IRSTI 28.23.15 https://doi.org/10.26577/JPCSIT.2023.v1.i2.06

R.K. Imanbek*, Z.A. Buribayev, A. Yerkos

Al-Farabi Kazakh National University, Kazakhstan, Almaty
*e-mail: imanbek.rustem2000@gmail.com

PROCESSING OF ISCHEMIC HEART DISEASE DATA USING ENSEMBLE CLASSIFICATION METHODS OF MACHINE LEARNING

Abstract. The WHO 2019 statistics provide evidence that cardiovascular diseases are among the prevailing causes of death globally [1]. In this study, a combined dataset of coronary artery disease (CAD), also known as ischemic heart disease, was used as the dataset for analysis. To influence the outcome of the occurrence of cardiovascular diseases, it is important to find significant features that contribute to the presence of this disease. This article demonstrated that important features can be obtained through classification and their visualization in Tableau. Three classification models were built, and important features were identified for each model. Then, the top 10 important features were selected from each model, and through comparison, the 5 most important features were identified that may influence the disease outcome. The classification models achieved the following f1-score results: LGBM (93.2%), XGB (92.0%), and RF (89.1%).

Key words: Ischemic heart disease, Extreme Gradient Boosting (XGB), Random Forest (RF), Light Gradient Boosting Machine (LGBM), Machine Learning (ML), Tableau, important features.

Introduction

According to the statistics provided by the WHO in 2019, cardiovascular diseases stand out as a leading global cause of death [1]. This includes related diseases such ischemic heart disease, arterial hypertension, arrhythmia, myocardial infarction, stroke, and many others. In Kazakhstan, more than 2 million people are registered as suffering from cardiovascular diseases, and over 40,000 Kazakhstan citizens die from cardiovascular diseases annually. This is also supported by the WHO 2019 data, see Figure 1.

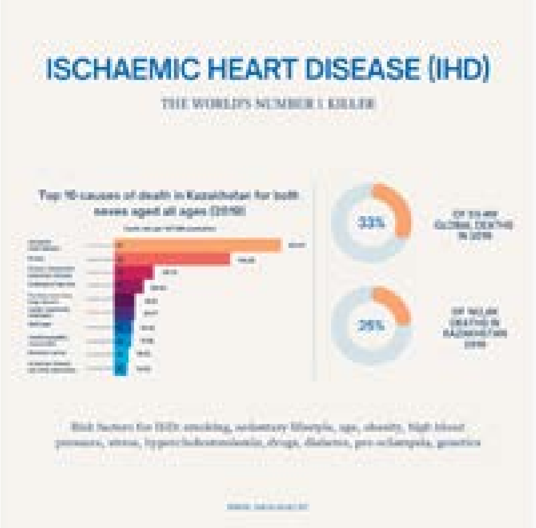


Figure 1 – Statistics from WHO 2019 on ischemic heart disease worldwide and in Kazakhstan

Prediction of outcomes in cardiovascular diseases is currently a highly relevant task in bioinformatics [2]. The influence on the outcome of this disease using various machine-learning algorithms has been extensively studied in scientific literature. For instance, in [3], a review of 49 papers was conducted, where algorithms such as Random Forest, k-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGB), Logistic Regression, Decision Tree, and others were employed. Another study [4] reviewed 58 papers with a focus on Artificial Neural Network (ANN), and [5] covered 63 papers primarily discussing Convolutional Neural Network (CNN), providing evidence for this statement.

In this study, we utilized three machine learning classification algorithms for the classification task: RF, XGB, and LGBM. This selection was based on the popularity of these algorithms. Specifically, in [6], XGB was applied to a similar task, achieving model performance metrics such as an accuracy of 91% and an f1-score of 90%. Similar cases were observed with RF [7] with an accuracy of 85.81%, and LGBM [8] with metrics: an accuracy of 84.48%.

By applying analytics and the Tableau platform, the dataset can be visualized, and by separating it using important features in a 2D space, one can verify the significance of these features [9].

Description of the combined dataset on cardiovascular diseases.

In this article, we used a publicly available combined dataset on cardiovascular diseases [10]. The dataset consists of 1190 patient records and 12 features. The target values are represented in the dataset as follows: 629 for patients with ischemia and 561 for healthy patients. In this case, class balancing was unnecessary in this particular situation. Within the dataset, two features had missing values: 'oldpeak' (14% missing) and 'cholesterol' (38% missing), which were imputed using KNNImputer. Three of nominal features ('chest pain type', 'resting ecg', and 'ST slope') had named values, because they have more than 2 unique values, we employed OneHotEncoder, resulting in an expanded dataset with 19 features. The percentage distribution of these nominal features is also provided in Table 1.

Table 1 – Percentage distribution of nominal features

Feature	Percentage of values in each column
male(ratio)	76.3
female(ratio)	23.6
male (mean age)	53.7
female (mean age)	53.4
asymptomatic (chest pain type)	52.5
non-anginal pain(chest pain type)	23.7
atypical angina(chest pain type)	18.1
typical angina(chest pain type)	5.54
flat (ST slope)	48.9
upsloping (ST slope)	44.2
downsloping (ST slope)	6.81
normal (resting ecg)	57.4
showing probable (resting ecg)	27.3
having ST-T wave abnormality (resting ecg)	15.2

For the implementation of this study, we utilized Python 3.7 and its associated libraries.

Methods

Random Forest, introduced by L. Breiman [11], is a supervised ensemble learning method, specifically based on bagging, that operates using decision trees. The trees are constructed by drawing subsets of training samples with replacement (a bagging approach). Additionally, RF reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data. Most commonly, as in this study, the Gini index is used. The Gini impurity measures how well a split classifies a randomly selected data point within a node. It is calculated as 1 minus the sum of the squared proportions of each class in the node presented in formula 1. The lower the Gini impurity, the "purer" (fewer mixed classes) the split will be.

Gini:

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (1)$$

where p_i- represents the relative frequency of class i.

XGB is a machine learning library based on the gradient boosting algorithm, proposed by Tianqi Chen [12]. The working principle of XGB is based on iteratively adding base models and optimizing the loss function using gradient descent. The main idea of gradient boosting is to sequentially add base models, each of which corrects the errors of the previous models. Unlike Gradient Boosting, XGB uses advanced regularization methods to combat model overfitting and control model complexity, thereby improving its generalization ability. In this article, we used L1 regularization, specifically Lasso Regression.

LGBM stands out for its high speed and good scalability compared to other algorithms. The main difference between LGBM and XGB is the binning method. Binning is the process of grouping continuous numerical features into discrete intervals. In LGBM, a histogram-based binning method is used. It divides the numerical feature values into several intervals (bins) using a histogram and then uses these intervals instead of the original feature values. This reduces the number of unique feature values and simplifies the model, which can improve performance and reduce memory usage.

Table 2 – Parameters of GridSearchCV

Model	Parameter	Range	Optimum Value
RF	'n_estimators'	100 to 400	150
	'max_depth'	4 to 9	5
	'min_samples_leaf'	10 to 15	10
	'min_samples_split'	10 to 20	10
XGB	'n_estimators'	100 to 400	350
	'max_depth'	4 to 9	7
	learning_rate	0.001, 0.01, 0.1, 0.2	0.2
	reg_lambda	0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5	1
LGBM	'n_estimators'	100 to 400	250
	'max_depth'	4 to 9	6
	learning_rate	0.001, 0.01, 0.1, 0.2	0.1
	reg_lambda	0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5	3.5

To create reliable models for RF, XGB, and LGBM, it was necessary to tune hyperparameters for the best results. GridSearchCV was used for optimizing model performance with 5-fold cross-validation. Furthermore, within each fold, a 5-fold cross-validation was implemented to evaluate the selected hyperparameter set. The parameters and ranges for the algorithm models, along with the optimized values determined using GridSearchCV, are presented in Table 2.

Results

As mentioned earlier, the classification algorithms RF, XGB, and LGBM were used. Three corresponding models were created. By analyzing the metrics presented in Table 3, it can be inferred that the LGBM model outperforms the other two models and emerges as the superior choice among them. In this study, we aimed to minimize the Type 2 error since if our model indicates a positive result for the presence of ischemia, it could lead to significant losses, as this type of error is highly significant. This error is unforgivable in the medical field since ensuring the patient's health is of preeminent importance.

Table 3 – Classification metrics results for each fold

Model	Fold	Accuracy	Recall	Precision	F1-score
RF		83.6	85.6	83.6	84.6
		82.3	88.9	80.0	84.2
		84.8	85.7	85.7	85.7
		83.6	85.7	83.7	84.7
		88.2	91.3	87.1	89.1
XGB		88.2	87.2	90.1	88.6
		90.3	89.7	91.9	90.8
		88.2	89.7	88.3	89.0
		91.5	91.3	92.7	92.0
		90.3	88.1	93.3	90.6
LGBM		90.7	89.6	92.6	91.1
		89.4	92.1	88.5	90.3
		89.9	91.3	89.8	90.6
		91.5	92.1	92.1	92.1
		92.8	92.0	93.0	93.2

In addition to these classification metrics, the roc_curve was used to evaluate and visualize the performance of the classification models. From the visualization of the roc_curve results for each fold of the three classification models shown in Figure 2, it can be concluded that the LGBM model exhibits the best performance.

To identify important features in the prediction classification models, the feature_importances_ function from the sklearn library was used. The top 10 most important features were selected for each model, and 5 common features were obtained by comparing them. Figure 3.

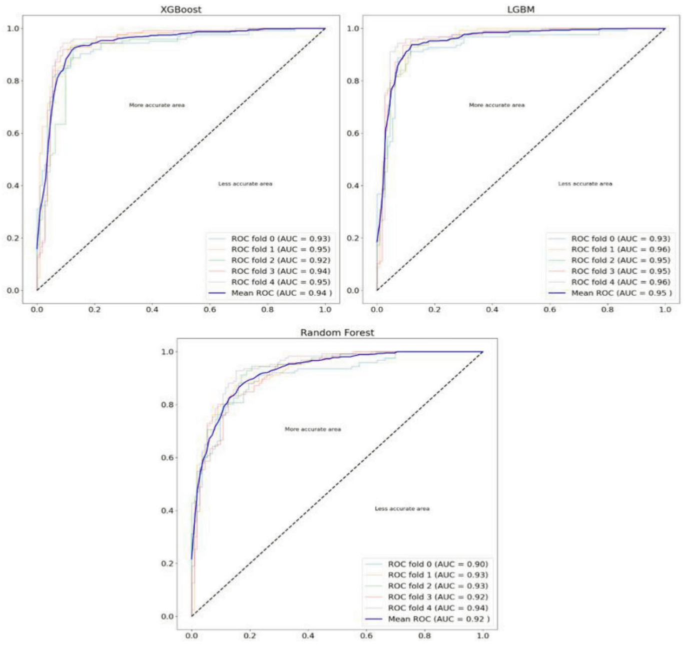


Figure 2 – Visualization of the roc_curve results of each model

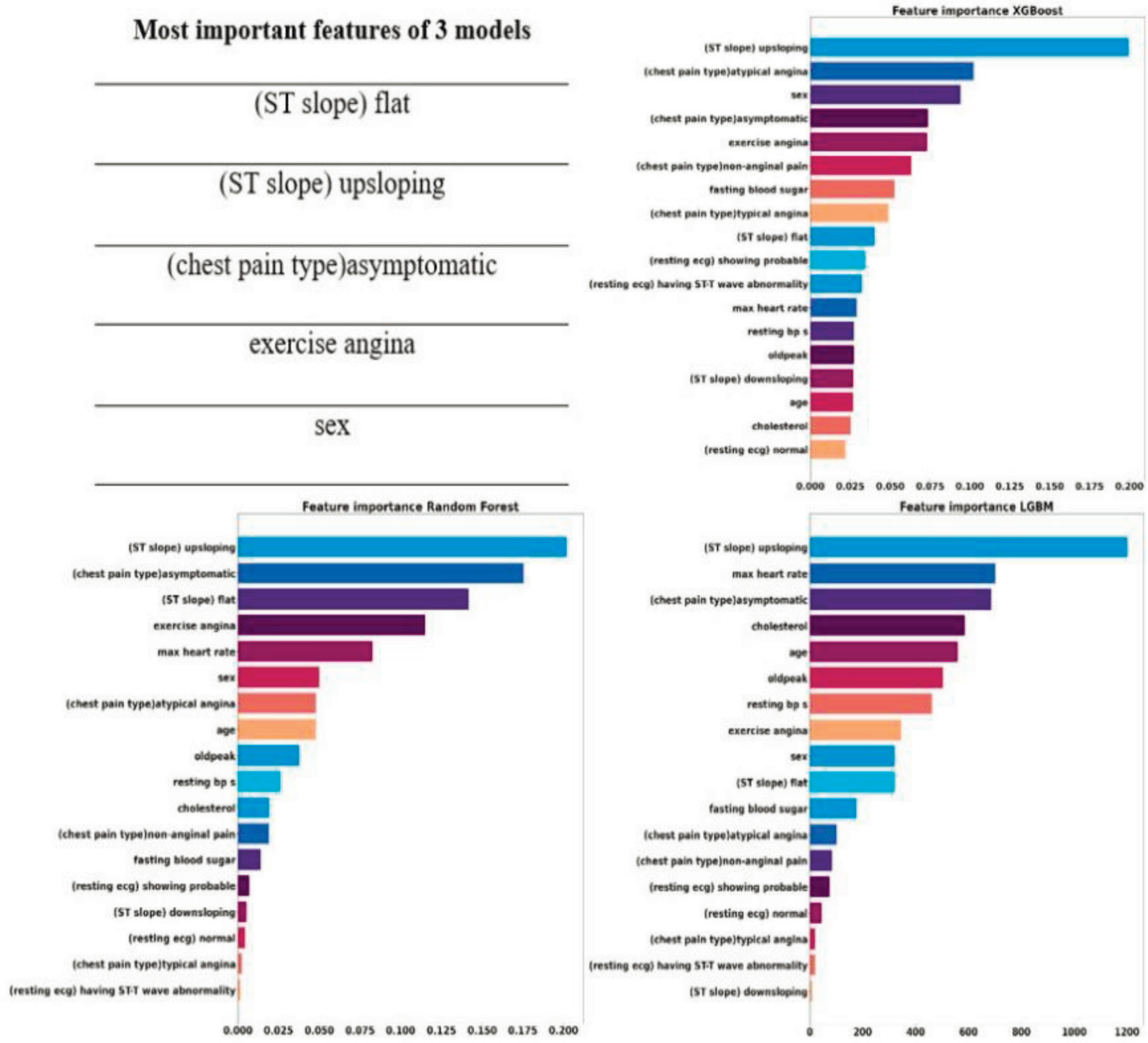
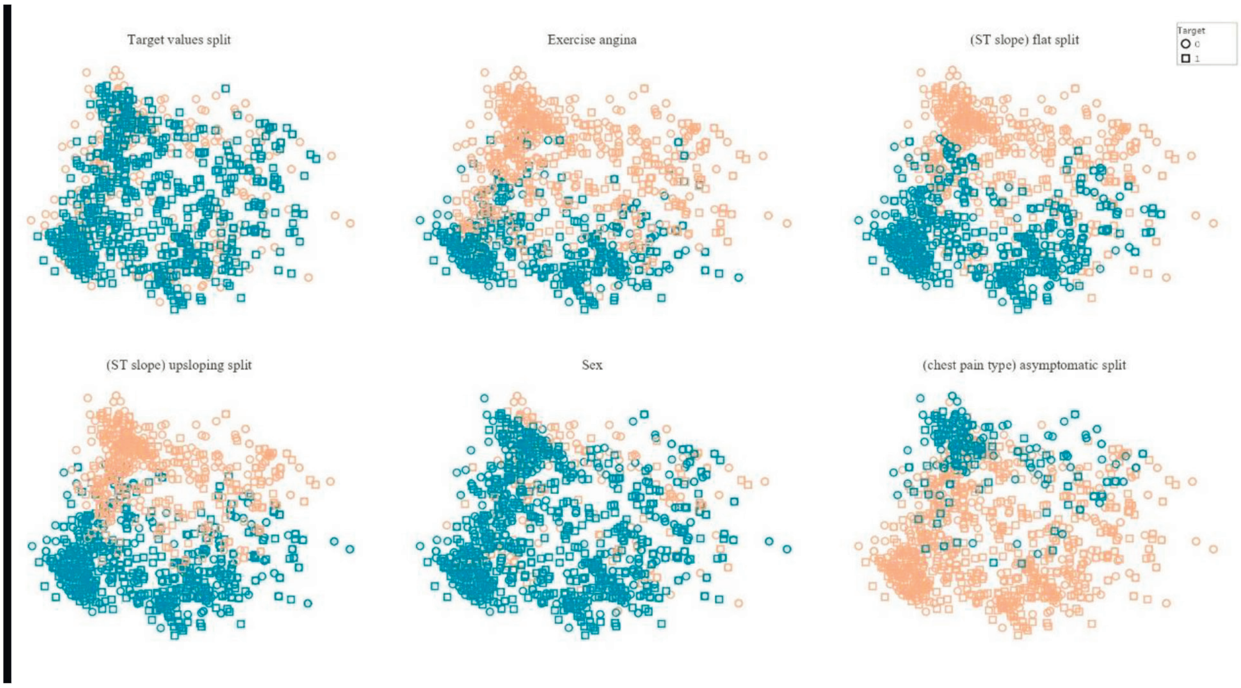


Figure 3 – Visualization of important features of each model

By leveraging the Tableau platform, the dataset can be visually represented, showcasing the division based on the utilization of the 5 important features, Figure 4 illustrates the division of the dataset into two distinct parts in the dimensional space. The PCA (Principal Component Analysis) algorithm was used for data compression. This demonstrates the importance of these features for prediction purposes.

For a more in-depth analysis, we utilized Tableau statistics and divided the dataset into two parts based on the target feature. Furthermore, we confirmed that the dataset could indeed be separated based on the features: 'ST slope flat', 'age, cholesterol', 'chest pain type asymptomatic', 'max heart rate', and 'resting bp s'.



Target values split

Exercise angina

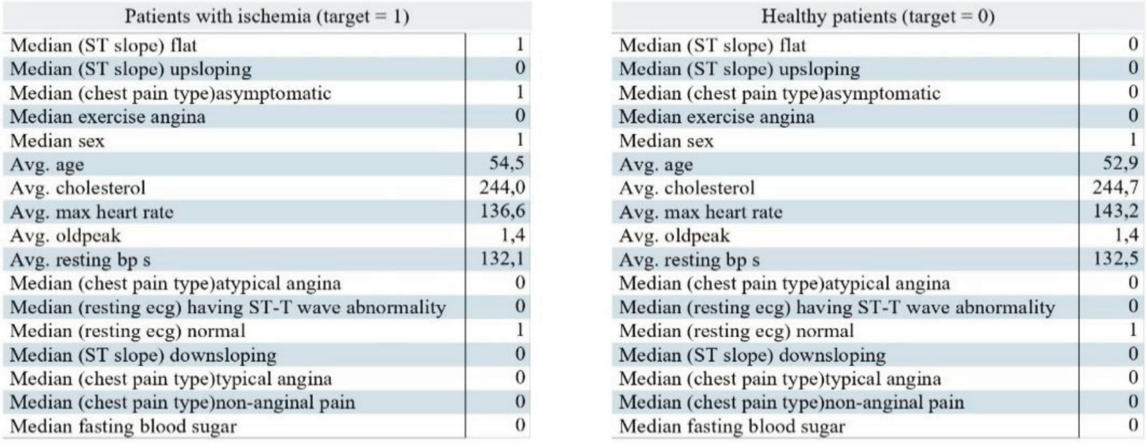
(ST slope) flat split

(ST slope) upsloping split

Sex

(chest pain type) asymptomatic split

Figure 4 – Visualization of tabular data in Tableau by splitting it based on important features



Patients with ischemia (target = 1)	
Median (ST slope) flat	1
Median (ST slope) upsloping	0
Median (chest pain type)asymptomatic	1
Median exercise angina	0
Median sex	1
Avg. age	54,5
Avg. cholesterol	244,0
Avg. max heart rate	136,6
Avg. oldpeak	1,4
Avg. resting bp s	132,1
Median (chest pain type)atypical angina	0
Median (resting ecg) having ST-T wave abnormality	0
Median (resting ecg) normal	1
Median (ST slope) downsloping	0
Median (chest pain type)typical angina	0
Median (chest pain type)non-anginal pain	0
Median fasting blood sugar	0

Healthy patients (target = 0)	
Median (ST slope) flat	0
Median (ST slope) upsloping	0
Median (chest pain type)asymptomatic	0
Median exercise angina	0
Median sex	1
Avg. age	52,9
Avg. cholesterol	244,7
Avg. max heart rate	143,2
Avg. oldpeak	1,4
Avg. resting bp s	132,5
Median (chest pain type)atypical angina	0
Median (resting ecg) having ST-T wave abnormality	0
Median (resting ecg) normal	1
Median (ST slope) downsloping	0
Median (chest pain type)typical angina	0
Median (chest pain type)non-anginal pain	0
Median fasting blood sugar	0

Figure 5 – Compare two targets with median and average

This suggests the hypothesis that important features can be identified using simple tools such as visualization on the Tableau platform, as well as median and average, without resorting to machine learning algorithms.

Conclusion

Cardiovascular diseases are a leading cause of premature death. The 5 most important features: 'exercise angina', '(ST slope) flat', '(ST slope) flat upsloping', 'sex', and '(chest pain type) asymptomatic' were obtained by combining the results of the 10 important features of 3 classification models in Figure 3. They are key factors in predicting heart disease. Using the Tableau platform, data visualization was demonstrated by dividing the dataset based on important features, the result of which proves the significance of these features. The LGBM model demonstrated the best performance among all models, based on the average ROC score (AUC=0.95). Based on the obtained visualization results in Tableau, there are plans to develop an algorithm on this platform in the future. This algorithm will allow the identification of important features that influence the target values without the need for machine learning algorithms.

References

1. World Health Organization (n.d.). World Health Statistics. www.who.int. Retrieved May 20, 2023, from https://www.who.int/

2. Smith J., Johnson A., & Williams B. (2023). Predicting Cardiovascular Diseases as the Leading Cause of Death. Journal of Health and Medicine, 123-135. https://doi.org/10.1234/jhm.2023.10.3.123

3. Brites, I. S. G., da Silva, L. M., Barbosa, J. L. V., Rigo, S. J., Correia, S. D., & Leithardt, V. R. Q. (2021, October). Machine learning and iot applied to cardiovascular diseases identification through heart sounds: A literature review. In Informatics (Vol. 8, No. 4, p. 73). MDPI. https://doi.org/10.3390/informatics8040073

4. Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 102289. https://doi.org/10.1016/j.artmed.2022.102289.

5. Bhushan, M., Pandit, A., & Garg, A. (2023). Machine learning and deep learning techniques for the analysis of heart disease: a systematic literature review, open challenges and future directions. Artificial Intelligence Review, 1-52. https://doi.org/10.1007/s10462-023-10493-5

6. Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University-Computer and Information Sciences, 34(7), 4514-4523. https://doi.org/10.1016/j.jksuci.2020.10.013.

7. Singh, Y. K., Sinha, N., & Singh, S. K. (2017). Heart disease prediction system using random forest. In Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11-12, 2016, Revised Selected Papers 1 (pp. 613-623). Springer Singapore. https://doi.org/10.1007/978-981-10-5427-3_63

8. Yildirim, O., & Pehlivan, N. (2021). Prediction of heart disease using LightGBM algorithm. Healthcare Informatics Research, 27(1), 37–48. doi:10.4258/hir.2021.27.1.37.

9. Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart disease prediction using exploratory data analysis. Procedia Computer Science, 173, 130-139. https://doi.org/10.1016/j.procs.2020.06.017.

10. Manu Siddhartha. (2020). Heart Disease Dataset (Comprehensive). IEEE Dataport. https://dx.doi.org/10.21227/dz4t-cm36

11. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

12. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785