Model for predicting heart failure using Random Forest and Logistic Regression algorithms

Vedran Grgić, Denis Mušić, Elmir Babović
"Džemal Bijedić“ University of Mostar, Faculty of Information Technologies

{vedran.grgic} {denis} {elmir.babovic}@edu.fit.ba

Abstract. The paper analyzes the cardiovascular parameters of patients with heart disease. The aim of this study was to predict death in a patient with cardiovascular disease based on 12 parameters, using Random Forest and Logistic Regression algorithms. Parameters were tuned for both algorithms to determine the best settings. The most significant factors in the process predicted were found using the FEATURE SELECTION method of both algorithms. By comparative analysis of the obtained results, the highest accuracy of 90% was obtained using the Random Forest Algorithm.

1. Introduction
Heart failure is a serious condition in which the amount of blood expelled from the heart every minute is insufficient to meet the body's normal needs for oxygen and nutrients. Heart failure can affect the right, left, or both sides of the heart. Due to the reduced ability of the heart as a pump, there may be an amount of blood in other parts of the body, and it is mistakenly believed that the term heart failure means that the heart has stopped, the term heart failure means that the heart has weakened and lost its ability to work. Heart failure occurs in all age groups but is statistically more common in the elderly with pre-existing conditions. Most often, the condition worsens with age, but with heart disease you can live undisturbed for years if you follow the recommended lifestyle habits. In the United States, about 550,000 new patients develop heart failure each year, and more than half of those who develop heart failure die within 5 years of diagnosis [18][27].

Heart failure is a global pandemic affecting at least 26 million people worldwide and is increasing in prevalence. Problems and mortality caused by heart failure will be dramatically associated with an aging population. Cardiovascular diseases are the leading cause of death in the world. More people die annually from cardiovascular diseases than from any other disease [17].

Most cardiovascular diseases can be prevented, it is necessary to monitor and eliminate risk factors. It is very important that people who have heart disease or extremely high values of the risk factors, one or more, such as hypertension, diabetes, hyperlipidaemia, etc. discover these facts in time, and undergo adequate treatment. Treatment at such an early stage can be only with drugs or only with changes in life habits (elimination of risk factors). When it comes to prevention, there are two ways to prevent cardiovascular diseases, population-wide, and individual.

Examples of comprehensive intervention against cardiovascular diseases can be seen in taxation of goods that cause illness or risk factors, tobacco, certain foods, salts, sugars, etc. Then it is necessary to animate people to healthy living habits, such as walking, playing sports. Find ways to reduce the distribution and consumption of alcohol. It is very important to implement all the above examples in
education and teach children healthy living habits and in a passive way to influence the reduction of the number of patients. At the individual level, for the prevention of first heart attacks and strokes, individual health-care interventions need to be targeted to those at high total cardiovascular risk or those with single risk factor levels above traditional thresholds, such as hypertension and hypercholesterolemia.

Efforts should be made in individual prevention and early recognition of disease in which the most significant support is expected from the application of modern technologies and algorithms. The Machine Learning area achieves the most significant results in data processing, grouping, pattern recognition and prediction making. Considering previously mentioned results, which imply some form of learning, applications of Machine Learning within medicine area is suitable for: Disease Diagnosis (key for prevention), Personalized Treatment/Behavioural Modification, Drug Discovery/Manufacturing, Clinical Trial Research, Radiology and Radiotherapy, Smart Electronic Health Records, etc.

The algorithms used within Machine Learning are most divided into two groups: supervised and unsupervised. Supervised machine learning refers to techniques in which a model is trained on a range of inputs (or features) which are associated with a known outcome. In medicine, this might represent training a model to relate a person’s characteristics to a certain outcome. Once the algorithm is successfully trained, it will be capable of making outcome predictions when applied to new data. The supervised algorithms are used to solve problems with classification and regression [5][7][9][16][20][21][24][26][30].

Model used in this paper produces discrete values, a patient died or not, so we used classification algorithms. In practice, classification algorithms return the probability of a class. This paper provides an example of a classification algorithm in which a heart failure is predicted. A model that returns a prediction of a continuous value is known as a regression algorithm.

This paper describes the process of training and validating an algorithm to predict mortality based on heart disease parameters [35], using Random Forest and Logistic Regression algorithms.

1.1. Overview of research area

Considering the importance of the area, a significant number of research addresses this topic. Emphasizing the significance of prevention, some of the previous research has focused on early diagnosis in order to prevent heart damage and lethal outcomes. Research [32] proposed model which uses weighted voting of Logistic Regression, Random Forest, K Nearest Neighbour, Gaussian Naive Bayes, and Artificial Neural Network. They used 6 different metrics (Accuracy, Precision, Sensitivity, Specificity, F Measure, and Error) to increase the prediction accuracy and precision. Authors of research [33] investigated different ML-based models to predict HF (heart failure) readmission or death. They reported improvement in performance by using the MLP (Multi-Layer Perceptron) and concluded that the proposed approach is superior to other ML and regression techniques for the prediction of 30 day HF readmission or death. In the context of our research, it is important to mention the paper [34] which used Logistic Regression and Random Forest Algorithm to classify heart disease. By comparing the accuracy of the aforementioned algorithms, they concluded that Logistic Regression generates slightly better results on used data set.

1.2. Random Forest

Random Forest is a supervised machine learning algorithm, used mostly for classification but can also be used for regression. Classification is method when we have predicting of outcome, based on input data. We can have several predictors X’s, and the one response Y. We can make prediction of death based on input parameters. Two possible model outcomes (death or life) we can present with 1 – death, 0 – life. Classification is used to understand relationship between predictors and response [6][8][19][29]. Random Forest Algorithm creates decision trees which are used for training with sample data, prediction and selection of the best solution by voting. More trees mean more robust forest.

Steps of Random Forest Algorithm [13]:

- Step 1 – selection of random samples from a given dataset.
• Step 2 – algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
• Step 3 – voting will be performed for every predicted result.
• Step 4 – select the most voted prediction result as the final prediction result

1.3. Logistic Regression
Logistic Regression is supervised machine learning algorithm or technique “borrowed” from mathematical statistics. Logistic Regression is used for binary classification problems which enables us to get the answer, will patient live or die. Logistic Regression in one of the most important analytic tools in science. In machine learning is commonly used for classification problems and it has a very close relationship with neural networks [14][25]. It provides mechanisms to train classifier that can make decision about class based on input data. For this step of classification, we use Sigmoid function and with this function we mapped real values into values between 0 and 1. With Sigmoid function, given threshold value, the algorithm determines which class the input parameter belongs to 0 or 1. Everything below threshold value is one class (0), and everything above threshold value is another class (1). For inputs (x) we have linear model, which measure relationship between dependent and independent variables and estimating probabilities by using logistic function. Using Sigmoid function, every real number is map into value between 0 and 1, but it is never exactly 0 or 1, so we have threshold classifier which classifies given number to class 0 or 1.

2. Application of algorithms
For implementation of this project, we used Python and Jupyter Notebook application [2][4][10][11][15][31]. Our dataset contains 299 rows of patients and 13 columns (features) which provide clinical, body and lifestyle data. From a logical perspective, we can split the data into categorical and continues variables.
Categorical Variables:
• Anaemia - Decrease of red blood cells - 0 corresponds to false, 1 to true
• High blood pressure - If the patient has high blood pressure - 0 corresponds to false, 1 to true
• Diabetes - If the patient has diabetes - 0 corresponds to false, 1 to true
• Sex - If 0 the patient was a woman and if 1 the patient was a man
• Smoking - If the patient smokes - 0 corresponds to false, 1 to true
• DEATH_EVENT - If the patient died during the follow-up period - 0 corresponds to false, 1 to true

Continues Variables:
• Platelets - Thrombocytes in the blood - range (25.01k - 850.00k)
• Ejection fraction - % of blood leaving the heart with each heartbeat- range (14 - 80)
• Creatinine phosphokinase (CPK) - Level of CPK enzyme in the blood - range (23 - 7861)
• Serum creatinine - Level of creatinine in the blood - range (0.5 - 9.4)
• Serum sodium - Level of sodium in the blood - range (114 - 148)
• Age - Age of the patient - range (40-95) years
• Time - Duration of the follow-up period - range (4-285) days

Feature of interest is DEATH_EVENT, we are going to predict will patient die or live, based on other features. First of all, we have to split data in test and training sets for both algorithms. We are taking 30% of data for training. Then we used sklearn pre-processing feature to remove the mean and scale to unit variance - this leads to better prediction. The sklearn is Python library which contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction.
2.1. Random Forest

Random Forest Algorithm have many attributes, but we will consider only the following: n_estimators - the number of trees in the forest., criterion {“gini”, “entropy”), max_depth - the maximum depth of the tree, min_samples_split - the minimum number of samples required to split an internal node.

After applying Random Forest Algorithm on the dataset, we have prediction accuracy of death event with 87% accuracy.

| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0         | 0.85   | 0.87     | 52      |
| 1         | 0.90   | 0.64     | 28      |

| accuracy  | 0.87   | 90       |
| macro avg | 0.88   | 0.81     | 83      | 90       |
| weighted avg | 0.87 | 0.87     | 0.86   | 90       |

**Figure 1.** Random forest accuracy with default parameters.

For Random Forest Algorithm accuracy depends on data, but also it depends on setting attributes/parameters of algorithm. For this prediction, default parameters are used, and they are represented on Figure 2.

```python
{‘bootstrap’: True,
 ‘ ccp_alpha’: 0.0,
 ‘class_weight’: None,
 ‘criterion’: ‘gini’,
 ‘max_depth’: None,
 ‘max_features’: ‘auto’,
 ‘max_leaf_nodes’: None,
 ‘max_samples’: None,
 ‘min_impurity_decrease’: 0.0,
 ‘min_impurity_split’: None,
 ‘min_samples_leaf’: 1,
 ‘min_samples_split’: 2,
 ‘min_weight_fraction_leaf’: 0.0,
 ‘n_estimators’: 100,
 ‘n_jobs’: None,
 ‘oob_score’: False,
 ‘random_state’: None,
 ‘verbose’: 0,
 ‘warm_start’: False}
```

**Figure 2.** Random Forest default parameters.

Also, we can see which features from dataset are more important than others. Figure 3 shows that smoking, blood pressure, diabetes and sex, are almost irrelevant but they are also used for prediction.
Random Forest hyperparameter tuning, is done using GridSearchCV [12][22][23][28]. That is process where we obtain the best parameters for Random Forest Algorithm. GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit model on our training set. So, in the end, we can select the best parameters from the listed hyperparameters [9].

![grid_forest.best_params_](image)

**Figure 4.** Best parameters for Random Forest.

After applying GridSearchCV parameters we have prediction accuracy of 90%.

![table](image)

**Figure 5.** Random Forest accuracy with tuned parameters.

The best score which is achieved with Random Forest Algorithm is 90% accuracy.

### 2.2. Logistic Regression

Logistic Regression has ability to show feature importance’s too. We can see which features from dataset are more important than others.
In first phase we tested Logistic Regression with default parameters, shown in the Figure 7.

\[
\begin{aligned}
\{ 'C': 1.0, \\
'\text{class_weight'}: \text{None}, \\
'\text{dual': False,} \\
'\text{fit_intercept': True,} \\
'\text{intercept_scaling': 1,} \\
'\text{l1_ratio': None,} \\
'\text{max_iter': 100,} \\
'\text{multi_class': 'auto',} \\
'\text{n_jobs': None,} \\
'\text{penalty': 'l2',} \\
'\text{random_state': None,} \\
'\text{solver': 'liblinear',} \\
'\text{tol': 0.0001,} \\
'\text{verbose': 0,} \\
'\text{warm_start': False}
\end{aligned}
\]

Figure 7. Logistic Regression default parameters.

After applying Logistic Regression algorithm with default parameters, we achieved accuracy of 81%.

|            | precision | recall | f1-score | support |
|------------|-----------|--------|----------|---------|
| 0          | 0.82      | 0.94   | 0.87     | 62      |
| 1          | 0.79      | 0.54   | 0.64     | 28      |

accuracy 0.81
macro avg 0.80 0.74 0.75 90
weighted avg 0.81 0.81 0.80 90

Figure 8. Logistic Regression accuracy with default parameters.

For Logistic Regression we also used hyperparameter tuning by GridSearchCV library. After applying GridSearchCV, we found the best parameters for Logistic Regression showed on Figure 9.

Best parameters) \{ 'C': 1, 'multi_class': 'ovr', 'penalty': 'l1', 'solver': 'liblinear' \}

Figure 9. Best parameters for Logistic Regression
After applying best parameter for Logistic Regression algorithm, we got prediction accuracy of 82%.

| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0         | 0.82   | 0.95     | 0.88    | 62      |
| 1         | 0.83   | 0.54     | 0.65    | 28      |

accuracy: 0.82, 90
macro avg: 0.83, 0.74, 0.77, 90
weighted avg: 0.82, 0.82, 0.81, 90

Figure 10. Logistic Regression accuracy with tuned parameters.

This can’t be considered as significant improvement, but it is worth emphasizing.

3. Comparative Analysis

After finishing project with both algorithms, we can compare them by accuracy in prediction, and using AUC (Area Under the Curve) and ROC (Receiver Operator Characteristic Curve) graph. Regarding accuracy in prediction, Random Forest have accuracy of 90%, Logistic Regression 82% with tuned parameters. With default parameters, Random Forest have accuracy 86%, and Logistic Regression have 81% accuracy. The results with tuned parameters are observed below in ROC and AUC graph. ROC graph provides support in deciding the best threshold value of algorithm. ROC graph consists of x-axis which represents False Positive Rate, and y-axis which represents True Positive Rate, and the curve is generated by changing the threshold on the confidence score. AUC and ROC are important evaluation metrics for calculating the performance of any classification model’s performance.

![ROC Graph](image)

Figure 11. ROC graph.

Green diagonal line is used to show when is proportion of correctly classified samples is equal to the proportion of incorrectly classified samples. The area under the ROC curve - AUC is a measure which determine efficiency of a classifier. The AUC makes it easy to compare the ROC curve of one model to another. Higher AUC the better [1].
After the analysis, all the above leads us to a conclusion that the Random Forest is better option for heart failure prediction. Graph (Figure 11) shows that Random Forest line is closer to 1 and have a greater AUC on this dataset.

4. Conclusion
The use of Machine Learning in medicine is gaining in importance for the last decade. There are many ways to use machine learning algorithms in medicine and outcome prediction as one of the most used. Random Forest and Logistic Regression algorithms can help predict deaths in heart patients, and the give very good prediction results. But Random Forest Algorithm has better accuracy for default and tuned parameters and have a better AUC result. The main disadvantage of this model is the complexity, in this case, Random Forest Algorithm creates 100 trees in Python sklearn library. To do so, this algorithm requires much computational power and resources. We analysed small dataset which can not affect algorithm performance in noticeable level. To have an even more precise model, it is necessary to increase the amount of data, so not only to get a more reliable model, but also to get more accurate parameters when doing hyperparameter tuning.

References
[1] Pandey P 2020 Understanding the roc and auc metrics towards data science [cited 2020Nov11] https://towardsdatascience.com/understanding-the-roc-and-auc-curves-a05b68550b69
[2] Li S 2019 Building a logistic regression in python step by step towards data science [cited 2020Dec14] https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-beced4d56c9e8
[3] Donges N 2021 The logistic regression algorithm Experfy Insights [cited 2021May18] https://www.experfy.com/blog/bigdata-cloud/the-logistic-regression-algorithm/
[4] Brownlee J 2020 How to calculate feature importance with python Machine Learning Mastery [cited 2020Dec4] https://machinelearningmastery.com/calculate-feature-importance-with-python
[5] Overfitting in machine learning: What it is and how to prevent it 2020 Elite Data Science [cited 2020Dec5] https://elitedatascience.com/overfitting-in-machine-learning
[6] Kreiger J R 2020 Evaluating a random forest model Analytics Vidhya [cited 2020Dec4] https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56
[7] Liyanapathirana L 2018 Machine learning workflow on diabetes data Part 02 Towards Data Science [cited 2020Dec2] https://towardsdatascience.com/machine-learning-workflow-on-diabetes-data-part-02-11262b77a5c

[8] Fraj M 2017 In depth: parameter tuning for random forest all things AI [cited 2020Nov23] https://medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d

[9] Borkowski A 2020 Predicting heart failure using machine learning, part 2 Analytics Vidhya [cited 2020Oct30] https://medium.com/analytics-vidhya/predicting-heart-failure-using-machine-learning-part-2-b343471dbde8

[10] How to optimize hyper parameters of a logistic regression model using grid search in python DeZyre [cited 2020Nov2] https://www.dezyre.com/recipes/optimize-hyper-parameters-of-logistic-regression-model-using-grid-search-in-python

[11] Dwivedi R 2020 Guide to hyperparameters tuning using GRIDSEARCHCV and RANDOMIZEDSEARCHCV Analytics India Magazine [cited 2020Dec1] https://analyticsindiamag.com/guide-to-hyperparameters-tuning-using-gridsearchcv-and-randomizedsearchcv

[12] Classification algorithms - random forest Tutorialspoint [cited 2020Dec19] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm

[13] Cancer linear regression model tutorial 2017 Dataset by NRIPPNER Data World [cited 2020Oct17] https://data.world/nrippner/cancer-linear-regression-model-tutorial

[14] Stojiljkovic M 2019 Linear regression in python Real python [cited 2020Nov19] https://realpython.com/linear-regression-in-python/#implementing-linear-regression-in-python

[15] Harrison O 2018 Machine learning basics with the K-nearest neighbors algorithm Towards Data Science [cited 2020Dec29] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[16] Cardiovascular diseases World Health Organization [cited 2020Dec9] https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[17] Heart and vascular conditions and treatments Heart Failure Statistics [cited 2020Dec30] https://www.emoryhealthcare.org/heart-vascular/wellness/heart-failure-statistics.html

[18] Probst P, Wright M N and Boulesteix A L 2019 Hyperparameters and tuning strategies for random forest Data Mining and Knowledge Discovery. doi: 9. 10.1002/widm.1301

[19] Babić D 2017 Postupak optimiziranog izbora oglasa na temelju stabala odlučivanja Master thesis (Ostijek: Josip Juraj Strossmayer University, Faculty of Electrical Engineering, Computer Science and Information Technology)

[20] Gregorić M 2019 Strojno učenje kao alat za zaključivanje Undergraduate thesis (Zagreb: University of Zagreb, Faculty of Humanities and Social Sciences)

[21] Probst P, Boulesteix A L and Bischl B 2019 Tunability: Importance of hyperparameters of machine learning algorithms, Journal of Machine Learning Research 20 pp 1-32

[22] Rijn J, Hutter F 2018 Hyperparameter Importance Across Datasets (NY: Association for Computing Machinery) 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining pp 2367–2376

[23] Jovanović S, Bajić D, Jovanović M and Milovanović B 2019 Primena tehnika mašinskog učenja u klasifikaciji pacijenata posle infarkta miokarda 18th International Symposium INFOTEH-JAHORINA

[24] Prasad R, Anjali P, Adil S and Deepa N 2019 Heart disease prediction using logistic regression algorithm using machine learning International Journal of Engineering and Advanced Technology 8 3S

[25] Tanvir A, Assia M, Bhatti S H, Aftab M, Raza M 2017 Survival analysis of heart failure patients: A case study PLoS ONE 12(7) e0181001. doi:10.1371/journal.pone.0181001

[26] Sidey G J and Sidey G C 2019 Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19 64 https://doi.org/10.1186/s12874-019-0681-4
[27] Selthofer R K and Drenjančević I 2014, Čimbenici rizika za kronično zatajivanje srca, Review article Medix 112

[28] Schratza P, Muenchowa J, Iturritxab E, Richterc J and Brenninga A 2018 Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data Ecological Modelling 406 pp 109-120 arXiv:1803.11266v1

[29] Louppe G 2015 Understanding random forests from theory to practice PhD thesis Machine Learning (stat.ML) arXiv:1407.7502

[30] Hrastić D 2019 Implementacija klasiifikacijskog algoritma podizanja gradijenta stablima odluke s primjenom na predikciju ishoda nogometnih utakmica Undergraduate thesis (Zagreb: University of Zagreb, Faculty of Electrical Engineering and Computer Science) https://www.bib.irb.hr/1009326/download/1009326.Dominik_Hrastic_Zavrsni_final.pdf

[31] Tokic L 2017 ROC krivulja – krivulja odnosa specifičnosti i osjetljivosti klasiifikatora Master thesis (Zagreb: University of Zagreb, Faculty of Science) https://repozitorij.pmf.unizg.hr/islandora/object/pmf%3A4490/datastream/PDF/view

[32] Chowdary G J, Suganya G and Premalatha M 2020 Effective prediction of cardiovascular disease using cluster of machine learning algorithms Journal of Critical Reviews 7 19 pp 1865-1875. doi:10.31838/jcr.07.19.227

[33] Awan SE, Bennamoun M, Sohel F, Sanfilippo F M and Dwivedi G 2019 Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics ESC Heart Failure 6 doi: 10.1002/ehf2.12419.

[34] Latifah F and Slamet I S 2020 Comparison of heart disease classification with logistic regression algorithm and random forest algorithm AIP Conference Proceedings 2296 020021 doi: 10.1063/5.0030579

[35] Chicco D and Jurman G 2020 Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone BMC Medical Informatics and Decision Making 20 16