Heart Disease Prediction Using Machine Learning

D. M. Khalid Mahmud
ID:1901004
Department Of IRE
Session:2019-20
1901004@iot.bdu.ac.bd
Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

MD. Al Morsaline
ID:1901020
Department Of IRE
Session:2019-20
1901020@iot.bdu.ac.bd
Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

ABSTRACT

In light of a recent study conducted by the World Health Organization (WHO), it has become evident that the prevalence of heart-related diseases is on the rise, resulting in a staggering 17.9 million annual fatalities worldwide. This alarming trend is exacerbated by a burgeoning global population, making it increasingly challenging to identify and initiate early-stage treatments.Fortunately, the field of healthcare has seen a surge in advancements driven by the application of Machine Learning (ML) techniques. These technological innovations have unleashed a wave of research endeavors aimed at tackling heart disease. The principal objective of this paper is to construct a robust ML model for predicting heart disease based on a comprehensive set of relevant parameters.For this research, we harnessed a dataset specifically designed for heart disease prediction, containing 14 distinct attributes associated with cardiac health. We deployed a range of ML algorithms, including Random Forest, Support Vector Machine (SVM), Naive Bayes, and Decision Trees, to develop our predictive model.Furthermore, our research explored the intricate correlations between various attributes within the dataset using standard ML methodologies, harnessing these insights to enhance the accuracy of our heart disease predictions. Our results reveal that among the ML techniques employed, Random Forest stands out as the most accurate and time-efficient approach for heart disease prediction.This meticulously crafted model has the potential to serve as a valuable decision support system for medical practitioners in clinical settings. It offers a promising solution to address the growing challenges posed by heart-related diseases in a world where early diagnosis and intervention are of paramount importance.

Author Keywords

Heart Disease, Machine Learning,KNN, Support Vector Machine, Classification,SVM, Naïve Bayes, Random Forest.

INTRODUCTION

The heart is one of the most essential organs in the human body. The primary function of the heart is to power the flow of blood and run it to all parts of the body. Heart disease, a common circulatory disorder also known as cardiovascular disease, encompasses a variety of heart diseases. Heart disease is today the most severe disease that threatens human life. According to World Health Organization estimates, heart disease kills 12 million people worldwide each year, and on average it takes the life of one person every 34 seconds in the United States alone. Coronary heart disease, arrhythmias, and myocardial infarction are the most common heart diseases, affecting 8 billion people worldwide. Take coronary heart disease, for example, according to recent statistics from the American Heart Association, coronary heart disease accounted for 13 percent of deaths in the United States in 2018.

The sudden onset of heart disease and the short time available for resuscitation lead to very few incidents in which heart attack patients are successfully resuscitated. In many cases, the patients die on the spot. Therefore, the rapid and effective prediction of potential patients with heart disease has become a critical and challenging task in the medical field. Current researchers have been interested in predicting heart disease, and they have developed several prediction methods for different heart diseases. Machine learning is currently one of the most rapidly developing subfields of artificial intelligence. It is an effective intelligence tool for rapid analysis of data and is used in many areas of life, including health care.

Many medical institutions worldwide and hospitals worldwide count data on various characteristics of heart patients, such as the patient's gender, age, heartbeat per minute, blood pressure, and other common data from daily medical examinations. A large amount of patient data is collected. Human observation alone still cannot obtain valid information from these large amounts of data or derive heart patients' characteristics. Machine learning algorithms can learn from existing patient cases and quickly analyze the data for heart disease, which is why machine learning is widely used in the medical field to analyze disease data.

Classification, the most basic type of machine learning, enables quickly dividing the data as required. In predicting heart disease, effective classification can promptly classify the indicated person into two categories by different features,

whether they have heart disease or not. For example, the well-known UCI heart disease dataset can be used to train and test classifiers

The accuracy ofheart disease prediction has been a concern for researchers because the prognosis of the condition affects the judgment of physicians and the self-protection of potential patients, and incorrect prediction can have incalculable consequences.

The rest of this paper will present heart disease and coronary heart disease predictions based on different datasets and summarize the data results and predictions obtained by researchers through various machine learning methods and compare the performance of different machine learning methods in heart disease prediction.

LITERATURE REVIEW

Lots of research work have been done for assessment of the classification accuracies of different machine learning algorithms by using the Cleveland heart disease database which is uninhibitedly accessible at an online data mining repository of the UCI. Authors of achieved 77accuracy by applying logistic regression algorithm on this dataset. In this study, authors did an enhanced work by doing comparison of global evolutionary computation approaches and thus they observed higher prediction accuracy. Authors Bayu Adhi Tama, et.al in their work suggested a research related to the identification of diabetes malady with utilization of ML procedures. This disease was viewed as incredibly a thrust area of ML. Roughly 285 million individuals around the globe were experiencing diabetes as per a study directed by International Diabetes Federation (IDF). As a matter of fact, detection of type 2 diabetes at beginning phase isn't a simple undertaking, yet research done by the authors, in which data mining was used on the grounds that it gives the best results, helped in the disclosure of information from accessible data. In their research, they utilized SVMs for the mining of related information of various patients from the previous records. The on-time acknowledgment of type 2 diabetes gave assistance in the taking of legitimate treatment and avoid the risk of expanding.

Yu-Xuan Wang, et.al. have explored different applications that demonstrated the significance of the ML methods in various areas. They proposed a new technique for the designing of a working framework. The approach used the distinct machine learning procedures. After getting the proper result from the data miner, the whole information assembled from the structure was inspected. In light of the various tests, it was seen that proposed approach gave proficient results. Zhiqiang Ge, et.al, (2017) proposed a work on analytics and data mining applications, which was done prior. These procedures were used in business area for various purpose of perspectives. Here they have explored 8 unsupervised and 10 supervised learning algorithms. In their research, they showed an application work for the semi-supervised type learning algorithms. In industry method, it was seen that roughly 90and supervised machine learning procedures. Consequently, it was portrayed that the Machine Learning methods play an indispensable part in the planning of different novel applications for domains like medical services and industry.

PROPOSED METHODOLOGY

The main task is the methodology for the detection and prediction of heart disease. According to literature studies, this task can be classified as a classification problem to classify whether a person has heart disease or not.The task requires heart patient data to predict heart disease. For experimental work, a publicly available heart disease dataset from the Machine Learning Repository is used. The dataset has 1025 rows x 14 columns and 14 attributes, however all published experiments refer to using a subset of 14 of them. The patients are both male and female in the database. The patient's age ranges between 29-77 years. Most patients have chest pain typical of the angina type and the heart rate is between 71-262 bpm. This dataset is divided as two splits namely train and test split using python library scikit-learn. The step-wise methodology adopted for developing the heart disease prediction model is presented in Fig. 1 and discussed in detail.

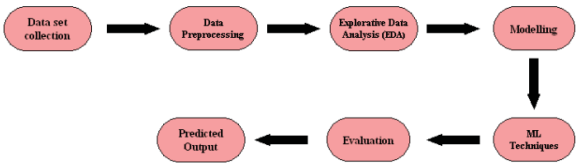

Figure 1. Heart Disease Prediction Methodology.

Dataset Collection

Data is an essential requirement for the prediction of heart disease. The performance of various ML algorithms will be evaluated using a publicly available dataset, the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository. A brief description of each of the features is given in Table 1 for a better understanding. There are 14 features, of which 13 are considered independent variables/features, and the one, namely the Target feature, is known as the dependent variable/feature.

```
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
```

Figure 2. Dataset details for heart disease prediction.

Data Pre-Processing

In general, medical records that are not always complete may contain missing and unwanted data. Data pre-processing is used to remove the number of discrepancies associated with the data, remove duplicate records, normalize values, account for missing data, etc. The primary step in this data pre-processing is to check for null values and treat them by filling in or dropping them. After importing a dataset using the Python library pandas, common data pre-processing methods such as data cleaning, data transformation, efficient processing, and classification are performed. No unique method of data processing is used in this work.

Feature Selection

In this section, I discuss the process of feature selection, an essential step in building a predictive model. Feature selection involves choosing the most relevant variables (features) from the dataset to use in the model. However in this case I selected all the features from my dataset but glucose is the most important feature here.

Data Splitting

Once the data has been pre-processed, it is segregated into two distinct sets, namely the training set and testing set. The dataset is divided into 80and testing. The training set is employed to train the machine learning models while the testing set is utilized to gauge the model's efficacy.

Model Selection

For predicting Heart disease , I am employing machine learning techniques. Decision Tree Classifier (DTC),Random Forest Classifier (RFC),Logistic Regression (LR),K-Neighbors Classifier (KNN), Support Vector Machines (SVM) are some of the algorithms used here.

MACHINE LEARNING TECHNIQUES

We have chosen some popular ML techniques to develop the heart disease prediction model. Details of these techniques are as follows:

Support Vector Machine

Support Vector Machine is a classification technique of Machine learning to, which is used to analyze data and discover patters in classification and regression analysis. SVM is typically mull over when data is characterized as two class problem. In this strategy, data is characterized by finding the best hyper plane that isolates all data points of one class to the other class. The higher separation or edge between the two classes is, the better is the model, considered. The data points lying on limit of the margin are called as support vectors. The actual basis of SVM is mathematical methods used to design complex real-world problems. We have chosen SVM for this experiment because our dataset - Cleveland Heart Disease Dataset CHDD has multi class to predict based on various parameters. In SVM, the mapping of training data is to be done with a function called kernel (Kernels of SVM), these are - linear kernel, quadratic kernel, polynomial kernel, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. Apart from the kernel's functionalities in SVM, few more methods are available such as quadratic programming, sequential minimal optimization, and least squares. While building up the model with SVM, most challenging thing is kernel selection and method selection to evade the issue of overfitting and underfitting. Since our dataset is having enormous number of parameters and instances too. So, we had choice of selecting the RBF or linear kernel. Thus, final model developed by SVM requires tested and validated against actual data.

Decision Tree

Decision Tree algorithm in Machine Learning is used to develop the Classification models. This classification model is based on the tree-like structure. This comes under the category of supervised learning, where the target result is already known. Both the categorical and numerical data can be applied on Decision tree algorithm. Decision tree consists of root node, branches and leaf nodes. Data is evaluated on the basis of traversing path from the root to a leaf node. For our dataset - CHDD, a total of 283 tuples were assessed down the decision tree. They potentially came to a positive or negative assessment for the heart disease prediction. These were compared to the actual parameters to check for the false positives/false negatives which show the accuracy, specificity, and sensitivity of the model.

Logistic Regression

Logistic Regression is a statistical and machine learning technique used for binary classification problems. It's a type of regression analysis that is well-suited for predicting the probability of a binary outcome (yes/no, 1/0, true/false). Logistic Regression models the relationship between a dependent variable (the binary outcome) and one or more independent variables (predictors) using the logistic function, which produces values between 0 and 1. These values represent the probability that the dependent variable belongs to one of the two classes.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and effective machine learning algorithm used for both classification and regression tasks. It classifies data points based on their similarity to neighboring data points in a training dataset. In KNN, "k" represents the number of nearest neighbors considered, and the algorithm assigns the class of the majority of these neighbors to a new data point, making it a non-parametric and instance-based learning method. KNN is intuitive and easy to understand, making it a popular choice for various classification problems, but its performance may be sensitive to the choice of the value of "k."

Naive Bayes

This supervised machine-learning algorithm is based on the Bayes' Theorem, which consider that features are statistically independent to each other's. The Naïve Bayes Classifier is used with high dimensionality of inputs data. Naive Bayes method is highly useful in computer vision application. In particularly, it has proven itself to be a classifier with good results.

Random Forest Classification

Random Forest is a troupe of unpruned classification based trees. It gives amazing performance with concern to number of real-life problems, as it is non effective to noise in the

dataset and risk of overfitting is also very less. In comparison to many other tree-based algorithms, it works faster than others and generally improves accuracy for testing and validation data. Random forests are the aggregation of the predictions of individual decision tree algorithm. There are various choices to tune the performance of random forest when constructing a random tree.

Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. A good model is one which has high TP and TN rates, while low FP and FN rates.

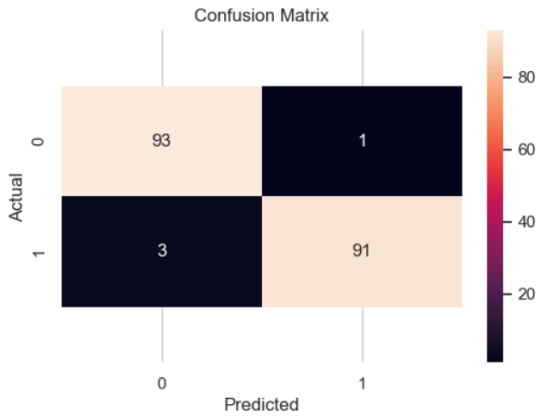

Figure 3. Confusion Matrix .

Receiver Operating Characteristic Curve

Description of ROC Curve (Receiver Operating Characteristic Curve) An ROC Curve is a graph showing the performance of a classification model at all classification thresholds. It measures a model's ability to distinguish between classes, for example 0 and 1. In our case, it is the model's ability to distinguish whether a patient has a heart disease or not. The curve is plotted base on the True Positive Rate and False Positive Rate.

Description of AUC Score (Area Under the ROC Curve Score) AUC measures the entire two-dimensional area under the ROC Curve. It provides an aggregate measure of performance across all possible classification thresholds. A model that has an AUC Score of 0.0 makes every prediction incorrectly, whereas a model that has an AUC Score of 1.0 makes every prediction correctly.

Correlation Matrix

A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship.

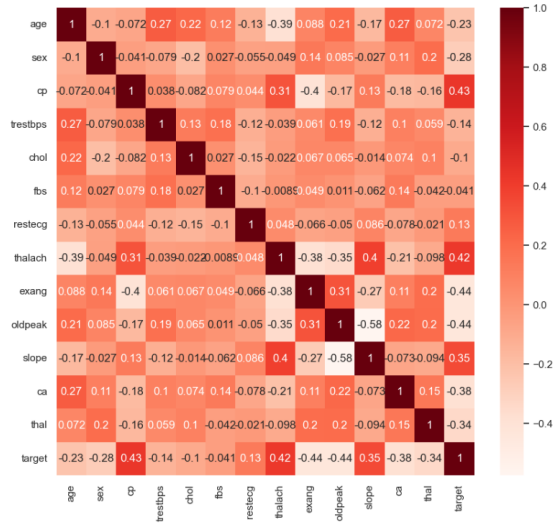

Figure 4. Correlation Matrix.

A scatter plot is a graphical representation that displays individual data points on a two-dimensional plane, with one variable on the x-axis and another on the y-axis. It's useful for visualizing the relationship between two continuous variables and identifying patterns, correlations, or trends in the data.

A bar chart, on the other hand, is a visual representation that uses rectangular bars to represent data. The length or height of each bar is proportional to the value it represents, making it a great choice for displaying and comparing discrete categories or groups. Bar charts are commonly used to illustrate data distribution, comparisons, or trends among different categories

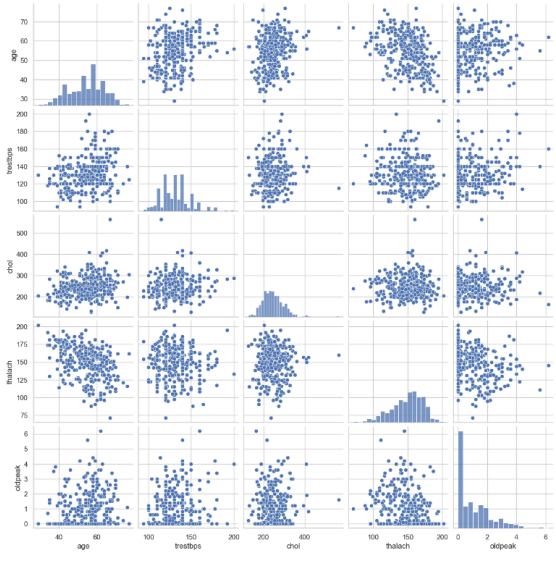

Figure 5. Scatter plot and bar chart plot.

Check for outliers

Checking for outliers involves identifying data points that significantly deviate from the overall pattern of the dataset. Outliers are values that are either unusually high (positive outliers) or unusually low (negative outliers) compared to the majority of the data. Detecting outliers is essential for data quality and statistical analysis, as they can skew results and distort interpretations. Various methods, such as the IQR (Interquartile Range) method, Z-scores, or visual inspection through box plots or scatter plots, can be used to identify and deal with outliers. Handling outliers may involve data transformation, exclusion, or understanding whether they represent genuine anomalies or errors in the data.

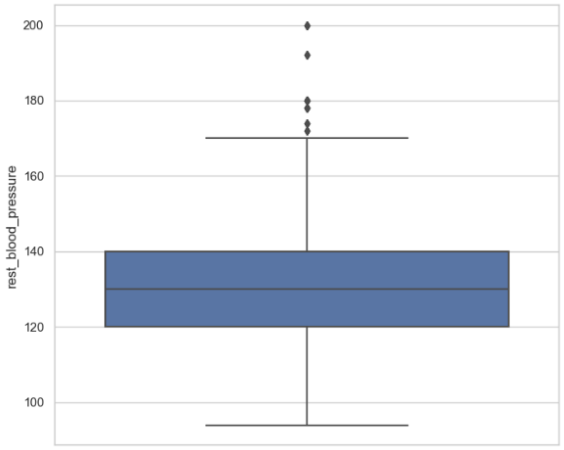

Figure 6. Blood Pressure.

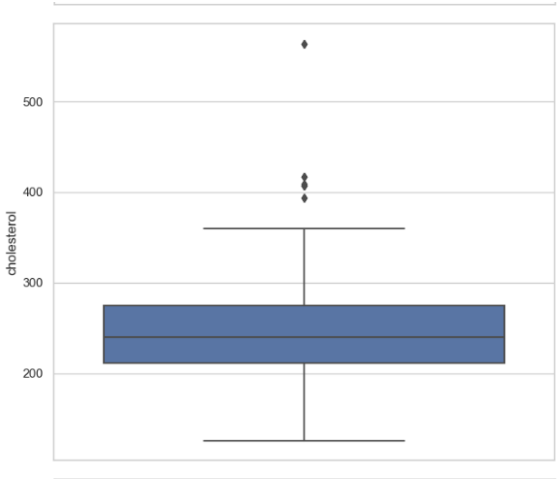

Figure 7. Cholesterol.

After removing outliers, the data has been cleansed, resulting in a more reliable dataset.

MODEL PERFORMANCE AND EVALUATION

The accuracy evaluations of various machine learning algorithms are as follows:

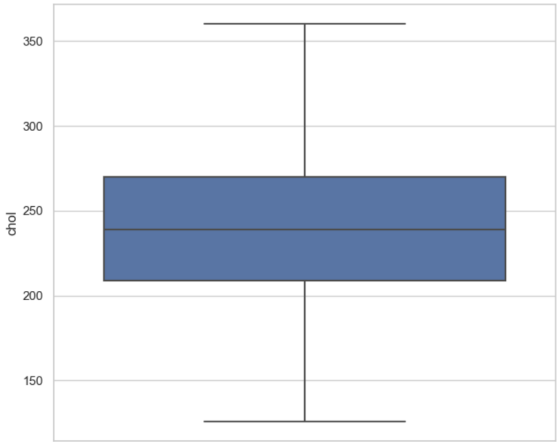

Figure 8. Remove Outliers of Cholesterol.

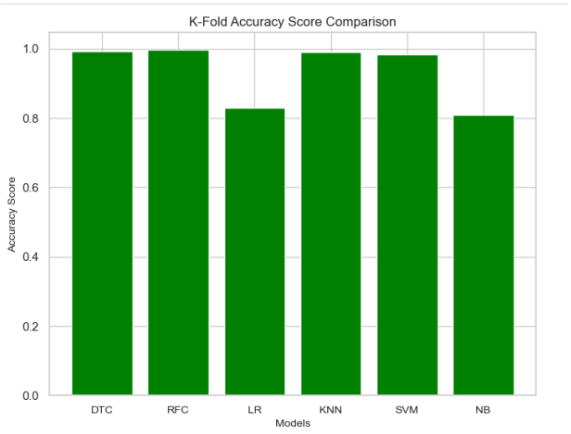

Figure 9. K-Fold Cross Validation Accuracy Score Visualization.

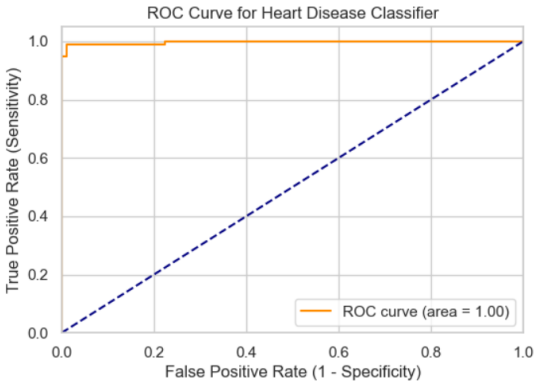

Figure 10. ROC Curve for Heart Disease Classifier.

Decision Tree Classifier (DTC) achieved an accuracy of 98.40% with a standard deviation of 1.56%. It performed well in classification tasks.

Random Forest Classifier (RFC) demonstrated strong performance, attaining an accuracy of 99.47% with a low standard deviation of 0.88%. This algorithm is highly reliable for classification tasks.

Algorithms	Accuracy
DTC	0.984018
RFC	0.994667
LR	0.827035
KNN	0.990684
SVM	0.976035
NB	0.827053

Figure 11. All algorithm accuracy evaluation.

Logistic Regression (LR) achieved an accuracy of 82.70%, with a standard deviation of 5.64%. While it showed lower accuracy compared to other algorithms, it can still be useful for certain tasks.

K-Nearest Neighbors (KNN) exhibited an accuracy of 99.07% with a standard deviation of 1.04%. It is a robust choice for classification tasks, providing high accuracy and consistency.

Support Vector Machine (SVM) achieved an accuracy of 97.60% with a standard deviation of 1.77%. It is a solid performer in classification problems, offering good accuracy.

Naive Bayes (NB) showed an accuracy of 82.71%, with a relatively high standard deviation of 5.74%. While it may not be the most accurate choice, it can still be applicable for certain classification scenarios.

RESULT AND DISCUSSION

we employed machine learning techniques to predict the risk of heart disease using a dataset of medical records and clinical features. The results of our analysis provide valuable insights into the potential for accurate heart disease prediction and underline the significance of machine learning in the field of healthcare.

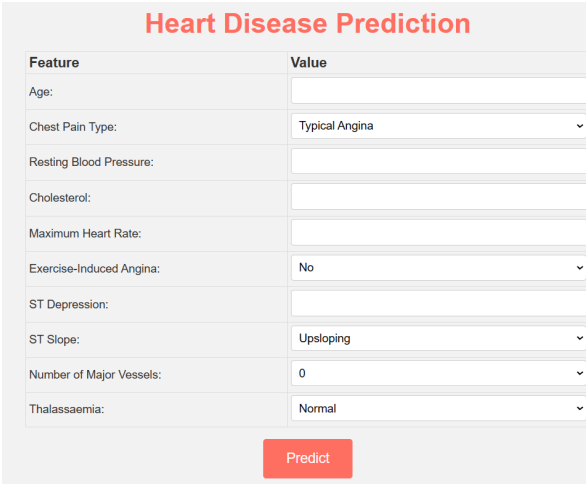

Figure 12. Graphical user interface heart disease

CONCLUSION

Through this research we have attempted to analyze the various machine learning techniques and anticipate if someone

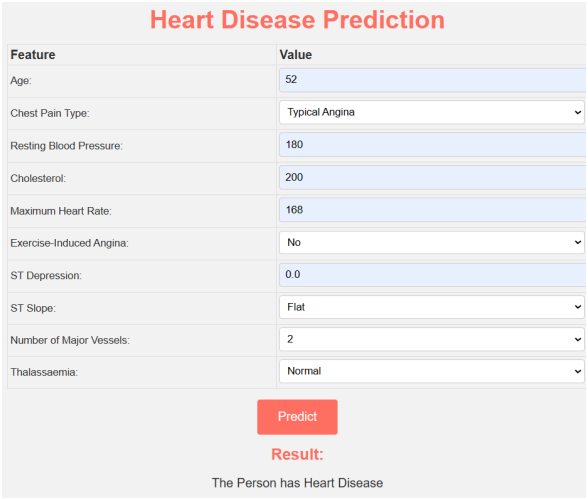

Figure 13. Output of the result

in particular, given different individual attributes and indications, will get coronary illness or not. The primary thought process of our report was to looking at the exactness and analyzing the reasons behind the variation of different algorithms. We have used heart diseases dataset for heart diseases which contains 1025 instances and used percent split to divide the data into two sections which are training and testing datasets. We have considered 14 attributes and implemented four different algorithms to examine the accuracy. By the end of the implementation part, we have discovered that Random Forest is giving the maximum accuracy level in our dataset which is 99 percent and Decision Tree is playing out the least with an accuracy level of 85 percent. Probably for other instances and different datasets other algorithm may work in better manner however for our situation, we have discovered this outcome. Also, on the off chance that we increment the number of training data, maybe we can find more accurate result but it will take more time to process and the system will be slower than now as it will be more perplexing and will be handling more data. In this way, considering these potential things we took this choice, which is better for us to work with.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all those who contributed to the successful completion of this project. We extend our thanks to our project advisors for their invaluable guidance and support throughout this journey. Our heartfelt appreciation goes to the research and medical experts who provided valuable insights and data. We are also thankful to our team members for their dedication and hard work. Lastly, we would like to acknowledge the countless patients and individuals who participated in this study, as their data and cooperation were instrumental in developing the machine learning model for heart disease prediction.

REFERENCES FORMAT

[1] Fahd Saleh Alotaibi, “Implementation of Machine Learning Model to predict Heart Failure Disease,” International Journal of Advanced Computer Science and Applications (IJACSA), vol.10, no.6, pp.261- 268, 2019.

[2] S.Nandhini, Monojit Debnath, Anurag Sharma and Pushkar, "Heart Disease Prediction using Machine Learning," International Journal of Recent Engineering Research and Development (IJRERD), ISSN: 2455-8761, vol.3, pp.39-46, 2018.

[3] Dhai Eddine Salhi, Abdelkamel Tari M-Tahar Kechadi Using Machine Learning for Heart Disease Prediction.

[4] https://en.wikipedia.org/wiki/Machine$_{l}earning$.

[5] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.

[6] S. Dhar, K. Roy, T. Dey, P. Datta and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531

[7]Thenmozhi K., Deepika P., "Heart Disease Prediction Using Classification with Different Decision Tree Techniques, Intemational Journal of Engineering Research and General Science 2(6), (2014).

[8]Jingchao Zhao, Yi Li, "Research on heart disease prediction algorithm based on optimized random forest," Journal of Qingdao University of Science and Technol, pp. (2021).

[9]Seyedamin Pouriyeh, et. al. "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (Iscc 2017): Workshops - ICTS4eHealth (2017).

[10]Ramalingam V.V., Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques: a survey, International Journal of Engineering Technology, 7 (2.8) pp. 684-687, (2018) [11]Hongyan Yu, Qian Feng, "Research review of random forest algorithm,"Journal of Hebei Academy of Sciences, (2019).

[11] Coffey, Sean, et al. "Global epidemiology of valvular heart disease,"Nature Reviews Cardiology 18.12 pp. 853-864, (2021) Rani, Pooja, et al. "A decision support system for heart disease prediction based upon machine learning." Journal of Reliable Intelligent Environments 7.3 pp. 263-275, (2021).

[12] KSasipriya V. R, Deepa E. "Heart diseases detection using naive Bayes algorithm," International Journal of Innovative Science, 2, pp.441-444,(2015).

[13] Tan Teoh, Yu Q. et al. hybrid evolutionary algorithm for attribute selection in data mining," Expert Systems with Applications, 36, pp.8616-8630, (2009).

[14] Chaurasia V, Pal S. "Data mining approach to detect heart disease," International Journal of Advanced Computer Science and Information Technology, 2(4) pp. 56-66, (2013).

[15] Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis," Health and Technology 11.1 pp.87-97, (2021).

[16] Baudet, Daugareil, Laulom, et al. "Cardiovascular diseases," Annales cardiology et angiology, 68 (l), pp.49-52, (2018).