Assessment of Heart Patients Dataset Through Regression and Classification Algorithms

Eliane Somavilla(✉) and Gustavo Medeiros de Araujo

PGCIN, Federal University of Santa Catarina, Florianópolis, Brazil
eliane.somavilla@gmail.com, gustavo.araujo@ufsc.br

Abstract. This study focused on the survival analysis of patients with heart failure who were admitted to the Institute of Cardiology and Allied Hospital of Faisalabad-Pakistan during April and December 2015. All patients had left ventricular systolic dysfunction, belonging to classes III and IV of the classification carried out by the New York Heart Association. Several Machine Learning algorithms capable of analyzing data through regression and classification techniques were used to predict the mortality rate of future patients with similar problems. Characteristics such as age, ejection fraction, serum creatinine, serum sodium, anemia, platelets, creatinine phosphokinase, blood pressure, diabetes and smoking were considered as potential contributors to mortality. All characteristics were analyzed in order to identify the minimum set of information necessary for a quick and efficient diagnosis of heart failure.

Keywords: Heart patients · Regression algorithms · Classification algorithms · Feature select

1 Introduction

Annually, cardiovascular disease kills millions of people worldwide and manifests itself primarily as myocardial infarctions and heart failure. Heart failure is characterized by the heart's inability to act as a pump, either by a deficit in contraction and/or relaxation, compromising the body's functioning. When untreated, the quality of life is reduced, which can lead to the patient's death [1].

The main cardiovascular risk factors are directly related to factors such as high blood pressure, diabetes, smoking, dyslipidemia (fats in the blood) and sedentary lifestyle. The presence of these factors alone or together in an organism causes the development of coronary disease that can lead to acute myocardial infarction or decreased heart performance [1].

Other causes include diseases that affect the heart valves (degenerative or inflammatory, such as kidney and rheumatic diseases), congenital diseases, genetic diseases, autoimmune, inflammatory (Eg: during peripartum), toxicity (cancer treatment) and also infectious (Eg: AIDS). The abuse of alcoholic beverages and drugs such as cocaine also increases the risk of developing heart disease which is now responsible for 31% of deaths worldwide [1].

A. L. Pinto and R. Arencibia-Jorge (Eds.): DIONE 2022, LNICST 452, pp. 178–190, 2022.
https://doi.org/10.1007/978-3-031-22324-2_14

Diagnosis of heart failure is made clinically by a medical professional specialized in cardiology, through the history told by the patient. The report may include intolerance to exertion, shortness of breath at bedtime, and swelling in the lower limbs and abdomen. The diagnosis usually also considers the physical examination of the accumulation of blood in the lungs and in the body as a whole [1].

Left ventricular systolic involvement is responsible for most cases of chronic heart failure and can be diagnosed echocardiographically by a left ventricular ejection fraction equal to or less than 0.40. Diastolic dysfunction is characterized by symptoms of heart failure with preserved ejection fraction (usually > 0.45) [2].

Heart failure can be established in any age group and its prevalence is estimated at 1 to 2% of the population. In addition, it affects progressively with increasing age, and after 70 years of age, more than 10% of the population is affected. After the age of 55, there is an approximate 30% risk of developing heart failure [1].

Aiming to support health professionals in the diagnosis of their patients, the objective of this study is to estimate the mortality rates due to heart failure and prove its connection with some of the main risk factors pointed out by the Hospital Israelita Albert Einstein in the city of São Paulo/ SP-BR and SCIELO-Brasil, choosing a reliable database for the study and application of Machine Learning algorithms specialized in data analysis through regression and classification techniques.

Also considering that the reality of many hospitals and clinics in Brazil and in several other countries does not contribute to a complete and detailed diagnosis for patients, in this study we assess what are the minimum characteristics that a health professional must have access to make a quick diagnosis and effective with respect to heart failure.

2 Related Work

The original version of the dataset used in this study was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab and Muhammad Ali Raza (Faculty of Government University, Faisalabad, Pakistan) and made available in July 2017. The current version of the dataset was compiled by David Chicco (Krembil Research Institute, Toronto, Canada) and donated to the University of California Machine Learning Repository under the same copyright in January 2020 [3].

In the study published in 2017 by Ahmad, Munir, Bhatti, Aftab and Raza, the collected data were analyzed using the COX regression model. Using the COX regression model, the risk of death of an individual was linked with one or more variables and the significance of these variables was also tested. The results were validated by calculating the calibration slope and the model's discrimination ability via bootstrapping [5].

As a result, it was observed that the chances of death from heart failure increase with increasing age. The risk of death increases by 4% for each additional year of age. Ejection Fraction (EF) was another significant factor, the hazard ratio among patients with EF ≤ 30 was 67% and 59% higher compared to patients with EF 3045 and EF ≥ 45, respectively. Other data observed indicates that the risk of death doubles when serum creatinine is above its normal level. The study shows that an anemic patient is 76% more likely to die compared to a non-anemic one. Increasing serum sodium reduces the risk of death by 6%. Data on sex, smoking, high blood pressure, diabetes, creatinine phosphokinase and platelets were not significant. Through graphs to measure the model's calibration, an accuracy of 77% was proven, considering the death event occurred with an average of 50 days of medical follow-up and 81% when considering the death event occurred with 250 days of treatment doctor. This shows that the Cox model's skill is greater at a longer follow-up time [5].

In 2020 researchers Chicco and Jurman published a new study with this same database using machine learning, whose focus was on the use of the Random Forests algorithm compared to several biostatistics techniques, that is, statistical methods focused on the biological and medical field.. The methods used were the Mann-Whitney U test, the Pearson correlation coefficient and the chi-square test to compare the distribution of each characteristic between the two groups (surviving individuals and dead patients). In addition to the Shapiro-Wilk test to verify the distribution of each resource [4].

As a result, traditional biostatistical analysis selected ejection fraction and serum creatinine as the two most relevant characteristics. The same result was obtained with machine learning. In addition, our approach has shown that machine learning can be used effectively for binary classification of electronic health records from patients with cardiovascular disease [4].

3 Methodology

In this study, we analyzed a database with real information about patients with heart failure using machine learning algorithms. In this way, computers have the ability to earn according to expected responses through associations of different data.

Another benefit achieved refers to the selection of resources, because, through specific functions of the algorithms, it is possible to automatically select the subset of data considered most relevant to define the result of the computational analysis. In this way, we improve computational efficiency and reduce the model's generalization error by removing extraneous data or noise.

In this study we use several Machine Learning algorithms to analyze the result achieved through different perspectives. We made use of the Python programming language and the main library used was sklearn, which offers a set of classification and regression algorithms. Before applying the algorithms, we carry out the initial processing of the database.

3.1 Pre-processing

After initial analysis of the data that make up the dataframe, we divided the database composed of 299 records into two sets, X and y, respectively. The set X corresponds to the independent variables of the dataframe, which are the data from the blood tests and the echocardiogram of each patient. The set corresponds to the dependent variable, y, indicates whether the patient died or not considering the result of the combination of variables independently assigned to each patient (Fig. 1).

Next, we use the SelectKBest resource belonging to the feature_selection module of the sklearn library. This feature removes all variables, except K features with the

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
0	75.0	0	582	0	20	1	2	1.9	130	0
1	55.0	0	7861	0	38	0	2	1.1	136	0
2	65.0	0	146	0	20	0	2	1.3	129	1
3	50.0	1	111	0	20	0	2	1.9	137	0
4	65.0	1	160	1	20	0	2	2.7	116	0

Fig. 1. Fragment of the set X

highest score, where K is a parameter that indicates the number of variables that should be evaluated by the resource used. [11].

The next step was the division of sets X and y in order to properly train the algorithms and perform machine learning. 80% of the data was used for training and 20% of the data was intended for testing in order to verify the behavior of each algorithm when receiving new data and make sure that the algorithm really learned based on the data used during the training.

Afterwards, we use the StandardScaler feature to normalize the values of variables X that will be used for training and testing. This feature transforms the data so that the distribution has a mean value of 0 and a standard deviation of 1 [14].

Finally, we ran the entire pre-processing process 10 times and the algorithms were run with increasingly smaller datasets considering the K most important characteristics pointed out by SelectKBest. The result obtained in each round was recorded for future comparison after applying other similar resources, but more suitable for each algorithm.

3.2 Regression Algorithms

Regression is one of the most used forecasting methods in the statistical field. The main objective is to verify how the variables of interest influence a response variable (y), in addition to creating a mathematical model capable of predicting y values based on new values of predictor variables X [15].

Multivariate linear regression is indicated when the set X includes more than one predictor variable. In our study, initially, 10 variables are considered in set X [16].

Once the pre-processing is executed, the recursive elimination of resources (RFE) is used to evaluate the best characteristics specifically for this algorithm through the RFE function. Given an external estimator that assigns weights to resources (for example, the coefficients of a linear model), the objective of recursive resource elimination (RFE) is to select variables considering smaller and smaller sets. First, the estimator is trained on the initial set of features and the importance of each feature is obtained through any specific attribute (such as coef_ or feature_importance). Then, the less important features are removed from the current set. This procedure is repeated on the remaining set until the desired number of features to be selected is reached [13].

Polynomial regression makes use of polynomials, which are algebraic expressions, used in cases where the model variables do not follow a linear form of data. In this case, the independent variables are expanded by the number of polynomials, as configured. In the present study, the best results for this algorithm were observed using the degree 3 of polynomials [15].

After pre-processing, we again use recursive resource elimination (RFE) to identify the importance of each variable from set X to the model.

The result obtained was expressed through the metric R2_score that calculates the coefficient of determination. This coefficient provides a metric of how well future samples are likely to be predicted by the model. The best possible score is 1.0 and can be negative because the model can be arbitrarily worse. The same metric was used with the Multivariate Linear Regression algorithm to measure the accuracy of the model during the test phase [12].

This entire process was performed 10 times for each of the regression algorithms. Each round considered the best variables according to the resource selection method used. The result obtained in each round was recorded for future comparison after applying similar resources, but more suitable for each algorithm.

3.3 Classification Algorithms

Classification algorithms aim to classify items or samples according to observed characteristics. Through algorithms that implement this method, it is possible to teach the computer to carry out the classification in order to identify which category a given data sample belongs to [15].

KNN (K-Nearest Neighbors) is a classification and regression algorithm whose objective is to determine to which group a sample belongs based on the distance to K neighboring samples [15]. In the present study, considering all the variables of the independent set, the best results for this algorithm were observed through the analysis of the 6 closest neighbors.

After the pre-processing, we use the Direct Sequential Selection through the SequentialFeatureSelector (SFS) function belonging to the feature_selection module of the sklearn library. SFS removes or adds one resource at a time based on the performance of the classifier until the desired K size resource subset is reached [17].

The RFC (Random Forest Classifier) is a classification and regression algorithm whose objective is to determine to which group a given sample belongs based on decision trees created at the time of training [15]. In the present study, considering all the variables of the independent set, the best results for this algorithm were observed through the creation of 7 decision trees. After pre-processing, we use the feature_importances function available to all ensemble algorithms, such as RFC. This method returns a score for each variable. The higher the score, the greater the importance of this variable for the result of the computational analysis [18].

The SVM (Support Vector Machine) is a classification and regression algorithm whose objective is to determine which group a given sample belongs by means of a separation line between two classes. This line is commonly called the Hyperplane. This algorithm has a theory a little more robust compared to other simpler algorithms such as KNN. The SVM can be configured to work with linear problems, using the Linear Kernel, or nonlinear ones, using the RBF Kernel (Radial Basis Function), which performs a mapping to a larger space [10].

After preprocessing the data, we used recursive resource elimination (RFE) to identify the importance of each variable from set X to the model.

In the first tests performed in this study, the Linear kernel reached 75% accuracy considering the entire set of initial independent variables and showed similar accuracy as the resources were recursively eliminated. Kernel RBF reached the same percentage from the moment that set X was made up of 8 variables. We chose to follow the study with the Linear Kernel.

The Logistic Regression, is a classification algorithm considered the most used statistical method to model categorical variables. Logistic Regression and Linear Regression are similar. While in Linear Regression we have a continuous independent variable, in Logistic Regression the response variable is binary [15].

After preprocessing the data, we used recursive resource elimination (RFE) to identify the importance of each variable from set X to the model.

The result obtained for all classification algorithms was expressed through a confusion matrix, which is a comparative table of the values that each algorithm brought as a prediction in relation to the real values that occurred [15]. That is, after training a model and applying the predictions on the separate dataset for testing, the result obtained is expressed in a column with the predictions.

This entire process was performed 10 times for each of the regression algorithms. Each round considered the best variables according to the resource selection method used. The result obtained in each round was recorded for future comparison after applying similar resources, but more suitable for each algorithm.

4 Exploratory Data Analysis

4.1 Dataset

The database used for the study is part of the Machine Learning Repository (UCI) file repository that maintains a collection of databases and data generators used by the community to feed machine learning algorithms. Created in 1987. Since then, it has been widely used by students, educators and researchers around the world as a primary source of machine learning datasets [3].

The data set chosen for the study contains the medical records of 299 patients with heart failure and all patients had left ventricular systolic dysfunction, belonging to classes III and IV of the classification performed by the New York Heart Association. Data collection was carried out during the follow-up period in 2015, specifically between the months of May and December. This data period was the most detailed dataset available at the UCI [3].

Each patient profile has 13 clinical features. Are they:

1. **Age**: patient age (years)
2. **Anemia**: decreased red blood cells or hemoglobin (boolean)
3. **High blood pressure**: if the patient has hypertension (Boolean)
4. **Creatinine phosphokinase (CPK)**: blood CPK enzyme level (mcg / L)
5. **Diabetes**: if the patient has diabetes (Boolean)
6. **Ejection fraction (EF):** percentage of blood that leaves the heart with each contraction (percentage)
7. **Platelets**: platelets in blood (kilo plates / mL)

8. **Gender**: female or male (binary)
9. **Serum creatinine:** blood serum creatinine level (mg / dL)
10. **Serum sodium:** blood serum sodium level (mEq / L)
11. **Smoking**: whether the patient smokes or not (Boolean)
12. **Time**: follow-up period (4 to 285 days, with an average of 130 days)
13. **Death event:** if the patient died during the follow-up period (Boolean)

4.2 Exploratory Analysis

The current study is based on 299 heart failure patients, 105 women and 194 men. For the quantitative assessment of disease progression, physicians typically use the New York Heart Association (NYHA) functional classification, which determines four classes ranging from the absence of symptoms of normal activity (Class I) to a stage in which any activity causes discomfort and symptoms occur at rest (Class IV) [4].

All patients were between 40 and 95 years old, had left ventricular systolic dysfunction, and had previous heart failure that placed them in class III or IV of the New York Heart Association (NYHA) classification. The disease was diagnosed in patients in this database by echocardiogram report, blood tests or notes written by the physician [4].

Follow-up time ranged from 4 to 285 days, with an average of 130 days. Follow-up time was not considered as an input parameter for this study, as it aimed to focus on data related to the cause of heart failure and not on the length of time the patient was under follow-up and medical treatment. The gender of patients (sex) was also not considered as an input parameter as it is understood that it is only a characteristic used to categorize patients into two distinct groups.

After disregarding the length of follow-up and gender variables, the other characteristics were used as potential variables to explain mortality from heart failure. Platelets were divided into three levels based on medical references per liter of blood [6]: values below 150000 indicate a low index, between 150000 and 450000 indicate a normal index and above 450000 indicate a high index related to the presence of thrombi or blood clotting. Serum creatinine greater than its normal level (1.5) is an indicator of renal dysfunction [4].

Among the 299 patients studied, 96 were smokers, 105 had high blood pressure, 67 had creatinine greater than 1.5, 125 had anemia, 129 had diabetes and 219 had ejection fraction less than or equal to 40, that is, the percentage of blood pumped out of a left ventricle with each heartbeat was up to 40%. A total of 252 patients had a normal platelet count, which means that there was no thrombus or blood clotting. A total of 96 patients died during the study period.

Sodium is a mineral that serves the correct functioning of muscles and nerves. An abnormally low level of sodium in the blood can be caused by heart failure [7]. Sodium is considered low when lower than 136 (mEq / L), and in these cases the patient has hyponatremia, which is the most common electrolyte disturbance in hospitalized patients. The presence of hyponatremia is associated with the need for admission to the intensive care unit, prolonged hospitalization and mortality. It is still unclear whether there is a direct causal relationship or whether hyponatremia is just a marker of underlying disease severity. However, it is known that the inadequate management of a patient with hyponatremia can cause severe neurological damage or death [8]. In the study, a total of 200 patients have normal levels of sodium in the body and 99 have low levels.

Finally, creatinine phosphokinase (CPK) indicates the level of the CPK enzyme in the blood. When muscle tissue, such as the heart and brain, is damaged, CPK flows into the blood. Therefore, high levels of CPK in a patient's blood may indicate heart failure or injury. Reference values for creatine phosphokinase (CPK) are 32 and 294 U/L for men and 33 to 211 U/L for women, but may vary depending on the laboratory where the test is performed [9]. Considering this reference, 91 men and 58 women have an indication of damaged muscle tissue. Graphically, we can see the following percentages (Fig. 2):

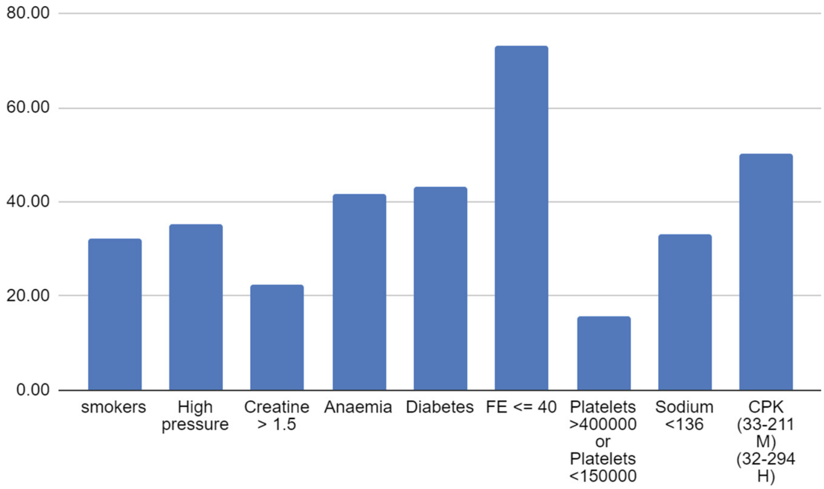

Fig. 2. Percentage of patients with a certain characteristic

5 Results and Discussion

When evaluating the results achieved through the SelectKBest resource, we can observe that the best accuracy was achieved by the classification algorithms, considering a minimum set of 3 to 4 variables for the model.

Among the 4 algorithms evaluated in this category, we observed that the KNN had a notable improvement in the accuracy rate when considering only 3 variables in the model, reaching 77% confidence. The other classification algorithms presented consistent data according to the elimination of variables from the model, obtaining a low variation in the accuracy rate. When considering only 2 variables, the model lost effectiveness for all algorithms (Table 1).

We noticed through this feature that the regression algorithms did not prove to be good options to define whether a patient will die because, in the case of Multivariate Regression, the accuracy of the model obtained a low percentage regardless of the number

Table 1. Result with SelectBest feature for evaluating variables

General Evaluation - Accuracy Using the Resource "SelectBest"									
Algorithms	10 Featur.	9 Featur.	8 Featur.	7 Featur.	6 Featur.	5 Featur.	4 Featur.	3 featur.	2 Featur.
KNN	0.65	0.65	0.66	0.66	0.7	0.7	0.65	0.77	0.7
SVM	0.75	0.73	0.72	0.73	0.72	0.73	**0.73**	0.72	0.63
Random forest	0.65	0.73	0.68	0.66	0.7	0.7	0.66	**0.72**	0.66
Logistic regression	0.72	0.73	0.72	0.72	0.7	0.72	0.73	**0.73**	0.65
Polynomial regression	**0.94**	0.32	0.75	0.64	0.49	0.45	0.39	0.33	0.14
Multivariate regression	0.15	0.16	0.17	0.17	0.17	0.17	0.17	0.16	0.06

of variables that made up the model. The Polynomial Regression Algorithm obtained accuracy greater than 90% when considering the 10 variables in the model, however, when eliminating variables the accuracy decreased and we did not verify linearity in the percentage obtained as a result. Thus, it is considered that this algorithm did not show consistency in the same way that we observed in classification algorithms. Specifically for the Polynomial Regression, it could be evaluated if the model presented overfitting, which occurs when the model fits well to the training dataset, but proves to be ineffective to predict new results (Fig. 3).

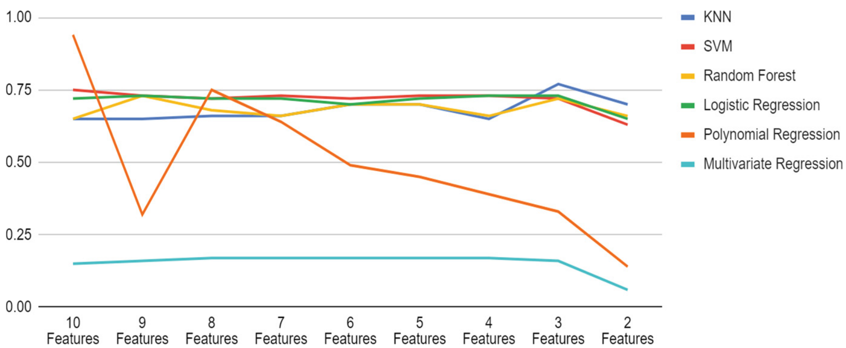

Fig. 3. Accuracy achieved through the SelectBest feature for evaluating variables

5.1 Evaluating Algorithms with Other Resources

When evaluating the results achieved through specific resources used to identify the best variables of each algorithm, we can observe that the best accuracy was achieved by each algorithm with a distinct amount of minimum variables (Table 2).

Table 2. Result with different ways to evaluating variables

General Evaluation - Accuracy Using Different Methods for each Algorithm									
Algorithms	10 Featur.	9 Featur.	8 Featur.	7 Featur.	6 Featur.	5 Featur.	4 Featur.	3 Featur.	2 Featur.
KNN	0.65	0.62	0.65	0.65	0.65	0.65	0.65	**0.68**	0.7
SVM	0.75	0.75	0.75	0.73	**0.75**	0.72	0.72	0.72	0.63
Random Forest	0.65	0.68	0.7	0.66	0.7	0.68	**0.75**	0.66	0.56
Logistic Regression	0.72	0.72	0.72	0.73	0.73	0.75	**0.75**	0.73	0.68
Polynomial Regression	**0.94**	0.72	0.72	0.69	0.57	−0.66	0.34	0.33	0.16
Multivariate Regression	0.1564	0.1548	0.1552	0.1511	0.1655	0.1667	0.1522	0.1598	−0.0077

We note that:

- The **KNN** algorithm, which used the feature_selection feature of the SequentialFeatureSelector function from the mlxtend library, obtained better accuracy when used with a minimum of **2 variables**. However, the algorithm maintained low variation, regardless of using the 10 model variables or only the reduced set of 2 variables.
- The **SVM**, which used the feature_selection feature of the RFE function from the sklearn library, obtained the best result with a minimum of **6 variables**. However, the algorithm kept low variation, regardless of using the 10 model variables or just the reduced set of 3 variables. Its accuracy dropped considerably when only 2 variables were evaluated.
- The **Random Forest Classifier**, which used the feature_importances resource (proper methods of assembly algorithms), obtained the best result with a minimum of **4 variables**. It is observed that as the variables were eliminated from the model, the accuracy increased, however, there was a reduction in the accuracy rate when evaluating the algorithm with a reduced set of 7 or 3 variables. When considering a set of only 2 variables, the accuracy drops considerably.
- The **Logistic Regression**, which used the feature_selection feature of the RFE function from the sklearn library, obtained better accuracy when used with a minimum of **4 variables**. However, the algorithm kept low variation, regardless of using the 10 model variables or just the reduced set of 3 variables. Its accuracy dropped considerably when only 2 variables were evaluated.

- The **Polynomial Regression**, which used the feature_selection feature of the RFE function from the sklearn library, obtained the best result when evaluating the total set of 10 variables. As the variables were eliminated from the model, the accuracy dropped considerably, reaching negative values
- The **Multivariate Regression**, which used the feature_selection feature of the RFE function from the sklearn library, did not present an accuracy higher than 16.5% at any time, even presenting negative values.

Graphically, we clearly observe the variations (Fig. 4):

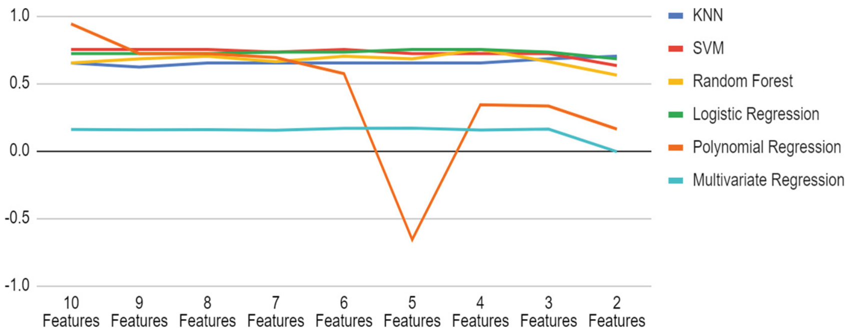

Fig. 4. Accuracy achieved through various variable evaluation features

5.2 Conclusion

All algorithms used to classify the data sample obtained consistent and linear results. Furthermore, it was confirmed that the Polynomial Regression and Multivariate Regression algorithms do not present efficient results to predict the death of a patient.

In general, age, ejection fraction (EF) and serum creatinine were the main variables to predict the mortality rate through classification algorithms. Data about anemia and diabetes, the amount of platelets in the blood and if the patient is a smoker or not were the variables that least influenced the results obtained. Variables related to high blood pressure, sodium and creatinine phosphokinase (CPK) proved to be important contributors but not determinants of the possibility of survival of a patient.

Thus, we can conclude that, through a reduced data set consisting of age, ejection fraction (EF) and serum creatinine, it is possible to predict the probability of death of patients considering an accuracy rate greater than 70% for all algorithms classification used in this study.

Considering the scenario of clinics and hospitals that often do not have all the resources needed to assess the condition of a patient with heart failure, this study directs health professionals to prioritize the data that are most relevant to the probability of survival of a patient.

6 Future Work

Considering the results obtained, a score could be developed to classify the patient's risk of dying, according to the data obtained for the variables studied. Variables that did not have a great influence on the model, such as anemia, diabetes, platelets and whether the patient is a smoker or not, could have a lower weight. On the other hand, Ejection Fraction (EF), age and serum creatinine would have a higher weight. Variables related to high blood pressure, sodium and creatinine phosphokinase (CPK) could have an intermediate weight. In this way, healthcare professionals could rely on a formula to predict the survival rate of patients with heart failure.

During the research studies were found on the influence of quality of life on treatment success for patients with heart failure. According to the Hospital Israelita Albert Einstein (2021), sedentary lifestyle is one of the causes of cardiovascular diseases that can be prevented with physical exercise.

Quality of life is also influenced by culture, income, access to information and location in which each person lives, as these variables can limit access to activities that help prevent cardiovascular disease [1]. If there was a database with information about the quality of life of patients, we could analyze the results of the study considering this reality.

References

1. Einstein, H.I.A.: https://www.einstein.br/especialidades/cardiologia/doencas-sintomas/insuficiencia-cardiaca. Accessed 03 July 2021
2. Scielo. Tratamento da insuficiência cardíaca -- aspectos atuais. Revista da Associação Médica Brasileira, vol. 44 (2020). https://www.scielo.br/j/ramb/a/npPgV7NzQ99tWdQhvZJvCfs/?lang=pt. Accessed 03 July 2021
3. UCI. Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. Accessed 01 July 2021
4. Chicco, D., Jurman, G.: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med. Inf. Decis. Mak. **20**(16) (2020). https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5. Accessed 02 July 2021
5. Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M., Raza, M.A.: Survival analysis of heart failure patients: a case study. Plos One (2017). https://doi.org/10.1371/journal.pone.0181001. Accessed 02 July 2021
6. Lemos, M.: https://www.tuasaude.com/plaquetas/. Accessed 03 July 2021
7. Stephens, C.: https://www.healthline.com/health/sodium-urine#purpose. Accessed 17 July 2021
8. Rocha, P.N.: Hiponatremia: conceitos básicos e abordagem prática. Revista da Associação Médica Brasileira (2011). https://www.scielo.br/j/jbn/a/ggcdv7X6mjHSyVRtcY8fTxS/?lang=pt&format=pdf. Accessed 07 July 2021
9. Lemos, M.: https://www.tuasaude.com/exame-cpk/. Accessed 06 July 2021
10. de Oliveira Júnior, G.M.: Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado. Trabalho de Conclusão de Curso (Graduação em Ciências da Computação) – Instituto de Matemática e Estatística da Universidade Federal de Pernambuco – Pernambuco (2010). https://www.cin.ufpe.br/~tg/2010-2/gmoj.pdf. Accessed 22 July 2021

11. Santana, R.: https://minerandodados.com.br/aprenda-como-selecionar-features-para-seu-modelo-de-machine-learning/. Accessed 08 July 2021
12. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html?highlight=r2%20score#sklearn.metrics.r2_score. Accessed 17 July 2021
13. Scikit-Learn. https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_digits.html#sphx-glr-auto-examples-feature-selection-plot-rfe-digits-py. Accessed 15 July 2021
14. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?highlight=standardscaler#sklearn.preprocessing.StandardScaler.fit_transform. Accessed 17 July 2021
15. dos Santos, H.G.: Comparação da Performance de Algoritmos de Machine Learning para Análise Preditiva em Saúde Pública e Medicina. Tese (Doutorado em Ciências) - Programa de pós Graduação em Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo, São Paulo (2018). https://www.teses.usp.br/teses/disponiveis/6/6141/tde-09102018-132826/publico/HellenGeremiasdosSantos_DR_ORIGINAL.pdf. Accessed 20 July 2021
16. Nogueira, F.E.: Modelos de Regressão Multivariada. Dissertação (Mestrado em Ciências) – Instituto de Matemática e Estatística da Universidade de São Paulo – São Paulo (2007). https://teses.usp.br/teses/disponiveis/45/45133/tde-25062007-163150/publico/dissertacao_4.pdf. Accessed 22 July 2021
17. Github. http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/. Accessed 15 July 2021
18. Brownlee, J.: https://machinelearningmastery.com/calculate-feature-importance-with-python/. Accessed 15 July 2021