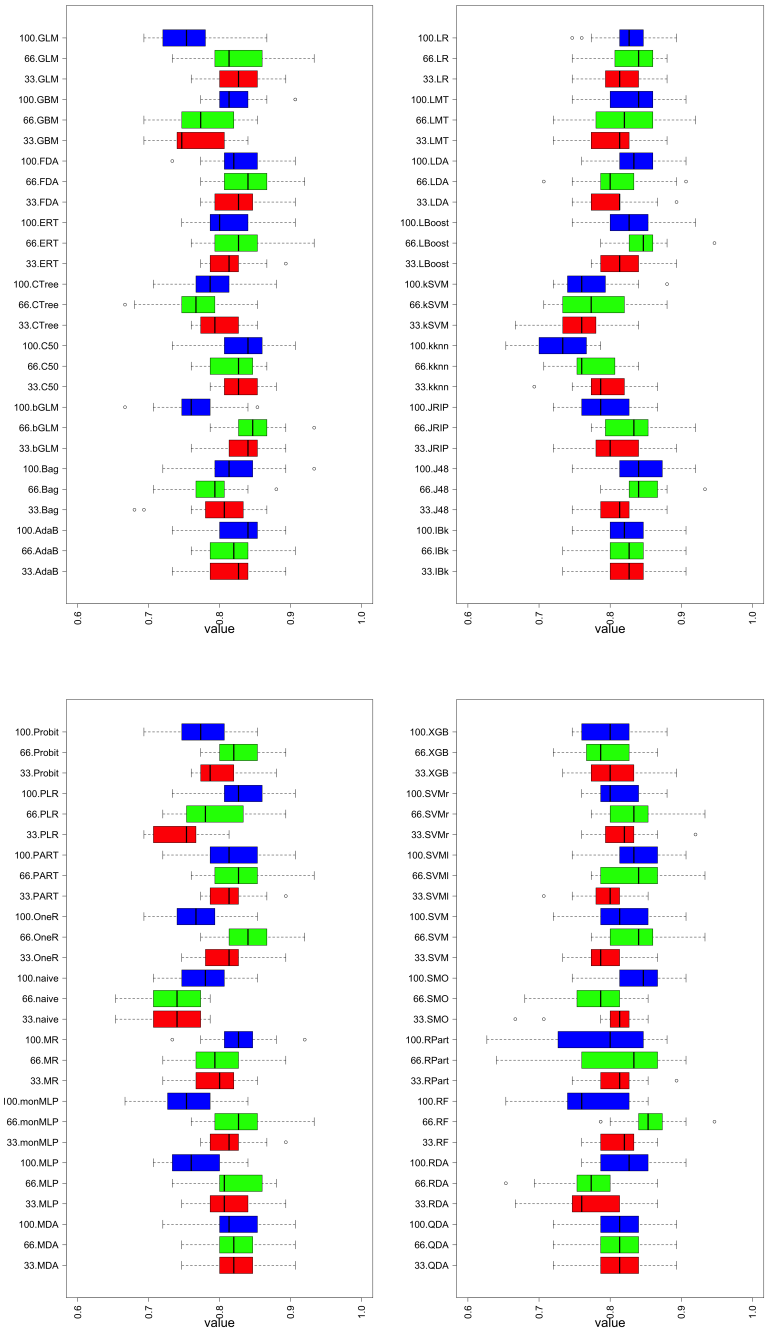

Figure 1. Comparison of classifier performance (Accuracy). features; blue - all, green - best 66%, red - best 33%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

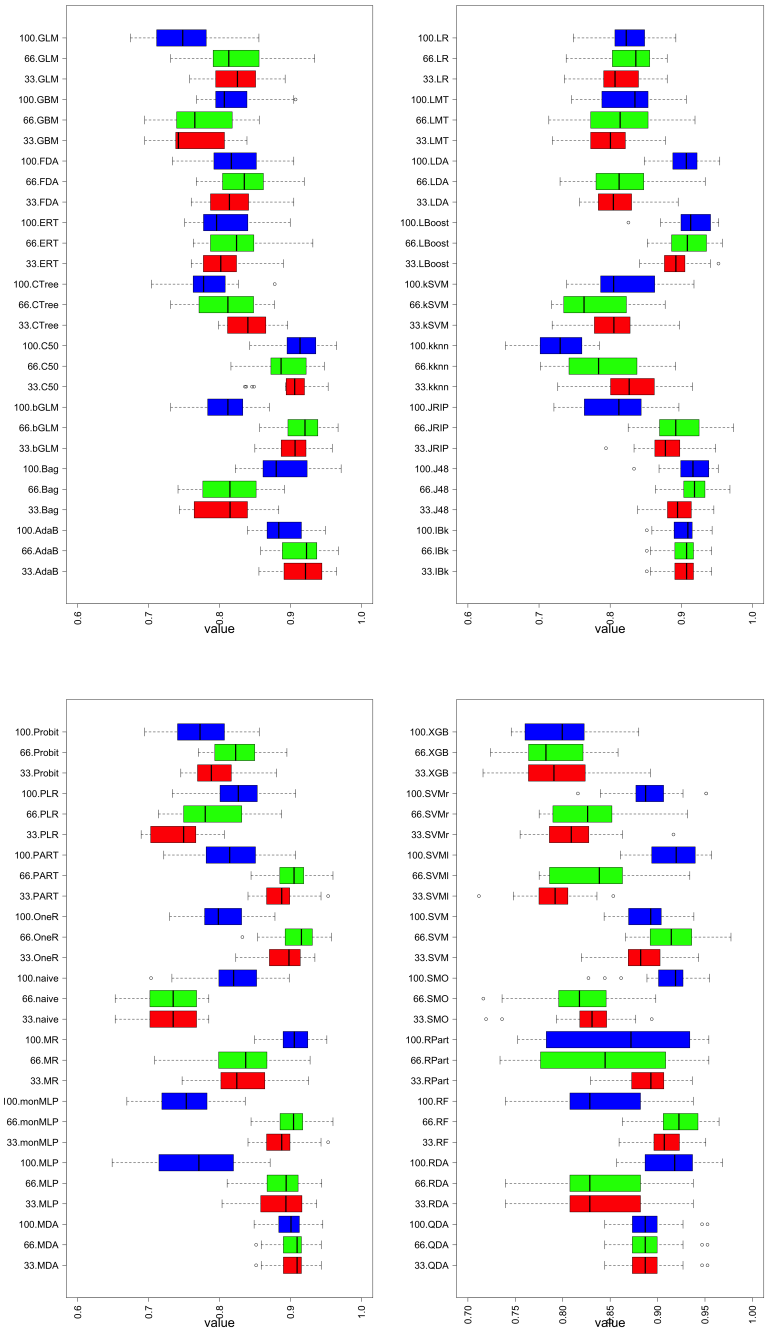

Figure 2. Comparison of classifier performance (AUC). features; blue - all, green - best 66%, red - best 33%

Regardless of the measure of quality, accuracy or AUC metrics, in general terms it is appreciated that better results are obtained by using the more features, although there are exceptions. For example, for discriminant analysis techniques (LDA, FDA) or SVM, better results are obtained with fewer features than using all of them. The reason may be due to the characteristics of the classification algorithms. They try to partition the classification space using hyperplanes to make the decision. Having more features, without a mechanism that can filter or weigh them, introduces noise to perform the partitioning of the classification space. On the other hand there are the algorithms that can perform progressive refinements of the model till the limit where overfitting occurs.Examples of this type are

 doi:10.20944/preprints202012.0318.v1

those that use boosting or bagging (random forest is very characteristic of this behavior) use more features getting better results because can adjust the model in a more complex way, the decision space is subdivided into hyper-rectangles. The disadvantage of these techniques is that they have better results with the cost of making it more difficult to interpret the model that guides its decisions. The decision trees are in an intermediate position, they do not achieve the performance values of the boosting techniques but they improve the classifiers based on hyperplane partitions, maintaining a high understanding of their models. It is for this reason that they will be used in the following discussion of the behavior of complexity versus the quality of the solution.

To see the effect of the number of features on the complexity of the results and the quality associated with them, we will see how AUC varies according to the depth of the tree. There are shown the results of AUC of trees obtained with RPART (Recursive Partitioning And Regression Trees) [37] and Ctree (Conditional inference trees) [38]. One hundred executions of the algorithms are performed using hold-out ($70 - 30\%$) representing in the following figure the mean value and the confidence interval at 95% of the AUC over the test set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

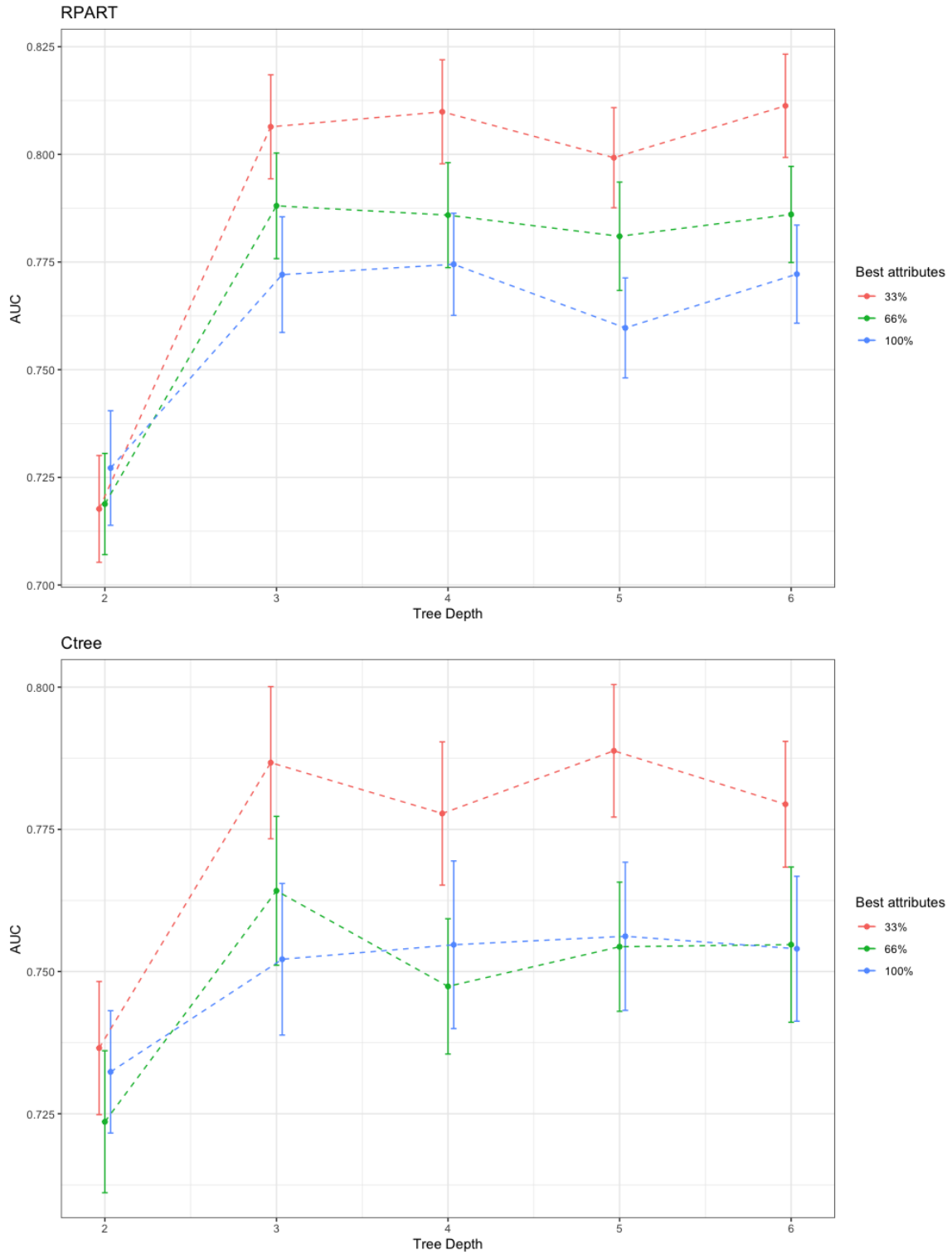

Figure 3. Variation of the performance (AUC) with the tree depth and the percentage of features used. Up- RPART, Down- Ctree.

The behavior of both algorithms is quite similar. Both behave like linear partitioning algorithms, adding features does not improve the quality of the solutions. The curve corresponding to the use of all features always appears below the curves with fewer features, the one with the highest value being the one with the least number of features. If you do not make a good choice of relevant features in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

nodes, then you make a mistake that is propagated to the decision of the class. This fact supports the importance of making a correct choice of features to be used when solving a classification task.

The behavior with the depth level of the best classification tree is also the expected one. It can be seen that as the depth of the tree increases, i.e. its complexity, then its behavior improves. The level of complexity is inversely associated with the ability to interpret the model, the more complex it is, the less interpretable it is. In the figure 3 when the depth of the tree increases, the AUC grows to a maximum level, although the complexity continues growing, it no longer improves the performance. An interesting result is the Pareto front structure that appears between complexity and performance, regardless of the number of features used. If it is assumed that complexity makes the classification model less interpretable, then a design decision must always be made between both of them. In this case, for the problem of classifying cardiovascular diseases, the price is small and the saturation of complexity occurs when depth 3 is reached. For problems with many more instances, a Pareto front with a smoother curve with depth saturation is to be expected. It can also happen that depending on the problem and the technique, the curves corresponding to the amount of features will be reversed. In Figures 4, 5, and 6, it is observed that the variation in AUC is 0.06 for RPART when passing the tree depth from 2 to 3 and has a lower value, approximately 0.05 for CTtree. It is up to the expert analyst to determine if he wants to have a more precise or more interpretable model. It is a design decision similar to finding the elbow point in a clustering problem in which precision must be balanced with the number of clusters obtained. The figure below shows the trees best accuracy corresponding to RPART with different number of features for a maximum depth of 6.

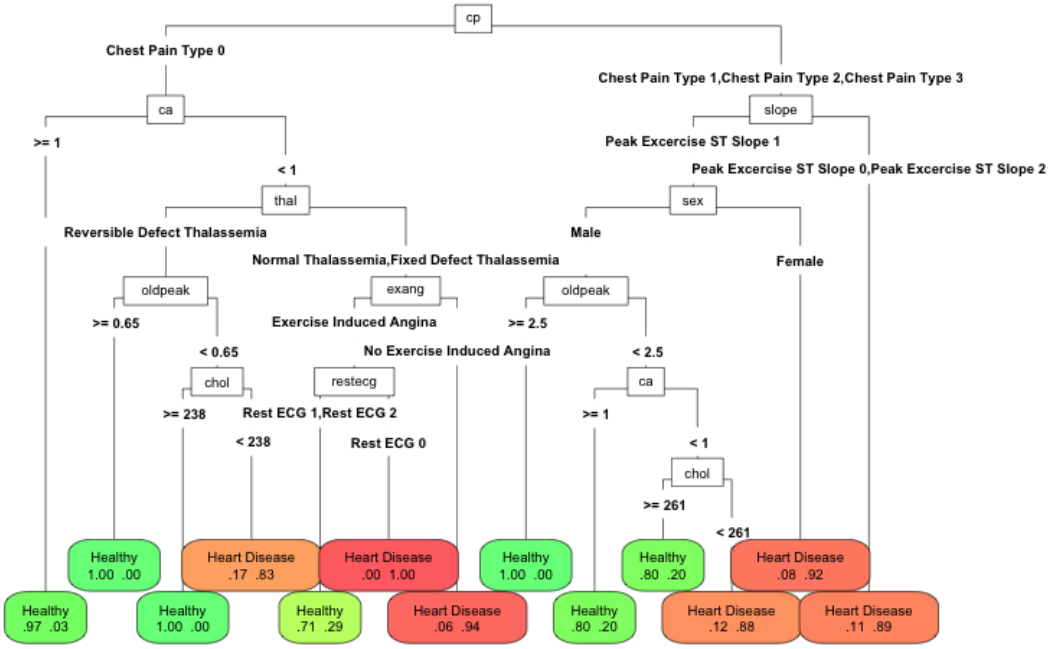

Figure 4. Performance of RPart 100% of features with maximum tree depth 6

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2020 doi:10.20944/preprints202012.0318.v1

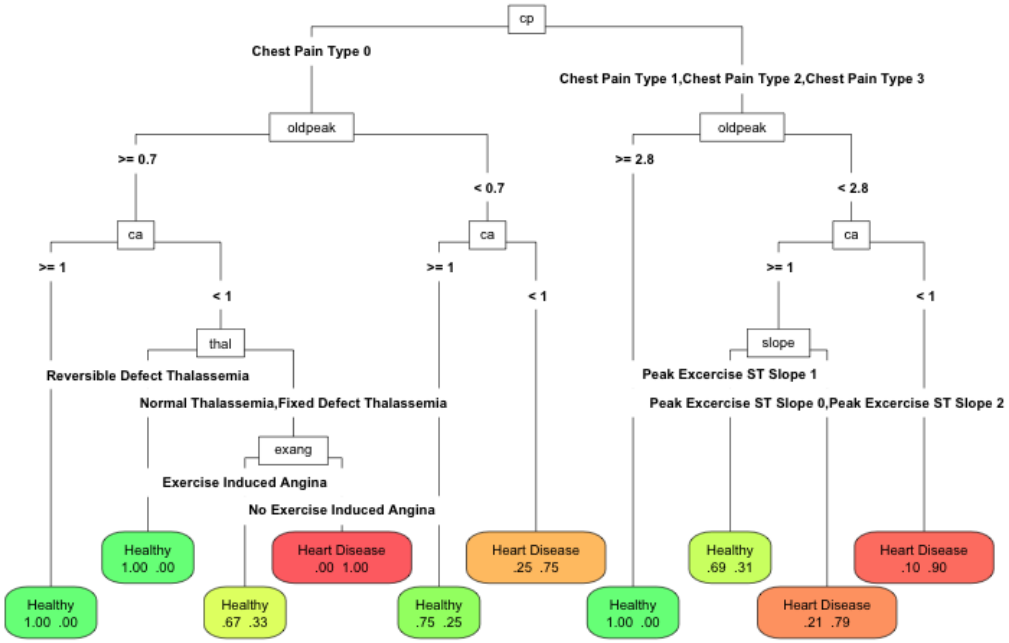

Figure 5. Performance of RPart 66% of features with maximum tree depth 6

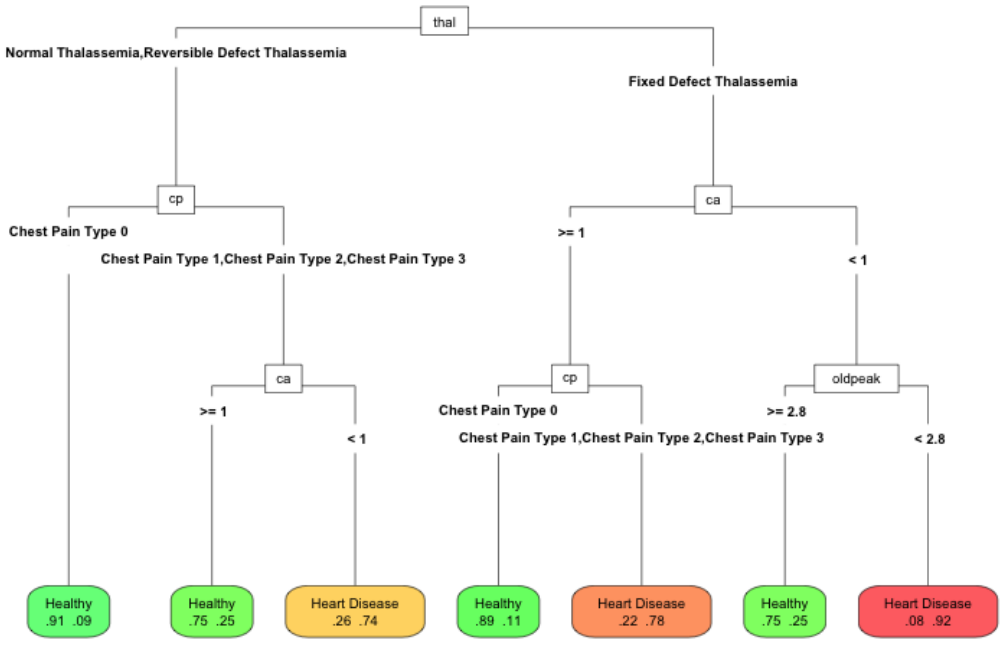

Figure 6. Performance of RPart 33% of features with maximum tree depth 6

The differences between them are minimal, in accuracy, it is about 0.07. With 100% and 66% of the features the depth of the tree is increased (25 and 19 nodes, 9 and 6 features, respectively), the algorithm tries to exploit all the available information. However, with 33% of the best features (13 nodes and 4 features involved), the depth is reduced, since it can no longer use the available features to improve the solution. Therefore, it is more appropriate to reduce the number of available features. In

 doi:10.20944/preprints202012.0318.v1

this case, the number of features is a proxy variable for maximum depth and therefore the complexity. The Pareto front obtained using the depth is similar to the one that would be obtained by modifying the amount of available features. Therefore, the compromise is in the number of features to make a model explainable. In this work we wanted to highlight the great importance of the choice of features to perform machine learning task. Their amount will determine the quality of the solution and the ease of interpretation of the obtained models. This is an issue that is currently being left aside in the era of big data. The computational capacity has grown so much that the algorithms are applied without making a meticulous reduction of the input features. But the need to make more interpretable the algorithm's decisions has become another factor to take into account when applying automatic learning techniques. This paper has shown the relationship between interpretability and the features used and how it is an additional factor for data scientists to consider.

4. Conclusions

This work shows the multi-objective nature of the problem of improving the accuracy of classification versus the "interpretability" of the obtained model. The problem of classifying a cardiovascular disease has been addressed with the Statlog Data Set following the steps of a data analysis process. First, the most relevant features have been characterized, applying a ranking that incorporates the importance of these according to different metrics. Then different classification algorithms are compared by applying different quality metrics with different amounts of features, so that it can be seen that for most techniques, as expected, better results are obtained using a greater number of features. Classification algorithms based on rules and trees have a similar behavior, but somewhat worse than algorithms that use "boosting" processes, but in counterpoint they allow to explain in a simpler way the underlying model. Finally it is shown that for two tree-based algorithms that the reduction of the number of input features not only improves the interpretability of the results but also improves the quality of the solutions. Although a compromise will always have to be found between the complexity that can be achieved with the features and the accuracy in the classification obtained. These results are promising and, in future work, are expected to be applied to other, more comprehensive data sets of a different nature.

Acknowledgments: This work was funded by the public research projects of the Spanish Ministry of Economy, Universidad Carlos III de Madrid, and Competitiveness (MINECO), references TEC2017-88048-C2-2-R, RTC-2016-5595-2, RTC-2016-5191-8 and RTC-2016-5059-8

Conflicts of Interest: "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

References

1. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics, 2019. doi:10.3390/electronics8080832.
2. World Health Organization. Fact sheet: Cardiovascular diseases (CVDs). *World Health Organization* **2017**.
3. Fagard, R.H. Predicting risk of fatal cardiovascular disease and sudden death in hypertension, 2017. doi:10.1097/HJH.0000000000001485.
4. King, R.D.; Feng, C.; Sutherland, A. Statlog: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence* **1995**, *9*. doi:10.1080/08839519508945477.
5. Ansari, M.F.; AlankarKaur, B.; Kaur, H. A prediction of heart disease using machine learning algorithms. Advances in Intelligent Systems and Computing, 2021, Vol. 1200 AISC. doi:10.1007/978-3-030-51859-2{_}45.
6. Turki, T.; Wei, Z. Boosting support vector machines for cancer discrimination tasks. *Computers in Biology and Medicine* **2018**, *101*. doi:10.1016/j.compbiomed.2018.08.006.
7. Nilashi, M.; Bin Ibrahim, O.; Mardani, A.; Ahani, A.; Jusoh, A. A soft computing approach for diabetes disease classification. *Health Informatics Journal* **2018**, *24*. doi:10.1177/1460458216675500.

8. Leslie, H.H.; Zhou, X.; Spiegelman, D.; Kruk, M.E. Health system measurement: Harnessing machine learning to advance global health. *PLoS ONE* **2018**, *13*. doi:10.1371/journal.pone.0204958.
9. Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics* **2020**, *21*. doi:10.1186/s12859-020-03626-y.
10. Fatima, M.; Pasha, M. Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications* **2017**, *09*. doi:10.4236/jilsa.2017.91001.
11. El Houby, E.M. A survey on applying machine learning techniques for management of diseases, 2018. doi:10.1016/j.jab.2018.01.002.
12. Bahadur, S. Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. *IOSR Journal of Agriculture and Veterinary Science* **2013**, *4*, 60–64. doi:10.9790/2380-0426164.
13. Liu, X.; Wang, X.; Su, Q.; Zhang, M.; Zhu, Y.; Wang, Q.; Wang, Q. A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. *Computational and Mathematical Methods in Medicine* **2017**, *2017*. doi:10.1155/2017/8272091.
14. Digital Single Market. Draft Ethics guidelines for trustworthy AI | Digital Single Market, 2019.
15. Zhang, Z.; Xie, Y.; Xing, F.; McGough, M.; Yang, L. MDNet: A semantically and visually interpretable medical image diagnosis network. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, Vol. 2017-January. doi:10.1109/CVPR.2017.378.
16. Hicks, S.A.; Eskeland, S.; Lux, M.; Lange, T.D.; Randel, K.R.; Pogorelov, K.; Jeppsson, M.; Riegler, M.; Halvorsen, P. Mimir: An automatic reporting and reasoning system for deep learning based analysis in the medical domain. Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, 2018. doi:10.1145/3204949.3208129.
17. Choi, E.; Bahadori, M.T.; Kulas, J.A.; Schuetz, A.; Stewart, W.F.; Sun, J. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems, 2016.
18. Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; Gao, J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, Vol. Part F129685. doi:10.1145/3097983.3098088.
19. Sha, Y.; Wang, M.D. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. ACM-BCB 2017 - Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2017. doi:10.1145/3107411.3107445.
20. Rögnvaldsson, T.; Etchells, T.A.; You, L.; Garwicz, D.; Jarman, I.; Lisboa, P.J. How to find simple and accurate rules for viral protease cleavage specificities. *BMC Bioinformatics* **2009**, *10*. doi:10.1186/1471-2105-10-149.
21. Che, Z.; Purushotham, S.; Khemani, R.; Liu, Y. Interpretable Deep Models for ICU Outcome Prediction. *AMIA ... Annual Symposium proceedings. AMIA Symposium* **2016**, *2016*.
22. Wu, M.; Hughes, M.C.; Parbhoo, S.; Zazzi, M.; Roth, V.; Doshi-Velez, F. Beyond sparsity: Tree regularization of deep models for interpretability. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2018.
23. Dua, D.; Graff, C. UCI Machine Learning Repository, 2017.
24. Abdar, M.; Kalhori, S.R.; Sutikno, T.; Subroto, I.M.I.; Arji, G. Comparing performance of data mining algorithms in prediction heart diseses. *International Journal of Electrical and Computer Engineering* **2015**, *5*. doi:10.11591/ijece.v5i6.pp1569-1576.
25. El-Bialy, R.; Salamay, M.A.; Karam, O.H.; Khalifa, M.E. Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science, 2015, Vol. 65. doi:10.1016/j.procs.2015.09.132.
26. Naushad, S.M.; Hussain, T.; Indumathi, B.; Samreen, K.; Alrokayan, S.A.; Kutala, V.K. Machine learning algorithm-based risk prediction model of coronary artery disease. *Molecular Biology Reports* **2018**, *45*. doi:10.1007/s11033-018-4236-2.
27. Chaurasia, V.; Pal, S. Data Mining Approach to Detect Heart Diseases. *International Journal of Advanced Computer Science and Information Technology* **2013**, *2*, 56–66.
28. Dr. S. Vijayarani1, M. Liver Disease Prediction using SVM and Naïve Bayes Algorithms. *International Journal of Science, Engineering and Technology Research* **2015**, *4*.

29. Zhao, Y.; Healy, B.C.; Rotstein, D.; Guttmann, C.R.; Bakshi, R.; Weiner, H.L.; Brodley, C.E.; Chitnis, T. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS ONE* **2017**, *12*. doi:10.1371/journal.pone.0174866.
30. Samuel, O.W.; Asogbon, G.M.; Sangaiah, A.K.; Fang, P.; Li, G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Systems with Applications* **2017**, *68*. doi:10.1016/j.eswa.2016.10.020.
31. Jin, B.; Che, C.; Liu, Z.; Zhang, S.; Yin, X.; Wei, X. Predicting the Risk of Heart Failure with EHR Sequential Data Modeling. *IEEE Access* **2018**, *6*. doi:10.1109/ACCESS.2017.2789324.
32. Long, N.C.; Meesad, P.; Unger, H. A highly accurate firefly based algorithm for heart disease prediction. *Expert Systems with Applications* **2015**, *42*. doi:10.1016/j.eswa.2015.06.024.
33. Pawar, U.; O'Shea, D.; Rea, S.; O'Reilly, R. Explainable AI in Healthcare. 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020, 2020. doi:10.1109/CyberSA49311.2020.9139655.
34. Ahmad, M.A.; Teredesai, A.; Eckert, C. Interpretable machine learning in healthcare. Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, 2018. doi:10.1109/ICHI.2018.00095.
35. Rudin, C. Please Stop Explaining Black Box Models for High Stakes Decisions. *NIPS 2018* **2018**.
36. Towards trustable machine learning, 2018. doi:10.1038/s41551-018-0315-x.
37. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and regression trees*; 2017. doi:10.1201/9781315139470.
38. Hothorn, T.; Hornik, K.; Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* **2006**, *15*. doi:10.1198/106186006X133933.
39. Bischl, B.; Lang, M.; Kotthoff, L.; Schiffner, J.; Richter, J.; Studerus, E.; Casalicchio, G.; Jones, Z.M. Mlr: Machine learning in R. *Journal of Machine Learning Research* **2016**, *17*.