Random Forest as a Method of Predicting the Presence of Cardiovasculars Diseases

Yurii Kryvenchuk, Alina Yamniuk, Iryna Protsyk, Lesia Sai, Andriana Mazur, Olena Sydorchuk

Lviv Polytechnic National University, Stepana Bandery Street 12, Lviv, 79013, Ukraine

Abstract

Heart diseases are considered to be one of the main reasons that lead to the death. The correct prediction of heart disease can prevent life threats, however incorrect prediction can be fatal at the same time.

The paper considers the ways to solve the problem of prediction cardiovasculars diseases. For this purpose, Random Forest method is considered. Both the advantages and disadvantages of using this method of predicting are investigated. The main problems that shape the work are defined. The paper provides step-by-step creation of a system and identifies the main requirements that it must meet. For training the model it is used classifier RandomForestClassifier. The results of the training are given as well. For improving the precision of the training model Grid Search is used. The metrics – error matrices and ROC-AUC curves are used in order to visualize the results of the research. To compare the results Gradient boosting algorithm is considered.

Such a model can be very useful in hospitals as an additional check of the diagnosis before prescribing treatment.

The object of the research is medical indicators and their importance for successful diagnosis of the disease. The data is taken from the open source. The subject of the research is the method of Random Forest for classification based on statistical data.

Keywords 1

Machine Learning, Classification, Random Forest, Decision Tree, Begging, Boosting, Diagnosis of Disease.

1. Introduction

Medicine is thought to be one of the most investigated areas for all the time. Due to the development of the information and computer technologies, this area and diagnostic process are rapidly being modernized [1]. Early detection of the disease in humans is an extremely important procedure, as it can prevent serious consequences. And the success or failure of treatment directly depends on timely and accurate diagnosis.

The scope of Machine Learning algorithms are increasing in predicting various diseases. Machine learning algorithms are often used for cardio disease prediction systems. Machine learning techniques help in identifying the data and automatically make the predictions.

Integrating Machine Learning into the healthcare ecosystem allows for a multitude of benefits, including automating tasks and analyzing big patient data sets to deliver better healthcare faster and at

COLINS -2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12-13, 2022, Gliwice, Poland
EMAIL: yurkokryvenchuk@gmail.com (Yu. Kryvenchuk); alinayamniuk@gmail.com (A. Yamniuk); iryna.s.protsyk@lpnu.ua (I.Protsyk); lesia.p.sai@lpnu.ua (L.Sai); andriana.v.mazur@lpnu.ua (A.Mazur); sallasky2009@gmail.com (O.Sydorchuk)
ORCID: 0000-0002-2504-5833 (Yu. Kryvenchuk); 0000-0003-1886-5902 (A. Yamniuk); 0000-0002-6270-1344 (I. Protsyk); 0000-0002-5081-4235 (L.Sai); 0000-0002-5985-5674 (A.Mazur); 0000-0002-9357-8690 (O.Sydorchuk)

a lower cost. Quickly obtaining patient insights helps the healthcare ecosystem discover key areas of patient care that require improvement.

There is a growing awareness of the importance of machine learning as a platform that can gather information from multiple sources into an integrated system. It significantly facilitates decision-making processes for highly skilled employees. Improvements in computing resources, as well as the storage and exchange of data over the last decade, have been a significant factor in harnessing the potential of machine learning systems in medicine.

Machine learning is a fast-growing trend in the health care industry, thanks to the advent of wearable devices and sensors that can use data to assess a patient's health in real time. The technology can also help medical experts analyse data to identify trends or red flags that may lead to improved diagnoses and treatment.

According to the World Health Organization, as of 2020, the leading cause of deaths is heart diseases. This disease is responsible for 16% of the world's total deaths. Since 2000, the largest increase in deaths has been for this disease, rising by more than 2 million to 8.9 million deaths in 2020. Also the number of deaths will add up by 24.5 million in 2030, because of the growth of cardiovascular risk factors such as high blood pressure, diabetes, obesity, and smoking [2].

Machine learning algorithms can help doctors diagnose, analyse X-rays, predict patient's health and so on. The exact and accurate analysis is normally attributed to the successful treatment. When doctors fail to make accurate decisions while examining a patient's disease, disease prediction systems that use ML algorithms can help.

An important aspect that distinguishes medical data from most others is that objectivity, accuracy, quality and timeliness of results are critical and must be constantly questioned. Thus, the problem of medical diagnosis can be solved with the help of the classification problem.

One of the classification methods is Random Forest [3]. It is used to create a classifier model that can predict disease with higher rates and accuracy.

By using classification methods, it is possible not only to cure people, but also to prevent the deterioration of their health in time. Appropriate and accurate prediction of cardiovascular disease has quite significant value.

The cost of error is quite high, so research in this area is always needed. This research proposes a prediction model to predict whether patient have a heart disease or not by using entered indicators and to give an awareness of heart disease.

An article primarily includes the following sections:

1. Introduction – provided actual problem, clear motivation of doing the research and possible improvements for this area.
2. Related work - introduced common works and theirs results, related to the theme.
3. Materials and methods – considered methods of the experiment, Random Forest, Decision Tree, Bagging or Bootstrap Aggregating, Gradient Boosting.
4. Experiments – investigated the data using Random Forest in order to predict the presence of cardiovascular disease for people with different health conditions.
5. Results – represented analysed data in figures and results of the study.
6. Conclusion – summarized overall findings.
7. References – source materials.

2. Related Work

To do the research on this specific topic, it is quite important to gain the results of the works that have been done earlier. For this it is necessary to comprehensively analyse the literary different type of sources.

Many research articles have been carefully studied in order to investigate the problem and systematize knowledge in this area.

In article [4] was proposed a model for prediction of cardiovascular disease using machine learning algorithm hybrid random forest with linear mode. Authors obtained 88.7% accuracy for prediction. The

dataset was collected from UCI repository site. Authors have chosen Cleveland dataset for this proposed study.

In article [5] authors used knn, decision tree, linear regression, support vector machine algorithms for prediction of heart disease and compared their accuracy. From the experimental result authors obtained best accuracy of 87% by using k-nearest neighbor algorithm followed by support vector machine 83%, decision tree 79% and linear regression of 78% accuracy among all these algorithms for prediction of heart disease.

In article [6] a study conducted to compare statistical, ML and data mining methods in terms of their ability to assist in predicting heart failure risks. The researchers compared the performance of statistical evaluation, Decision Trees, Random Forest, and convolutional neural networks, and they obtained prediction accuracy results of 85%, 80.1%, 85.38%, and 93%, respectively.

In article [7] was used different varieties of unsupervised clustering algorithms to determine their accuracy in terms of cardiac disease search and diagnosis. The algorithms were applied to the Cleveland dataset. The study results highlighted k-means as the most appropriate algorithm for cardiac disease diagnosis.

In article [8] was suggested model using ensemble approaches (boosting and bagging) with feature extraction algorithms (LDA and PCA) for predicting heart disease. The authors compared ensemble techniques (bagging and boosting) with five classifiers (SVM, KNN, RF, NB, and DT) on selected features from the Cleveland heart disease dataset. The results of the experiments indicated that the bagging ensemble learning method with DT and PCA feature extraction obtained the most outstanding performance.

3. Materials and methods

Random Forest is an ensemble machine learning method based on decision trees that involves creating multiple trees and then combining their results to improve model generalization capabilities.

Main features of Random Forest Algorithm:

- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of overfitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

Decision trees are the building blocks of a Random Forest algorithm. Decision trees [9] are a decision-making tool that uses a tree-like graph or decision model and their possible consequences. Decision trees seek to find the best distribution for a subset of data, and they are usually learned using a classification tree algorithm.

Decision trees are very easy to overfit. In the process of building a tree, so that its size does not become too large, is used special procedures that allow you to create optimal trees. The creation process continues until the stop criteria are met.

The most well-known measures of entropy and Gini index are used in the work [10].

During the study, the best criterion was chosen - the measure of entropy [11].

The measure of entropy in the construction of decision trees is a measure of the diversity of classes in the node. As a result of breakdown nodes with smaller variety of states of an initial variable are formed. Thus, the entropy decreases and the amount of internal information in the node increases. Formally, the entropy of a certain node T of the decision tree is determined by the formula:

$$Info(T) = -\sum_{j=1}^{n} p_j \cdot \log(p_j), \quad p_j = \frac{N_j}{N} \text{ – part of class j in the node T}$$

p – the probability of system being in the state i,
T – current node,
n – number of classes,

N – number of objects in the node

The entropy of the whole breakdown is the sum of the entropies of all the nodes multiplied by possibility of records of all nodes in the total number of records:

$$Info(S) = \sum_{j=1}^{n} \frac{N_j}{N} \cdot Info(T_j)$$

To select the split attribute, a criterion called information gain or entropy decrease is used:

$$Gain(S) = Info(T) - Info_s(T)$$

The best attribute to be used in the S breakdown is the one that provides the largest increase in Gain (S) information.

To reduce deviations, one of the main methods of aggregation in machine learning is used - Bagging or Bootstrap Aggregating [12].

Begging is a simple technique in which we build independent models and combine them using some model of averaging. Because, classification is carried out, aggregation occurs by majority voting. For this test observation, we can record the class predicted by each of the trees and take the majority of votes: the overall forecast is the most common class.

Begging has been particularly useful for decision trees. This is because begging avoids the high correlation between decision trees that occurs when training them using the same data.

One of the most popular boosting algorithms Gradient Boosting is used in the research [13]. Like begging, the main task of boosting is to transform a set of weak classifiers (that is, those that make many mistakes in the test sample) into a stronger one. Gradient Boosting works consistently, adding new ones to past models to correct mistakes made by previous predictors. This algorithm tries to teach new models on the residual error of the past (moving to a minimum loss function).

Let's evaluate the use of Random Forest and Gradient Boosting.

Like a Random Forest, Boosting Trees are a set of Decision Trees. The main differences between the algorithms are the way the trees are built and the results combined.

How trees are built: random forests build each tree independently; Gradient Boosting creates one tree at a time. This additive model (ensemble) works in stages, representing a weak student to improve the shortcomings of existing weak students.

Combining results: random forests combine results at the end of the process (by averaging or "majority rules"), while Gradient Boosting combines results along the way.

If you adjust the parameters carefully, Gradient Boosting can lead to better performance than random forests. However, Gradient Boosting may not be the best choice if noise is present, as it can lead to overfitting.

4. Experiments

For the practical part of this work, it was decided to investigate a dataset that contains data about patients and whether they have been diagnosed with cardiovascular disease. So that, the problem of classification will be solved.

This dataset consists of 70,000 patient records, which include (Figure 1):

1. age;
2. height;
3. weight;
4. gender;
5. blood pressure;
6. cholesterol;

7. blood glucose;
8. whether the patient smokes;
9. whether the patient drinks alcohol;
10. whether the patient is physically active.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Figure 1. Medical indicators

The first stage is to check the dataset for duplicate rows. Rejection of duplicates is necessary, because during training, the model will learn from the original data, and then re-study their duplicate. Therefore, the model will relearn the same sample of data. As a result, the model may be poorly generalized.

The next stage is to check the relationships between the target variable and other variables.

Figure 2. shows the number of patients who were diagnosed (yellow column) and were not diagnosed (green column) cardiovascular disease relative to their age (in years).

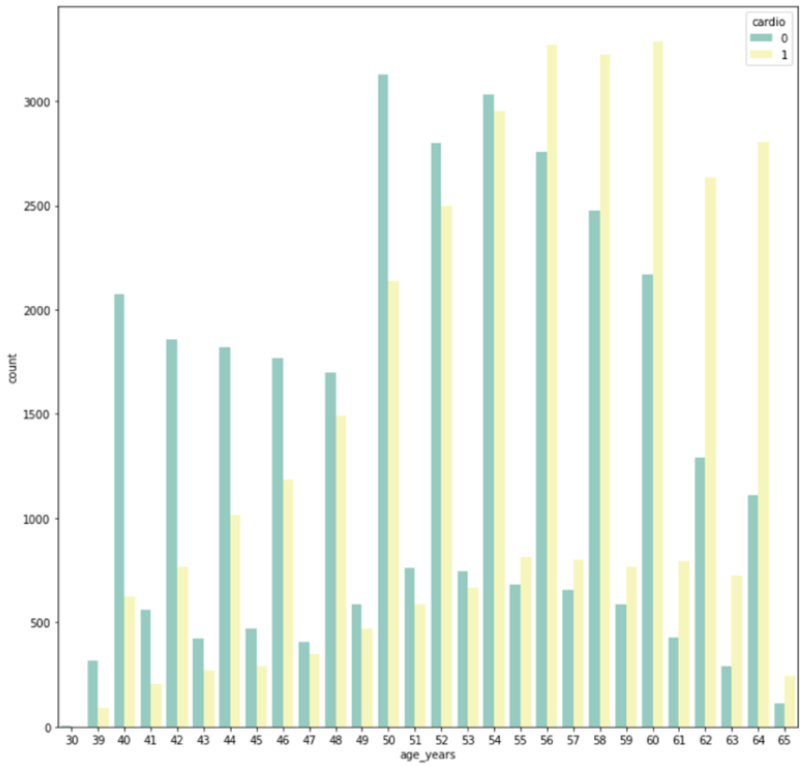

Figure 2: Dependence of cardiovascular disease on age

As a result, the ratio of patients to healthy patients increases with age.

The next stage is to check whether there are linear relationships between variables. To do this, we construct a correlation matrix that will contain the correlation values for all pairs of variables. If there is a high correlation between the variable and the target variable, it will be possible to find a linear relationship between these variables. If non-target variables have high correlation values, it means that they contain very similar information, and therefore one of these variables can be neglected, and thus reduce the complexity of the model.

In the Figure 3. the correlation matrix for the dataset is shown.

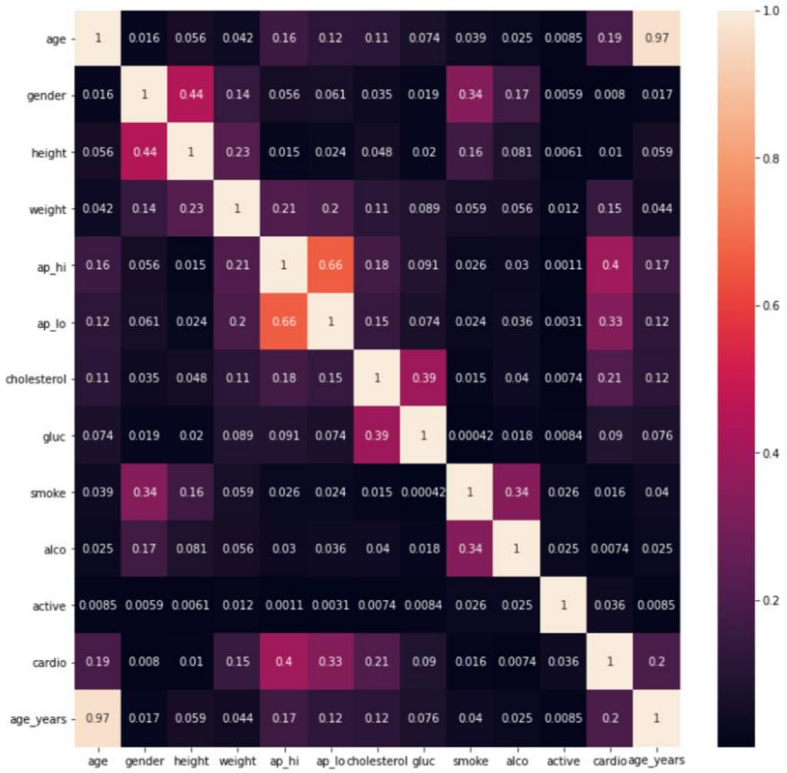

Figure 3: Correlation matrix.

There is no direct linear relationship between variables. However, it can be noted that the highest values of correlation relative to the target variable (cardio) have indicators that correspond to blood pressure (ap_hi, ap_lo).

The next stage is to analyse the data in order to find possible emissions (not typical data, extreme). To do this, statistic data about the columns in the dataset are displayed. To be more precise - the mean value, standard deviation, minimum and maximum values and quantiles for each indicator.

From this data, it can be analysed that blood pressure indicators (both ap_hi and ap_lo) have extreme points because the maximum value is very different, which means that the mean and median values are different as well.

To see the distribution of data visually, the box charts for this data are built. (Figure 4).

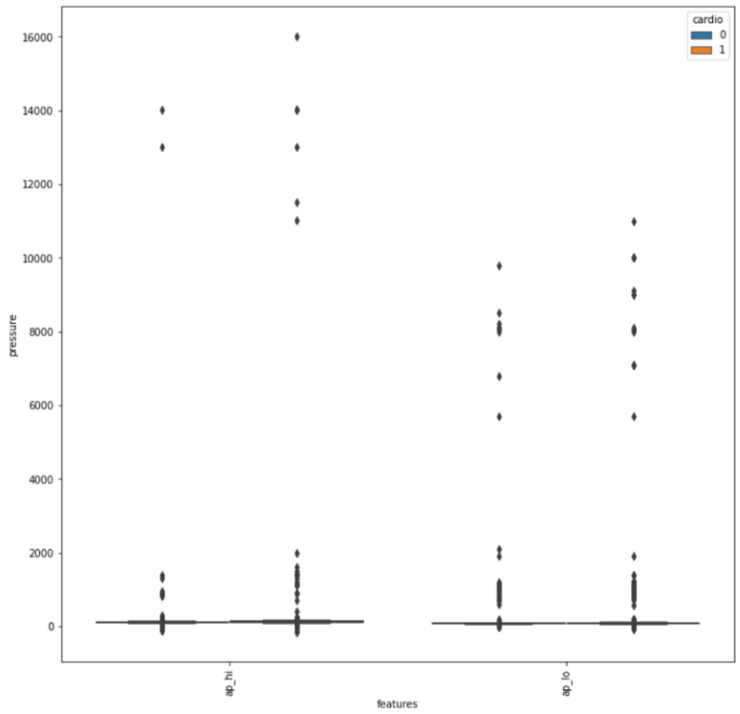

Figure 4: Box charts.

As a result, these indicators have extreme data, which are indicated on the graph by points that go beyond the interquartile range.

The next step is to delete the lines that contain the extreme points. On the graph you can see the emissions, which are indicated by points beyond the quarterly interval. According to the research, emissions are non-standard indicators of a patient's health. To minimize the risks of incorrect prediction, it is necessary to clean the dataset from emissions.

The categorical data is encoded with one hot encoding. Such indicators in dataset 2: "cholesterol" and "gluc". Each of them has three unique values, so after unary coding, each of them will turn into three different indicators with boolean values.

4.1. Training Random Forest Model

Before training, the dataset need to be divided into a train set and a test set. A ratio of 80% - train, 20% - test.

The RandomForestClassifier class from the scikit-learn library is chosen to train the random forest model. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.

Advantages of using the Random Forest Classifier [14]:

1. The random forest algorithm is significantly more accurate than most of the non-linear classifiers.
2. Ability to efficiently process data with a large number of attributes and classes.
3. The random forest classifier doesn’t face the overfitting issue because it takes the average of all predictions, canceling out the biases and thus, fixing the overfitting problem.
4. You can use this algorithm for both regression and classification problems, making it a highly versatile algorithm.
5. Both continuous and discrete features are treated equally well. There are methods of constructing trees according to data with omitted values of features.
6. This algorithm offers you relative feature importance that allows you to select the most contributing features for your classifier easily.

7. Ability to work in parallel in many threads.
8. Internal assessment of the model's ability to generalize (out-of-bag test).
9. Scalability.

Disadvantages of using Random Forest Classifier [14]:

1. This algorithm is slower than other classification algorithms because it uses multiple decision trees to make predictions. When a random forest classifier makes a prediction, every tree in the forest has to make a prediction for the same input and vote on the same. This process can be very time-consuming.
2. The algorithm tends to relearn on some tasks, especially with a lot of noise.
3. Large size of the received models. Requires O (NK) memory to store the model, where K is the number of trees.

In order to improve the accuracy of the model, the search for optimal Grid Search hyperparameters is used. It sorts the combinations from the given hyperparameters and chooses the best combination.

Hyperparameters that are used in the search:

- n_estimators - the number of trees used to build a random forest;
- criterion - the criterion of breaking a tree;
- max_depth - the maximum depth that can be reached by the tree, after which the construction stops;

As a result, after searching for hyperparameters, the model that showed the best predictions on a given metric will be returned.

5. Work results

Since the problem of binary classification has been solving, the most objective metric is roc-auc (area under the error curve) [15]. The ROC curve provides detailed information about the behavior of the classifier. The curve is the result of the True Positive Rate (TPR) and False Positive Rate (FPR) depending on the threshold.

As a result of Grid Search, the best combination of hyperparameters is:

1. the number of trees - 1000
2. the maximum depth of the tree – 10

To see the number of true and false predictions of the classifier, the error matrices for the train and test sets, accordingly are shown on the plot (Figure 5, Figure 6).

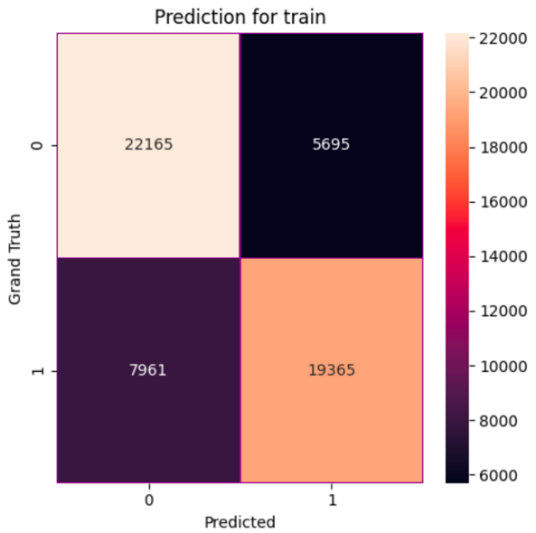

Figure 5: Error matrix for the train set

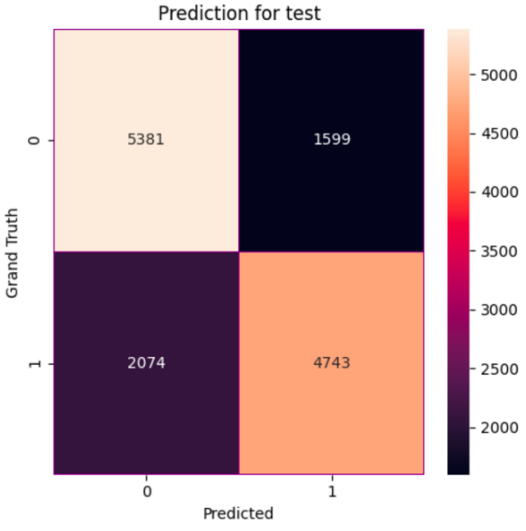

Figure 6: Error matrix for the test set

As a result, the relative distribution of errors has been preserved from the train to test sets, and therefore the model is well generated.

The model makes fewer FN errors (when the correct prediction is 1 and the model gives 0). False Negative errors - the algorithm did not recognize the disease and recognized the sick person healthy. The cost of error is very important, especially in medicine.

Obviously, ideally, we aim for the classification algorithm is to give zero errors of the FP and FN classes, but in real life this is rare, but each model should minimize the number of errors.

In the Figure 7. and Figure 8. the error curves (ROC) and the calculated area under them (roc-auc) are shown for the training and testing set, accordingly.

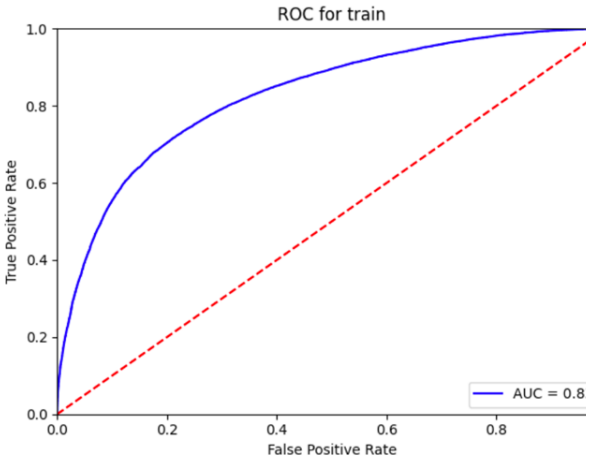

Figure 7: ROC-AUC for training dataset

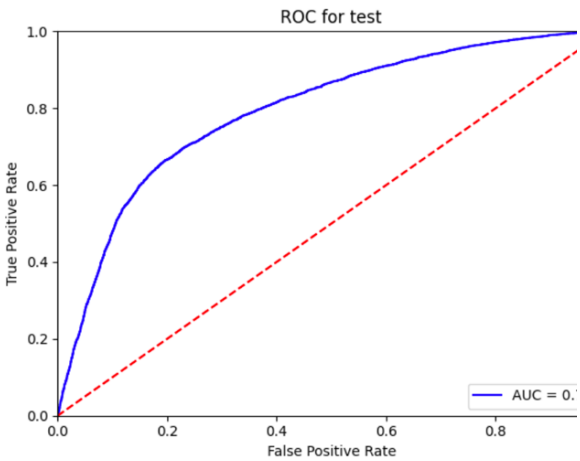

Figure 8: ROC-AUC for testing dataset

The ROC curve [16] illustrates the sensitivity of the classifier, showing how many correctly classified objects can be obtained, allowing more and more FP cases.

This metric shows the dependence of the fullness (recall) of predictions - the proportion of class 1 objects from all class 1 objects, that were correctly predicted, to the proportion of class 0 objects that were incorrectly predicted. The metric has possible values [0; 1]. As a result, the metric values are similar on the training and testing datasets, so the model is well generated.

1. It reduces overfitting in decision trees and helps to increase accuracy.
2. It is flexible for both classification and regression problems.
3. It works well with both categorical and continuous values.
4. Data normalization is not required because a rule-based approach is used.

6. Conclusions

To summarize the research, it should be said the goal of the study has successfully been achieved. Namely, with the help of one of the most popular methods - the method of Random Forest to predict the presence of cardiovascular disease in people with different health conditions.
The research was conducted on the basis of an open dataset cardio.csv, taken from the Internet.

Using the best classification model obtained through Grid Search from the scikit-learn library, the Random Forest training was committed. Then the trained model was used to classify patients. The accuracy of predictions is 80%. The same operations were performed for another ensemble algorithm - Gradient Boosting. When using this algorithm, the correct predictions were 73%.

The AUC metric for the test data showed a high prediction score of 0.79. So, the classifier worked quite well.

7. References

1. Croft, P., Altman, D. G., Deeks, J. J., et. al. The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice. BMC Medicine. 2015. Vol. 13, No. 1. P. 8. DOI: 10.1186/s12916-014-0265-4
2. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
3. Random Forest (2021) *https://en.wikipedia.org/wiki/Random_forest*
4. Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). "Heart disease prediction using data mining techniques". In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE
5. Singh, A., & Kumar, R. (2020). "Heart Disease Prediction Using Machine Learning Algorithms". 2020 International Conference on Electrical and Electronics Engineering (ICE3). doi:10.1109/ice348803.2020.9122958
6. Tiwaskar, S.A.; Gosavi, R.; Dubey, R.; Jadhav, S.; Iyer, K. Comparison of Prediction Models for Heart Failure Risk: A Clinical Perspective. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018.
7. Kodati, S.; Vivekanandam, R.; Ravi, G. Comparative analysis of clustering algorithms with heart disease datasets using data mining weka tool. Adv. Intell. Syst. Comput. 2019, 900, 111–117. [CrossRef]
8. X.-Y. Gao, A. A. Ali, H. S. Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, Article ID 6663455, 10 pages, 2021.
9. Kingsford, C., Salzberg, S. What are decision trees?. *Nat Biotechnol* 26, 1011–1013 (2008). https://doi.org/10.1038/nbt0908-1011
10. Dobashi, N.; Saito, S.; Nakahara, Y.; Matsushima, T. Meta-Tree Random Forest: Probabilistic Data-Generative Model and Bayes Optimal Prediction. *Entropy* 2021, *23*, 768. https://doi.org/10.3390/e23060768
11. Akhil Kadiyala, Ashok Kumar, Applications of python to evaluate the performance of bagging methods, 2018, https://doi.org/10.1002/ep.13018
12. Biau, G., Cadre, B. & Rouvière, L. Accelerated gradient boosting. *Mach Learn* 108, 971–992 (2019). https://doi.org/10.1007/s10994-019-05787-1
13. Polat K.A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems/ Polat K., Gunes S., – An International Journal of Expert Systemswith Applications. – 2016. – vol. 36(2). – pp.587-1592
14. Pavan Vadapalli (June 18, 2021). Random Forest Classifier: Overview, How Does it Work, Pros & Cons. https://www.upgrad.com/blog/random-forest-classifier/
15. Pawlak Z. Rough Sets / Pawlak Z. // International Journal of Computer and Information Science. – 1982. – No11. – pp. 341-356
16. Sarang Narkhede, Understanding AUC – ROC Curve, Jun 2018, https://medium.com/@narkhedesarang
17. Agajanian, S., Odeyemi, O., Bischoff, N., Ratra, S., and Verkhivker, G. M. (2018). Machine learning classification and structure-functional analysis of cancer mutations reveal unique dynamic and network signatures of driver sites in oncogenes and tumor suppressor genes. *J. Chem. Inf. Model.* 58, 2131–2150. doi: 10.1021/acs.jcim.8b00414
18. Desai, S. Chaudhary Distributed decision tree v.2.0, in: 2017 IEEE International Conference on Big Data (Big Data**)** Presented at the 2017 IEEE International Conference on Big Data (Big Data) (2017), pp. 929-934
19. Strobl, C., Boulesteix, A. L., and Augustin, T. (2017). Unbiased split selection for classification trees based on the gini index. *Comput. Stat. Data Anal.* 52, 483–501. doi: 10.1016/j.csda.2006.12.030
20. Bashir S. IntelliHealth: A medical decision support application using a novel weighted multi-

layer classifier ensemble framework / Bashir S., Qamar U., Khan F.H. // Journal of Biomedical Informatics. – 2016. – vol.59. – pp.185-200
21. Obuchowski NA, Bullen JA. Receiver operating characteristic (roc) curves: review of methods with applications in diagnostic medicine. Phys Med Biol. 2018;63(7):07–1.
22. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”. IEEE Access, 1–1. doi:10.1109/access.2019.2923707
23. Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining."In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5. IEEE, 2017.
24. Gnaneswar, B., and MR Ebenezar Jebarani. "A review on prediction and diagnosis of heart failure." In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-3. IEEE, 2017.
25. Shah, D., Patel, S., Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science, 1 (6). doi: https://doi.org/10.1007/s42979-020-00365-y
26. Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.
27. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.
28. Chaurasia V, Pal S. Data mining approach to detect heart diseases. Int J Adv Comput Sci Inf Technol (IJACSIT). 2014;2:56–66.