1st International Conference on Computational Intelligence and Sustainable Engineering Solution(CISES-2022)

A Comparison Based Study of Supervised Machine Learning Algorithms for Prediction of Heart Disease

Deepak kumar chohan*
Assistant Professor, Department of Computer Application
Graphic Era Hill University, Dehradun, Uttarakhand, India
dchauhan.geit81@gmail.com

Dinesh C Dobhal
Associate Professor, Department of Computer Application
Graphic Era Hill University, Dehradun, Uttarakhand, India
dineshdobhal@gmail.com

2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) | 978-1-6654-8004-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/CISES54857.2022.9844328

***Abstract*— Recent time's heart disease has been propagating at an intensified rate and has become a large cause of untimely deaths all over the world among non-communicable diseases. South Asian countries in particular seem to have a higher risk of heart disease suggest the studies. It is challenging to predict heart disease, the expertise required to do this is not easy to get and only highly experienced doctors have it. A large amount of data is available to us which can provide us with hidden information, therefore data analysis techniques can be used to make some effective decisions in this area. Not only would this provide us with the reliable way of predicting heart disease, but it would also reduce the pressure on the medical professionals. These algorithms can estimate the likelihood of a person developing heart disease based on factors such as age, gender, blood pressure, stress, and so on. We used a data set with 1025 samples of data including 13 attributes in our research. In this work, we used the Logistic Regression, Decision Tree, SVM, Naive Bayes, Random Forest, and KNN algorithms to predict heart disease, with the purpose of determining which method is the most reliable. In our case, Decision Tree came out at the top with 98.53% accuracy. In the end, we also compared the results of some other study with ours.**

Keywords*— *Classification Algorithm, Heart Disease, Logistic Regression, SVM, Random Forest, Decision Tree, Naïve Bayes, KNN.

I. Introduction

One of the leading causes of death in the globe is heart disease. According to a World Health Organization research, heart disease is responsible for one out of every three fatalities worldwide. Other organs would certainly stop operating if the heart did not function normally. Although the mortality rate of heart disease varies by country, one constant is that the number of patients has been continuously rising over time. Every year, 17.9 million people die from heart disease, accounting for nearly 31% of all fatalities worldwide. [1].

Data mining is a technique for extracting hidden information from data. If this data is appropriately analysed, machine learning algorithms can be used to produce more accurate results that can aid in the prediction of medical diagnoses. This would also help the doctors take correct measures in treating their patients in time.

Heart disease rates increase gradually depending on a person's lifestyle, for example habits like smoking, high fat injection, lacking physical mobility or lack of exercise can increase the chances of a person contracting a heart disease. In a report, US National Institute of health stated that heart rates vary from person to person depending on several parameters but on average range from 60-100 times per minute [3]. This research aims to find the best model for predicting heart disease.

II. RELATED WORK

This section presents some works that we would like to highlight in this study. Quite a bit of research has been done on cardiovascular disease as this affects a considerable amount of people. In their paper, "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction" authors conducted a study in 2019 and discovered that a system trained with Random Forest for diagnosing Heart illness may offer us 90% to 95% accuracy while employing 14 characteristics [4].

In 2018, Uma N Dulhare proposed a method for improving the accuracy of a Bayesian classifier for predicting cardiac disease. They used a data set containing 14 attributes and 270 instances. Their efforts yielded them an accuracy of 87.91% [5]. In [6], the authors did a study in 2018 to predict cardiac disease in a person. The dataset they used had 13 factors. They employed MLP on the UCI library dataset and obtained a precision of about 0.91.

Sonakshi Harjai1 and Sunil Kumar Khatri suggested a new model by using Correlation-based feature selection and a Multilayer Perceptron classifier. They tested 297 inputs and discovered that their model was 89.2% accurate [7]. To identify the heart, in [8] the authors utilised a variety of machine learning techniques before settling on a Support Vector Machine with a linear kernel approach. In [9], the authors suggested a Hybrid Random Forest with a linear model for predicting heart disease in 2019. The dataset for their investigation was obtained from the UCI machine learning library.

III. PROPOSED MODELS

The study addresses the topic of heart disease prediction.

We employ a dataset with 1025 samples and 13 attributes [10]. We used Logistic Regression, Decision Tree, Random Forest, Nave Bayes, KNN, and SVM classification algorithms and assessed their performance in predicting heart disease against each other. We divided our data set into

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on October 18,2022 at 14:21:32 UTC from IEEE Xplore. Restrictions apply.

different sets: training data which consisted 80% of the samples and testing data that had 20% of the samples. The samples were nominated randomly to avoid any bias. The figure below represents the process.

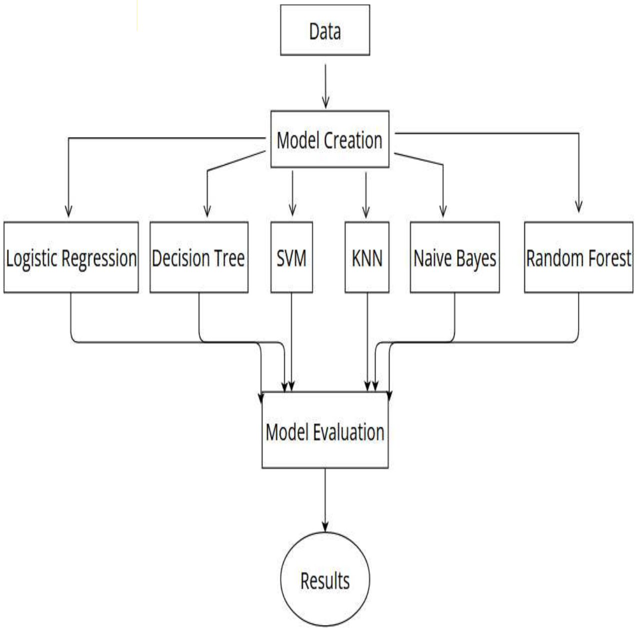

Figure 1. Process Chart

As previously stated, we used these six algorithms to predict an individual potentially having heart disease. Decision Tree produced the best results among the algorithms we tested. With logistic regression after training our model we found the relations between our two sets. In the end we got 78.53% accurate results with it. Naive bayes employs the Bayes theorem and disregards attribute correlation. It can, however, generate a bias for particular characteristics. In our study, we discovered that naive bayes was 80.0% accurate. Support Vector Machine is another algorithm that we used in this study, and with our data set got results accurate up to 80.48%. The Random Forest algorithm is commonly utilised in classification and regression issues. It operates by constructing. So, clearly a proper solution for prevention heart disease is required asap, model that can accurately predict the disease is the next best thing and researchers worldwide are very keen to work towards this aspect. decision trees from many samples and relying on a majority vote for classification and an average in the case of regression.

Random forest was determined to be 87.80% accurate in our testing. Another technique used for classification and regression is K Nearest Neighbor. It is highly versatile and may be used to enter missing values as well as resample datasets. It takes into account K Nearest Neighbors, which are Datapoints, and utilises them to predict the class or continuous value for the new Datapoint, and it proved to be 67.80% accurate with our data set. Finally, we applied the Decision Tree algorithm. It provided an accuracy of 98.53% in our study. In [12] authors defined the basic concept of decision tree. In this work, they discussed various types of decision trees, as well as equations and explanations.

Now we will be describing our dataset. This dataset was downloaded from Kaggle under the name of Dataset for Heart Disease. It contained 1025 samples along with 13 attributes. Below is a description of these attributes.

- **age:** your age
- **sex:** male/female (1 = male, 0 = female)
- **cp:** chest discomfort that is or has been experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- **trestbps:** total resting blood pressure (mm Hg)
- **chol:** milligram per decilitre cholesterol
- **fbs:** fasting blood sugar level (more than 120mg/dl, 1 = true, 0 = false)
- **restecg:** Resting electrocardiographic measurement (0 = normal, 1 = aberrant ST-T wave, 2 = probable or definite left ventricular hypertrophy according to Estes' criteria)
- **thalach:** attained maximal heart rate
- **exang:** Exertional angina (1 = yes, 0 = no)
- **oldpeak:** ST depression caused by activity compared to rest ('ST' refers to places on the ECG plot)
- **slope:** the peak exercise ST segment's slope (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- **ca:** the total number of major vessels (0-3)
- **thal:** thalassemia (blood disorder) (3 = normal, 6 = fixed defect, 7 = normal defect)
- **target:** cardiovascular disease (1 = yes, 0 = no)

We also visualized a heatmap in order to check if there are any features that are highly correlative. We did not find any such values in our dataset.

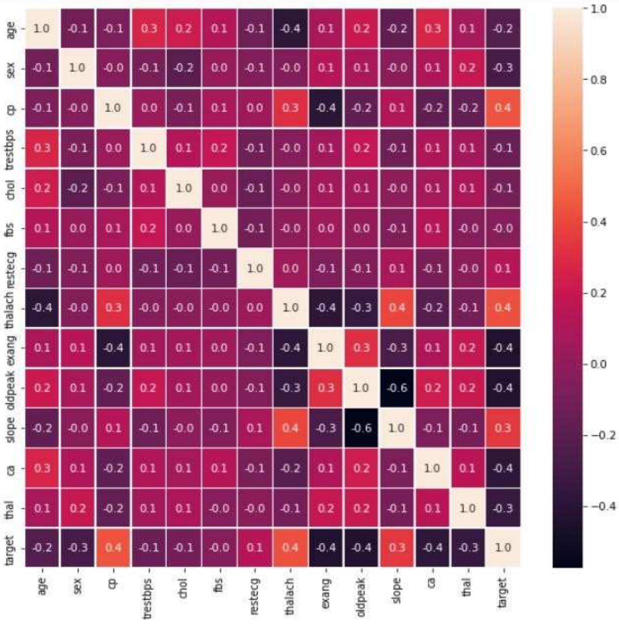

Figure 2. Heatmap

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on October 18,2022 at 14:21:32 UTC from IEEE Xplore. Restrictions apply.

Next, we produced violin plots for certain attributes to check for any outliers in the dataset. The attributes we made the plots for are trestbps (resting blood pressure), chol (cholesterol measurement) and thalach (person's maximum heart rate achieved).

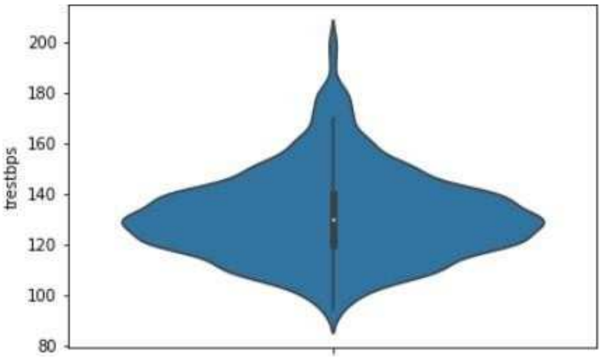

Figure 3. Resting BP Violin Plot

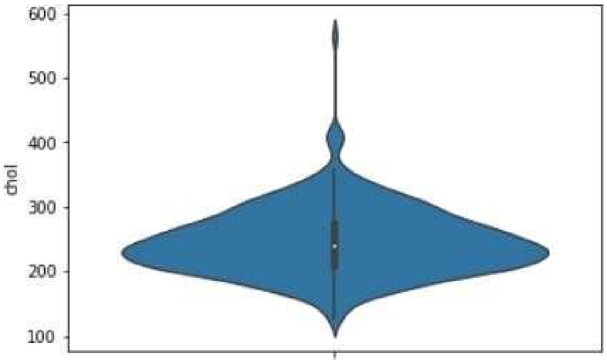

Figure 4. Cholesterol Violin Plot

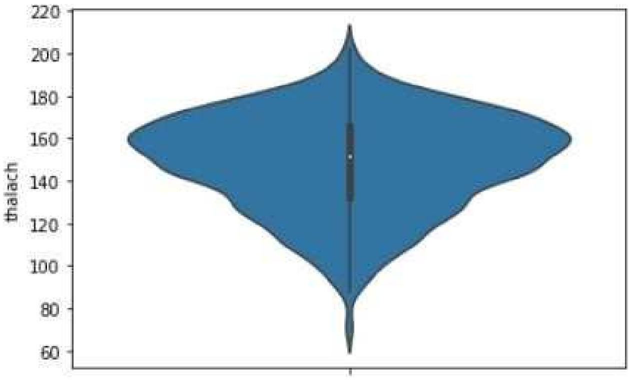

Figure 5. Max Heart Rate Violin Plot

IV. RESULT ANALYSIS

Our prediction system was built with 13 attributes, and the results of our research are displayed in the table below.

Table 1. Comparison of Accuracies of different models in our study

Classification Algorithm	Accuracy
Decision Tree	98.53%
Random Forest	87.80%
Support Vector Machine	80.48%
Naïve Bayes	80.00%
Logistic Regression	78.53%
K Nearest Neighbors	67.80%

V. COMPARATIVE STUDY

Here, we also observed some other works done in this field with similar datasets having the same 13 attributes, Senthil kumar Mohan et al. used UCI heart disease dataset while also applying these classification algorithms. The table below shows the results obtained by them and a comparison of it with our results.

Table 2. UCI Dataset result comparison

Classification Algorithm	Accuracy	Accuracy (Our)
Decision Tree	85%	98.53%
Random Forest	86.1%	87.80%
Support Vector Machine	86.1%	80.48%
Naïve Bayes	75.8%	80.00%
Logistic Regression	82.9%	78.53%

VI. CONCLUSION

Heart disease is one of the top causes of death worldwide, but timely alertness and treatment can help individuals a lot in taking preventive measures before the event of a major incident. We proposed an approach here to help in prediction of the risk of an individual having said disease. We also compared our results with that of another study that used UCI dataset on heart disease. Moving forward our wish is to be able to generate predictions for this disease with higher accuracy and to achieve this, data collection on a large scale is required. We believe that everyone deserves to live a good life in good health for the betterment of society.

REFERENCES

[1] "Cardiovascular diseases" Available: https://www.who.int/en/newsroom/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: 25-January- 2020]

[2] Wu, Ching-seh Mike, Mustafa Badshah, and Vishwa Bhagwat, "Heart Disease Prediction Using Data Mining Techniques." In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, pp. 7- 11. 2019.

[3] "What should my heart rate be?" https://www.medicalnewstoday.com/articles/235710.php#normal resting-heart-rate. Accessed: 21- January- 2020]

[4] N. Satish Chandra Reddy, Song Shue Nee, Lim Zhi Min & Chew Xin Ying "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction", International Journal of Innovative Computing, 2019, DOI: https://doi.org/10.11113/ijic.v9n1.210.

[5] Duraipandian, M. "Performance Evaluation of Routing Algorithm for MANET based on the Machine Learning Techniques." Journal of trends in Computer Science and Smart technology (TCSST) 1, no. 01 (2019): 25-38.

[6] Aditi Gavhane, Gouthami Kokkula, Isha Pandya & Prof. Kailas Devadkar (PhD) "Prediction of Heart Disease Using Machine Learning", ICECA 2018, IEEE Xplore ISBN:978-1- 5386-0965-1.

[7] Sonakshi Harjai1 & Sunil Kumar Khatri, "An Intelligent Clinical Decision Support System Based on Artificial Neural Network for Early Diagnosis of Cardiovascular Diseases in Rural Areas", AICAI, 2019, DOI: 10.1109/AICAI.2019.8701237.

[8] Nabaouia Louridi, Meryem Amar, Bouabid El Ouahidi "IDENTIFICATION OF CARDIOVASCULAR DISEASES USING MACHINE LEARNING", 7th Mediterranean Congress of

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on October 18,2022 at 14:21:32 UTC from IEEE Xplore. Restrictions apply.

Telecommunications (CMT), 2019, DOI: 10.1109/CMT.2019.8931411.

[9] Senthilkumar Mohan, Chandrasegar Thirumalai And Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", Special Section On Smart Caching, Communications, Computing and Cybersecurity For Information-centric Internet Of Things, IEEE Access Volume 7, 2019, DOI:10.1109/ACCESS.2019.2923707.

[10] "Dataset" Available:https://www.kaggle.com/johnsmith88/heart-disease-dataset

[11] Suthaharan S., "Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification". Integrated Series in Information Systems, vol 36. Springer, Boston, MA, 2016.

[12] Arundhati Navada, Aamir Nizam Ansari, Siddharth Patil, Balwant A.Sonkamble, "Overview of Use of Decision Tree algorithms in Machine Learning". IIEEE Control and System Graduate Research Colloquium, 2011.

[13] Gavhane, A., 2018. Prediction of Heart Disease Using Machine Learning. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), (Iceca), pp.1275−1278.

[14] R, T.P. Thomas, J., 2016. Human Heart Disease Prediction System using Data Mining Techniques

[15] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[16] Sheik A Abdullah and R R Rajalaxmi. Article: A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls (ICON3C 2012) ICON3C(3):22- 25, April 2012.

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on October 18,2022 at 14:21:32 UTC from IEEE Xplore. Restrictions apply.