International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Heart Disease Risk Prediction: A Comparison of Machine Learning Techniques

Ayşe Banu Birlik*

Department of Medical Services and Techniques, Beykoz University, Istanbul, Turkey
Graduate School of Engineering and Natural Sciences, Health System Engineering Program, Istanbul Medipol University, Istanbul, Turkey
Email: aysebanubirlik@beykoz.edu.tr

Abstract

Healthcare services have once again demonstrated their worldwide importance under the pandemic conditions. Under the leadership of Industry (4.0), data mining continues to develop in the field of health. Data mining prediction tool act a critical role in healthcare. Heart disease is the most dangerous noncommunicable disease in the world. To predict heart disease, a variety of data mining techniques are used. The study's goal is to use classification algorithms to predict the occurrence of heart disease in an individual. In the study, a dataset consisting of 14 variables belonging to 303 patients accessed from the Kaggle site was used. 75% of the dataset is split into training sets and 25% into test sets in order to train and test the model. Classification performances were compared using K-Nearest Neighborhood (KNN), Random Forest (RF) and Decision Tree (DT) algorithms. As a result of the study, it was determined that the classification accuracy of KNN, RF and DT algorithms was 88.16%, 89.47% and 84.21%, respectively.

Keywords: Classification Algorithm; Decision Making; Machine Learning; Risk Prediction.

1. Introduction

Diseases of the heart and blood arteries are collectively referred to as cardiovascular diseases (CVDs). Stroke and ischemic heart disease (IHD) are the main causes of morbidity and death worldwide. CVDs are also a major source of morbidity and mortality [1]. CVD is still a leading cause of death and increased healthcare costs [2,3].

---

* Corresponding author.

In 2019, cardiovascular disease, one of the noncommunicable diseases, was responsible for 9.6 million male deaths and 8.9 million female deaths, accounting for roughly one-third of all deaths worldwide. [4]. Although CVDs can be avoided through lifestyle changes and other preventive measures, various World Health Organization reports show that CVDs are on the rise globally, which is very concerning [5]. Estimating the risk of developing cardiovascular disease in individuals is important for preventive approaches and treatment. Identification of cardiovascular risk factors and determination of the cardiovascular risk of the patient is possible through lifestyle changes and appropriate medical treatment [6]. Parameters such as smoking, hypertension, diabetes, obesity and sedentary life are considered risk factors for heart disease. Some risk factors can be changed, but family history, age and gender are cardiovascular risk factors that cannot be changed [7]. The increase in deaths and injuries due to heart diseases (HD) on a global scale requires the health sector to find new approaches to these diseases. Techniques like data mining and machine learning (ML) are crucial for processing and analyzing medical data.

ML is widely used in healthcare decision support systems to aid clinical diagnoses and disease predictions [8]. Many areas of health informatics have recently benefited from the application of data mining and ML techniques. In order to address the shortcomings of traditional methods based on invasive detection of HD, researchers have attempted to create a non-invasive intelligent health care system based on predictive ML technologies such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), Random Forest (RF), and Decision Tree (DT) [9]. Cleveland heart disease datasets are widely used by researchers in the literature [10]. Different results have been obtained from various methods by applying data mining algorithms for diagnosis.

Pereira, N. [11], the study's findings show that Logistic Regression (LR) is a good model for fitting to the heart disease dataset (UCI Machine Learning). Budholiya K. and his colleagues [12], used the optimized XGBoost (Extreme Gradient Boost) classification algorithm to predict the risk of heart disease. In this predicted diagnostic model, hyperparameter optimization was performed with Bayesian, which is an efficient method. The performance of the proposed model was evaluated in the Cleveland heart disease dataset and compared with RF and DT algorithms. The proposed method had a 91.8% prediction accuracy. M. Anbarasi and his colleagues [13], using genetic algorithms, researchers attempted to identify the factors that cause heart disease. As a result of the research, the six most important factors among the factors grain were identified. The DT technique was found to perform better. Kalluri, H.K & Tulasi K.S. [14], suggest comparing traditional approaches to predicting the disease at an early stage, such as LR, NB, SVM and artificial neural networks (ANN) algorithms. Finally, the proposed network model predicted disease with the highest accuracy of 94.78% on the dataset. An accurate model of predicting cardiovascular disease is presented by K. G. Dinesh and his colleagues [15], by comparing the accuracy of applying the rules to the individual results of the gradient boost, SVM, RF, NB, and LR on a dataset. Additionally, it plans to use ML methods made available by the R programming language. The attributes gathered from datasets provide accuracy. The LR algorithm, with a rate of 92% accuracy, is the best algorithm for finding the prediction based on the accuracy produced by the algorithms. D. Krishnani and his colleagues [16], proposed using machine learning algorithms such as RF, DT, and KNNs to predict the risk of coronary heart disease. A comparison study was also conducted between these algorithms based on prediction accuracy, and k-fold cross validation was used to generate randomness in the data. These algorithms were tested

on the 4240-record "Framingham Heart Study" dataset. The accuracy of RF, DT, and KNN was 96.8%, 92.7%, and 92.89%, respectively, according to their experimental results. Padmaja and his colleagues [17] created a model that predicts heart disease using classification algorithms and ML techniques. Models, LR, RF, SVM, KNN, DT, gradient boosting classification algorithms, including NB, and other polynomial-based algorithms are developed. They used a feature selection algorithm called χ^2 (chi square) to select the basic features from the input dataset in order to reduce execution time and improve classifier performance. According to the results of the performance evaluation, the RF model provides good accuracy among all classifiers. Dhankhar & Jain [18] goal is to find the best algorithm for predicting HD using several ML techniques from the UCI heart disease datasets. To evaluate the accuracy of the classification model, the entire dataset is divided into two HD prediction sets: 80% and 20% test data. KNN, RF, DT, and SVM algorithms were used to assess the provided data, and it was discovered that RF, with 90% accuracy, is the best classification model. Guruprasad and his colleagues [19] present ML methods based on RF, DT, KNN, and NB algorithms, as well as many characteristics of heart disease. The accuracy of various algorithms is tested by using data from the Cleveland UCI HD patient repository and 303 cases with 14 features in the collection. Data are preprocessed before being used in the model. The result shows that KNN has the highest accuracy of 90.78%. Gao and his colleagues [20] improved the accuracy of heart disease prediction using ensemble learning techniques. They utilized the procedure of choosing important features of the UCI HD dataset using linear discriminant analysis (LCA) and principal component analysis (PCA) algorithms. They compared ensemble methods and five classifiers (KNN, SVM, NB, DT, and RF) based on selected features. They came to the conclusion that the DT and PCA feature extraction methods, as well as the bagging ensemble learning algorithm, performed the best. In order to increase accuracy, Mohan and his colleagues [21] suggested a hybrid model for CVD prediction that makes use of ML methods. The proposed model was employed as a combination of the LR and RF techniques. With an accuracy rate of 88.7%, the proposed hybrid model was successful in the investigation.

2. Material and Methods

2.1. Dataset

This study was conducted on a database of 303 samples with and without heart disease risk provided by the University of Cleveland (UCI machine learning repository) in the Kaggle database. The data set consists of 14 different columns. The data has processed with python programming using Pycharm (*JetBrains individual licenses).* Different python libraries such as numpy, pandas, matplotlib and scikit-learn are used in processing algorithms. There are 8 categorical attributes and 6 numeric attributes.

2.2. Data Preprocessing

Data cleaning is a process that must be done before data analysis is performed. It includes operations like filling in missing data, removing inconsistencies, and detecting outliers. No missing data values of different attributes have detected for the cardiovascular dataset used in this study. One of the most important transformations you must perform on your data is feature scaling. With a few exceptions, ML algorithms do not perform well when the scales of the input numerical attributes are very different [22]. Therefore, rescaling was made so that the values with different scales in the dataset change in the range of 0-1.

2.3. Modeling

For training and testing purposes, preprocessed data was divided into 75% training and 25% testing. These data were then tested using machine learning classifiers such as K-Nearest Neighbor (KNN), Random Forest (RF), and Decision Tree (DT).

2.3.1. K-Nearest Neighbor (KNN)

A supervised classification algorithm is the K-nearest neighbor algorithm. It calculates the target class based on how similar the data labels in the training set are to the model. The classification process is carried out by the algorithm by comparing the new data to the data in the training set.

The simplest method of changing the distance function in nearest neighbor classification is feature weighting. KNN computes the distance between data point features (like Euclidean distance). It collects unclassified data and classifies the new sample by identifying the k samples closest to the new sample in the training set [23].

2.3.2. Random Forest (RF)

A decision tree is a forest of randomly generated decision trees in a supervised machine learning algorithm. The algorithm's goal is to combine and present tree decisions trained in different training sets rather than a single decision tree. Rather than simply resampling the data, we also use random subsets of our predictors to find splits. As a result, a "forest" of trees is created [24, 25].

2.3.3. Decision Tree (DT)

The decision tree is one of the most intuitive tools for data classification. It is one of the most basic algorithms, but it is also one of the most effective and useful. Decision trees are used in medicine to diagnose diseases and make treatment decisions for individuals or communities.

A decision tree is a rooted, directed tree with the appearance of a flowchart. The data item begins at the root and builds the classification tree. It must be determined which fields in the training data will be used and the order in which the tree will be constructed. The tree is mapped or plotted using two different methods: the Gini index and the entropy rule. The item of data is repeated until it reaches a leaf node [23].

3. Results

3.1. Data Distribution

The dependent variable is the target variable, which indicates whether the patient has heart disease. Figure 1 depicts the class distribution, with 303 observations showing 165 cases of heart disease and 138 cases of no heart disease.

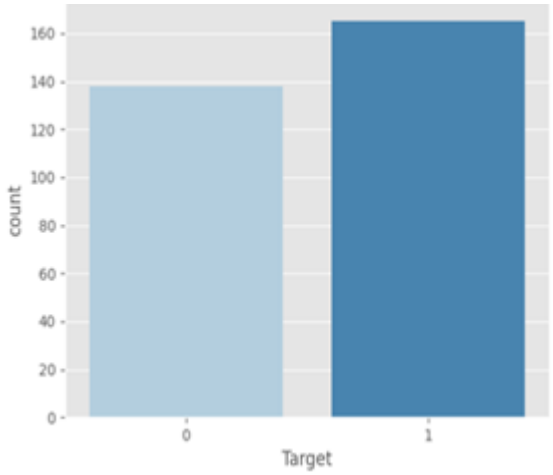

Figure 1: The target distribution of dataset.

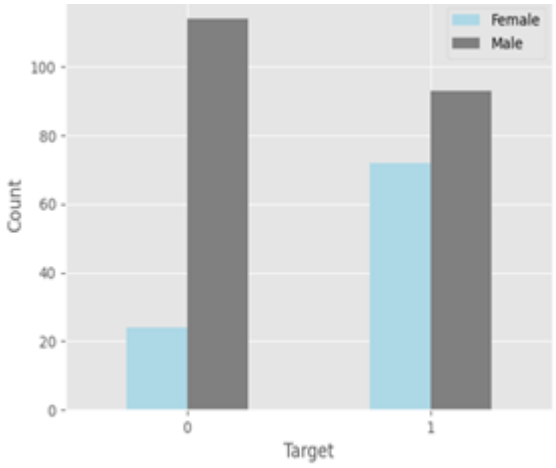

Figure 2: Heart disease frequency for sex.

Figure 2 depicts the relationship between sex and heart disease, while Figure 3 depicts the data frame attributes as histograms using the data visualization procedure.

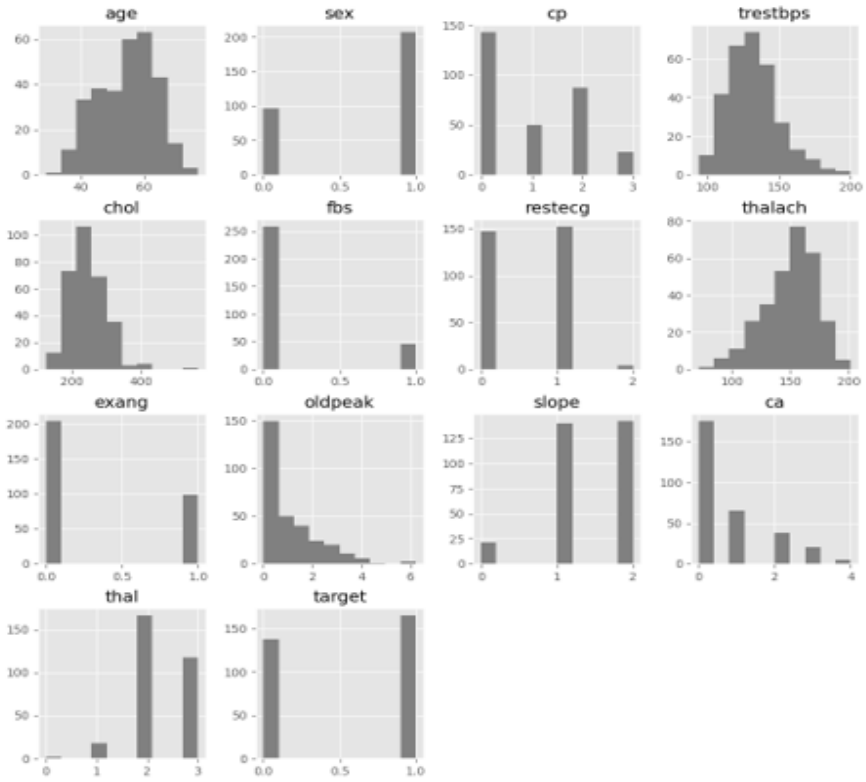

Figure 3: Histograms of data frame attributes.

The maximum heart rate attained decreases with increasing age in patients without heart disease, according to the relationship graph between maximum heart rate and age (Figure 4). Furthermore, regardless of age, the maximum heart rate reached in patients with heart disease is between 140-160 beats per minute.

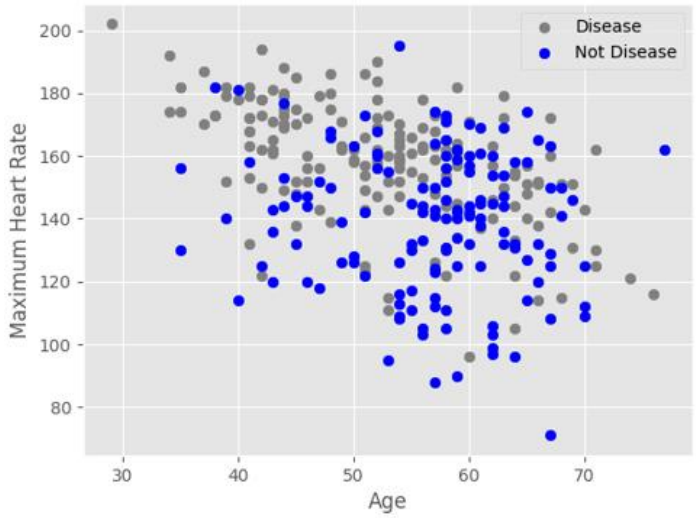

Figure 4: Relation between maximum heart rate and age.

The correlation between the variables is presented in Figure 5. According to the color scale, the correlation relationship between the columns close to red is high, while the correlation relationship decreases gradually in the colors towards the blue. In the data set, it is seen that cp, thalach and slope variables have the highest correlation with the patient's heart disease status. In addition, a correlation above ±0.5 was not detected between the dependent variable (target) and other variables. This shows that no variable can be used independently in the estimation.

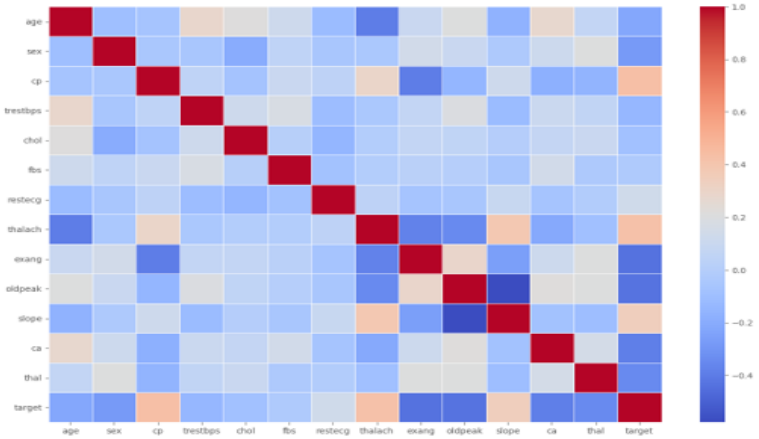

Figure 5: Correlation matrix.

3.2. Performance Evaluation

According to the confusion matrix, the KNN algorithm correctly predicted 67 of 76 test data. However, there are 4 people who are not predicted to have heart disease, and 5 people who are considered to be not heart patients. When Random Forest algorithm is applied on data, accurate estimation was made for 68 people in 76 test data. Two people were estimated to be patients when they were not heart patients, and 6 were classified as healthy,

although they were heart patients. In the application made with the decision tree algorithm, 64 people were correctly classified in the test set of 76 people according to the confusion matrix. 7 people were classified as patients, although they did not have heart disease. Although 5 people have heart disease, they are classified as healthy (Figure 6).

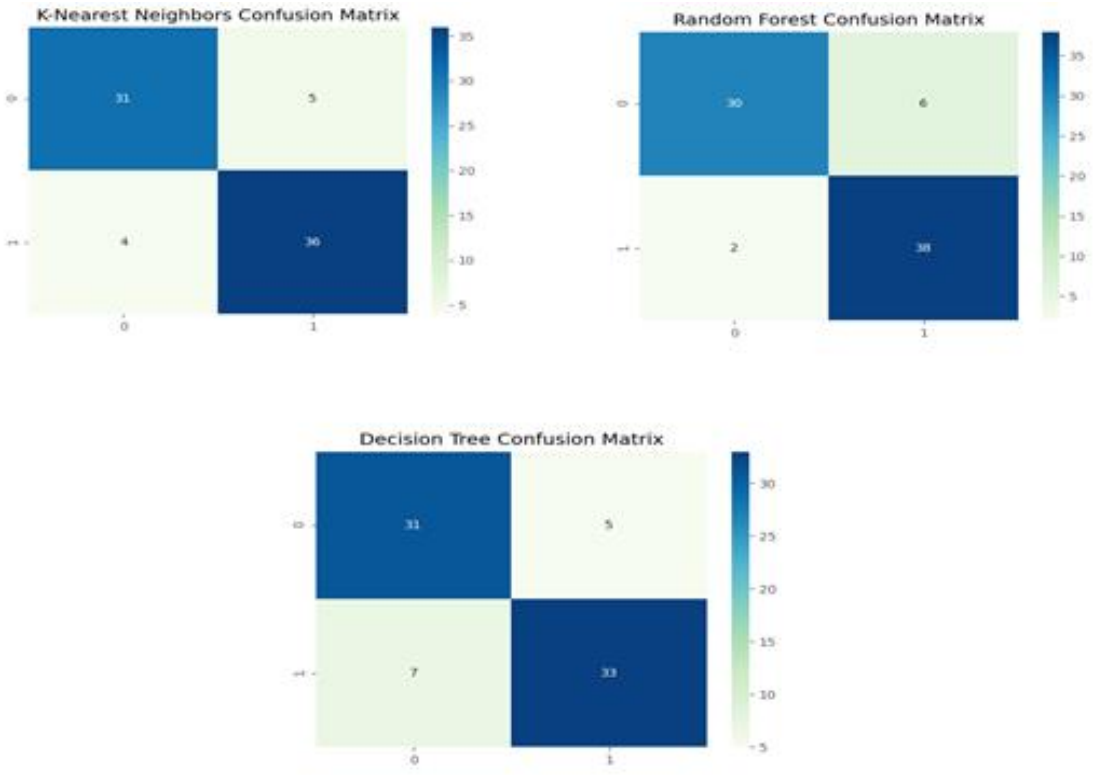

Figure 6: Confusion matrix.

Figure 7 presents KNN varying number of neighbors and model accuracy. The KNN algorithm was trained up to 9 neighbors. For k = 7, the maximum test accuracy is seen. Therefore, by training the model for the 7 nearest neighbors, the accuracy increased from 84.21% to 88.16%.

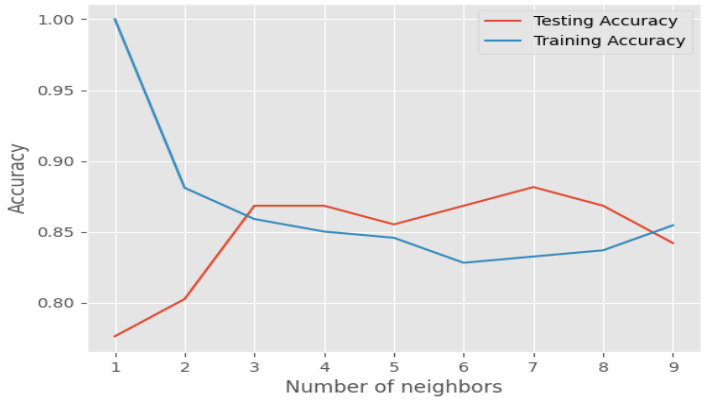

Figure 7: KNN varying number of neighbors.

The Random Forest algorithm predicts diseases with the highest accuracy rate of 89.47% of the three algorithms used to predict heart diseases using data, compared to K-Nearest Neighbors (88.16%) and Decision Tree (84.21%) (Table 1).

Table 1: Accuracy of the classification algorithms.

Classification Algorithm	Accuracy in (%)
K- Nearest Neighbors	88.16
Random Forest	89.47
Decision Tree	84.21

Figure 8 clearly shows that RF (AUC = 0.95) is the best classifier, followed by KNN (AUC = 0.90) and DT (AUC = 0.88).

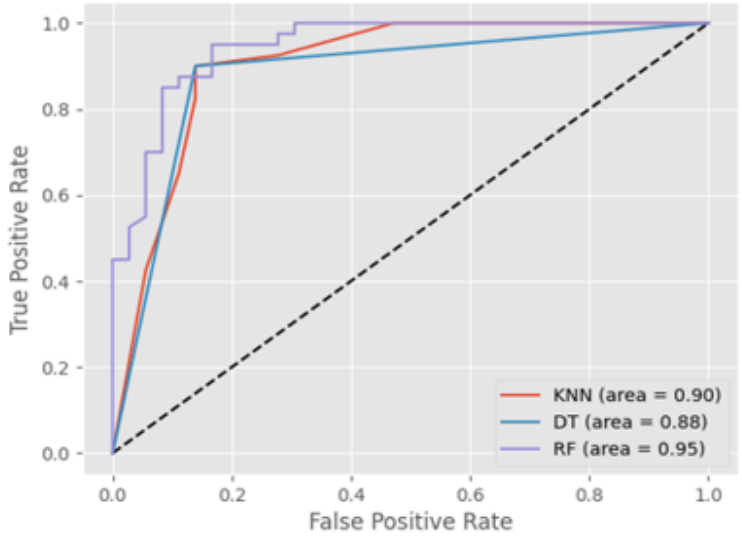

Figure 8: ROC curves from the investigated models.

4. Conclusion

Data mining algorithms are crucial in detecting heart disease and determining risk factors. Machine learning algorithms such as KNN, RF, and DT were used in this study to predict the risk of heart disease. The classification algorithms are compared in terms of correlation and confusion matrix, accuracy and ROC. When the data mining classification algorithms were examined in terms of accuracy, the best result was obtained from the RF classification method with an accuracy rate of 89.47%. It is clear that the study's findings are consistent with the findings of other studies investigated and demonstrate success within the reference ranges specified in the studies. The fact that clinical data variables lack details or risk factors are insufficient indicates that medical data sets are limited. This situation necessitates the collection of more data and analysis in order to accurately predict heart diseases. The study is limited because the amount of heart disease data used is insufficient to

adequately address the issues, and the best-performing model has not been externally validated in an unused cohort of patients. This disparity in clinical severity may have an impact on prediction accuracy. Given that machine learning algorithms perform better with larger data sets, it is anticipated that the study can be expanded by working with a larger data set. In the future, multiclass classification of cardiovascular disease datasets can be considered by including coronary artery angiography and coronary artery calcium scoring parameters in the dataset, which are important in the early diagnosis of cardiovascular diseases. Since prediction models play an important role in the decision-making process, there is a need to evaluate new methods with more complex mathematical assumptions. A roadmap needs to be drawn in order to obtain meaningful information in the cumulative data.

References

[1]. Gregory A. Roth, e. a. (2020). Global Burden of Cardiovascular Diseases and Risk Factors 1990–2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology, 76*(25), 2982-3021.

[2]. G.A. Mensah, G. R. (2019). The global burden of cardiovascular diseases and risk factors: 2020 and beyond . *J Am Coll Cardiol*, 2529-2532.

[3]. T. Vos, S. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019 Lancet. 1204-1222.

[4]. Gregory A. Roth, G. A. (2020). The Global Burden of Cardiovascular Diseases and Risks: A Compass for Global Action. *Journal of the American College of Cardiology, 76*(25), 2980-2981.

[5]. Judith Mackay, G. M. (2004). *The atlas of heart disease and stroke* . World Health Organization.

[6]. Tekin, A. (2018). The calculation of cardiovascular mortal risk and the evaluation of the level of knowledge of cardiovascular risk factors with score equality between 40-65 years age. İzmir: İzmir Katip Çelebi Üniversitesi Tıp Fakültesi Aile Hekimliği Anabilim Dalı.

[7]. C. Beulah Christalin Latha, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked, 16*, 100203.

[8]. Lamido Yahaya, N. D. (2020). A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. *American Journal of Artificial Intelligence., 4*(1), 20-29.

[9]. Azmi, J., Arif, M., Nafis, M. T., Alam, M. A., Tanweer, S., & Wang, G. (2022). A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering and Physics*, *105*(February), 103825. https://doi.org/10.1016/j.medengphy.2022.103825

[10]. Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific Reports*, *10*(1), 1–17. https://doi.org/10.1038/s41598-020-76635-9

[11]. Nestor, P. (2020). Using Machine Learning Classification Methods to Detect the Presence of Heart Disease. Technological University Dublin

[12]. Kartik Budholiya, S. K. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University Computer and Information Sciences*.

[13]. M. Anbarasi, E. A. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm. *International Journal of Engineering Science and Technology*, 5370-5376.

[14]. Kalluri, H. K. (2020). A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network. *Revue d intelligence artificielle*, 601-606.

[15]. K. G. Dinesh, K. A. (2018). Prediction of Cardiovascular Disease Using Machine Learning Algorithms. *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, (s. 1-7).

[16]. D. Krishnani, A. K. (2019). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms . *TENCON 2019 - 2019 IEEE Region 10 Conference* , (s. 367-372).

[17]. Padmaja, B., Srinidhi, C., Sindhu, K., Vanaja, K., Deepika, N. M., & Krishna Rao Patro, E. (2021). Early and Accurate Prediction of Heart Disease Using Machine Learning Model. *Turkish Journal of Computer and Mathematics Education*, *12*(6), 4516–4528.

[18]. Dhankhar, A., & Jain, S. (2021). Prediction of diabetes disease using machine learning algorithms. In N. Gupta, P. Chatterjee, & T. Choudhury (Eds.), *In Smart and Sustainable Intelligent Systems* (pp. 115–126). Scrivener Publishing LLC. https://doi.org/10.1002/9781119752134.ch8

[19]. Guruprasad, S., Mathias, V. L., & Dcunha, W. (2021). Heart Disease Prediction Using Machine Learning Techniques. *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, ICEECCOT 2021 - Proceedings*, *1*(6), 762–766. https://doi.org/10.1109/ICEECCOT52851.2021.9707966

[20]. Gao, X. Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. *Complexity*, *2021*. https://doi.org/10.1155/2021/6663455

[21]. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

[22]. Géron, A. (2017). *Hands On Machine Learning with Scikit-Learn and TensorFlow.* O'Reilly Media.

[23]. J., C. C. (2015). *Data Classification Algorithms and Applications.* New York, USA: Taylor & Francis Group.

[24]. Göktaş, M. Y. (2020). Veri Bilimi Uygulamalarının Hastalık Teşhisinde Kullanılması: Kalp Krizi Örneği. *Journal of Information Systems and Management Research*, 26-32.

[25]. Matloff, N. (2017). *Statistical Regression and Classification From Linear Models to Machine Learning.* USA: Taylor & Francis Group.