2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) | 978-1-6654-3215-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/PARC52418.2022.9726613

Effective Study of Machine Learning Algorithms for Heart Disease Prediction

Mihir J. Gaikwad
Department of Information Technology,
G H Raisoni College of Engineering,
Nagpur, India- 440016
gaikwad_mihir.it@ghrce.raisoni.net

Prathmesh S. Asole
Department of Information Technology,
G H Raisoni College of Engineering,
Nagpur, India- 440016
asole_prathmesh.it@ghrce.raisoni.net

Prof. Leela S. Bitla
Assistant Prof., Department of
Information Technology,
G H Raisoni College of Engineering,
Nagpur, India- 440016
leela.bitla@raisoni.net

***Abstract*—Heart disease has been a major public health concern in recent years, excessive alcohol consumption, cigarette, and a sedentary lifestyle are the primary factors, and it is the leading cause of mortality among patients. Medically, heart disease is known for being difficult to forecast, detect, and diagnose. To treat heart diseases, hospitals and other clinics are giving costly therapies and treatments. According to a recent WHO research, heart disease is on the rise. In 2019, 17.9 million people die as a result of this. It becomes more difficult to diagnose as the population grows. As a result, detecting cardiac disease early on will benefit people all across the world, allowing them to receive necessary therapy before it becomes critical. Thanks to recent technical breakthroughs, machine learning has shown to be effective in making decisions and predictions from a big set of data provided by the health-care sector. In this paper, some of the supervised machine learning techniques used in this prediction of heart disease which are Support Vector Machines (SVMs), Gradient Boosting Classifier (GB), Decision tree (DT), Random forest (RF), Logistic Regression (LR) on the "UCI Machine learning repository for Statlog (Heart) Data Set". Furthermore, the findings of these algorithms are reported, and a proposal is made to employ the algorithm with the highest accuracy for predicting Heart Disease on a web application. This application will be used as a decision support system by medical practitioners in their clinics as well as people at home.**

Keywords—Heart Disease, Machine Learning, Support Vector Machines (SVMs), Gradient Boosting Classifier (GB), Decision tree (DT), Random forest (RF), Logistic Regression (LR), UCI Machine learning repository, Classification, Web Application.

I. INTRODUCTION

This part outlines the primary characteristics and reasons that led to the creation of this work, as well as the major contributions it makes. As a result, we provide an insight of how the concept of prior knowledge can be implemented with machine learning and in the context of web development.

As per the World Health Organization, heart disease remains the major cause of mortality in the world, resulting for 17.9 million deaths in 2019 [1]. The disease demands special attention since it is one of the largest and most important organs in the human body. Because the majority of illnesses are connected to the heart, it's critical to predict heart disease, necessitating a comparative research on the topic. Medical diagnosis has been difficult due to the many factors that can cause heart disease [2], and because 20-40% of heart attacks occur in patients who have previously been unable to diagnose their diseases due to a lack of instrument accuracy [3], there is a need to learn about more efficient disease prediction algorithms [4.] As a result, developing an ideal machine-learning model for heart-disease prediction to aid diagnosis is critical and medically significant, and various researchers have attempted to achieve pinpoint model accuracy in the past [5].

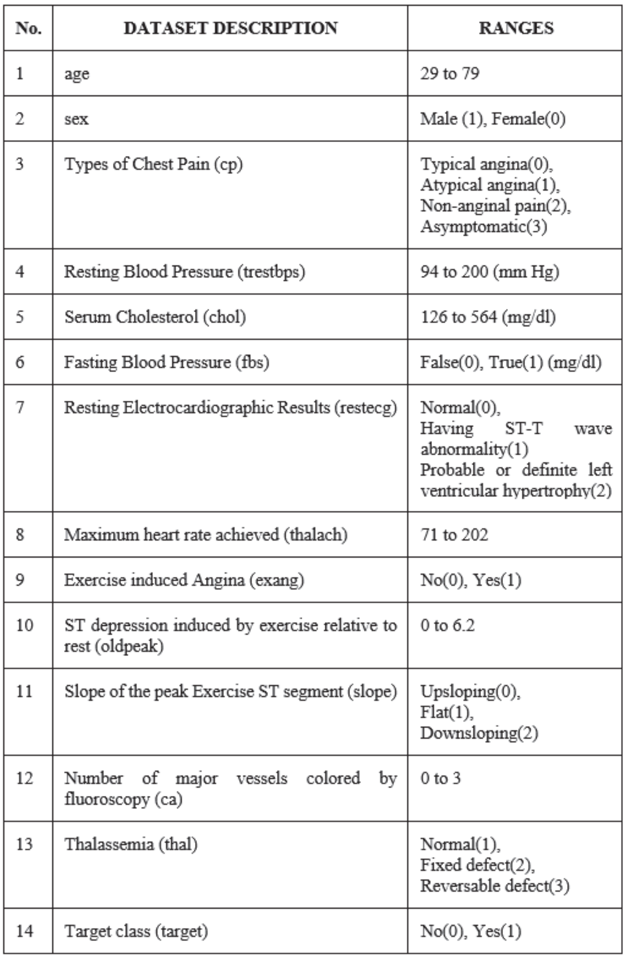
Heart disease is one of the top causes of critical mortality rates throughout the world, and many people rely on the healthcare system to deliver accurate data in a timely manner. The healthcare facility creates and gathers a significant amount of data on a daily basis. Data innovation enables for the automated extraction of data, resulting in unique discoveries. As soon as possible, heart illness should be diagnosed. Many visual representation and machine learning methods have been developed to achieve this goal. Machine learning may also be used to analyse the most important factor in heart disease.

Machine learning is becoming increasingly popular in the medical diagnosis field, where analysis may minimize manual error and enhance accuracy. With machine learning algorithms, a disease's diagnosis is quite accurate. We proposed some algorithms that can aid in the early identification of patients with heart diseases, using attributes such as age, serum cholesterol, fasting blood sugar, chest pain type, sex, and so on. We measured the accuracy of five different machine learning algorithms based on these findings. We also determine which of them is the best. Furthermore, by preserving the best model in a pickle file, we use it for prediction on the web application. The use of machine learning techniques in the medical business is crucial since diseases can be predicted at an early stage, saving money on medicine, improving people's health, predicting mainly correct outcomes, and improving healthcare value while also saving lives.

Clinical choices are frequently made on the basis of a doctor's intuition and experience rather than professional proof. As a result, there are more errors and expenditures, as well as a decrease in the quality of medical services. Doctors may make better clinical judgments with the support of analytic tools and data modelling. As a result, we've created a website to assist professionals in diagnosing heart diseases.

The heart-disease dataset is hosted on the UCI (University of California, Irvine) Machine Learning Repository, and it contains 270 patient records with 14 unique features, which are described in Table I.

TABLE I. DESCRIPTION OF DATASET ATTRIBUTES



No.	DATASET DESCRIPTION	RANGES
1	age	29 to 79
2	sex	Male (1), Female(0)
3	Types of Chest Pain (cp)	Typical angina(0), Atypical angina(1), Non-anginal pain(2), Asymptomatic(3)
4	Resting Blood Pressure (trestbps)	94 to 200 (mm Hg)
5	Serum Cholesterol (chol)	126 to 564 (mg/dl)
6	Fasting Blood Pressure (fbs)	False(0), True(1) (mg/dl)
7	Resting Electrocardiographic Results (restecg)	Normal(0), Having ST-T wave abnormality(1) Probable or definite left ventricular hypertrophy(2)
8	Maximum heart rate achieved (thalach)	71 to 202
9	Exercise induced Angina (exang)	No(0), Yes(1)
10	ST depression induced by exercise relative to rest (oldpeak)	0 to 6.2
11	Slope of the peak Exercise ST segment (slope)	Upsloping(0), Flat(1), Downsloping(2)
12	Number of major vessels colored by fluoroscopy (ca)	0 to 3
13	Thalassemia (thal)	Normal(1), Fixed defect(2), Reversable defect(3)
14	Target class (target)	No(0), Yes(1)

Section 2 of the following chapters provides a quick review of the paper's background knowledge in the subject of machine learning. Furthermore, the prediction system, algorithms, Block diagram for proposed model, heat map, correlations, and pair plot are all included in section 3. In chapter 4, the accuracy results were shown, as well as the output from the web application. Finally, section 5 brings this paper's work to a conclusion.

II. Related Work

Nowadays, big data analytics, particularly healthcare analytics, has become a hot topic in a slew of studies. Many recent studies on heart disease prediction and analysis have been conducted by researchers. Some of these works are discussed farther down.

Using a machine learning algorithm to predict heart disease has been the subject of several academic articles. As seen by positive performance on the UCI dataset set [6] the use of train test split confirmation accompanied by logistic regression for diagnosis boosted the accuracy of heart disease prediction. Another study looked at the use of automated machine learning to predict cardiovascular disease risk. [7] To characterize this technique, they coined the term "autoprognosis." This algorithm chooses and fine-tunes Machine Learning models on your behalf. Using data from 423,604 persons, the model was put to the test. The results were compared to the ’Framingham Score, a well-known measure of risk prediction. The Autoprognosis model, with a 95 percent accuracy rate, outperformed the Framingham Score in terms of prediction. Triglycerides, inflammatory markers, and epinephrine (adrenaline were not included in the prognostication process.

A work that would use machine learning to assess CVD in diabetes patients [8] was published. With datasets from Italy and the United States, several Machine Learning approaches were put to the test. The Support Vector Machine (SVM) with RBF Kernel method generated the highest returns in both the Italian and American datasets, yielding 95.25 percentage points in the Italian dataset and 92.15 percent efficiency in the American dataset. The Italian dataset’s tilt, on the other hand, might affect forecast accuracy. To predict heart illness, the authors of [9] created a hybrid intelligence system. The system used classification techniques such as logistic regression, artificial intelligence, neural networks (ANN), SVM, KNN, decision trees, Naive Bayes & random forest. To improve prediction efficiency, relief, MRMR, and LASSO, three feature selection approaches, were used. These algorithms choose strongly linked traits that have a significant impact on target variable. Logistic Regression plus 10-fold validation generated 89 recognition rate using the Relief component selection approach, according to the findings. This model’s output might be improved by using neural network-based optimization approaches. [10] Proposed a support vector machine- based expert system (SVM). The first SVM is used to eliminate superfluous characteristics, while the second is utilized to forecast. They also used the HGSA to improve the two methods (hybrid grid search algorithm). By using this model, they were able to enhance accuracy by 3.3 percent over earlier SVM models.

The concept of multi-level risk evaluation was presented by the author in [11]. Smoking, a low physical activity, and fatness were added to the list of risk factors. The Decision Tree approach accurately predicted heart failure risk with an accuracy of 86.53 percent. The introduction of enhanced feature selection procedures to the model’s conclusion result may improve the model’s output. Many studies have been conducted on the classification accuracies of various machine learning approaches using the Cleveland heart disease database, which is publicly available via the University of California’s online data mining repository. The authors of [12] were able to achieve a prediction accuracy of 77 percent on this dataset using the logistic regression technique. The authors of [13] refined their work by comparing global evolutionary computing approaches and found that prediction accuracy increased as a consequence. The author of another work [14] proposed a study utilizing ML approaches to diagnosis diabetic malady. This illness was formerly believed to be a significant ML focus. Diabetes affects nearly 285 million people globally according to a survey done by the international diabetes federation (IDF).

Using the UCI machine learning data set which has 303 samples with 14 input characteristics, the author [15] investigated several machine learning and data processing methods and determined that SVMs is the best. Naive Bayes, KNN, and Decision Tree were among the other algorithms examined. The multilayer perceptron model for human cardiovascular disease prediction, as well as the accuracy of the algorithm employing CAD technology, were investigated in [16]. If more people utilize the prediction approach to anticipate their illness, disease awareness will increase, and the death rate of heart patient would drop. One or two illness prediction algorithms are being tested by certain researchers.

Comprehensive review of machine learning techniques for breast prediction and diagnosis [17]. On the Breast Cancer Wisconsin (Diagnostic) Dataset, trained and evaluated a method for identifying breast cancer using logistic regression and decision tree by the authors (Prateek P. Sengar, Mihir J. Gaikwad, Ashlesha S. Nagdive). Accuracy, recall, and precision are performance matrices that can be calculated. LR and DT accuracy were 94.40 percent and 95.10 percent, respectively, in the final result.

III. METHODOLOGY

The goal of the project is to create a machine learning-based heart disease analysis system that may be used to speed up the diagnosis process in hospitals. Our platform provides the detection capability over the Heart Disease. This section aims to focus upon the methodology employed throughout our experiments and development procedure.

Using cutting-edge technology, the risk of heart disease in humans can be predicted much more accurately and effectively. Machine learning will be used to anticipate heart problems. There are approximately 270 data entries (rows) and 14 attributes in the dataset (columns). Using multivariate regression, our system will be taught to predict people with and without heart disease. We'll use a total of five algorithms to converge the attributes so that our machine can predict faster: Support Vector Machine, Random Forest, Gradient Boosting, Decision Tree, and Logistic Regression Technique. Our data will be used to train our machine 75% of the time (202 records), while the remaining 25% (68 records) will be used to test it, confirming that our computer correctly predicts the type of cancer. Machine learning packages such as scikit-learn, numpy, pandas, matplotlib.pyplot, seaborn, and others will be used to write all of our algorithms in Python.

The steps in Figure.1. will be performed to get our machine to forecast whether or not the patient has cancer:

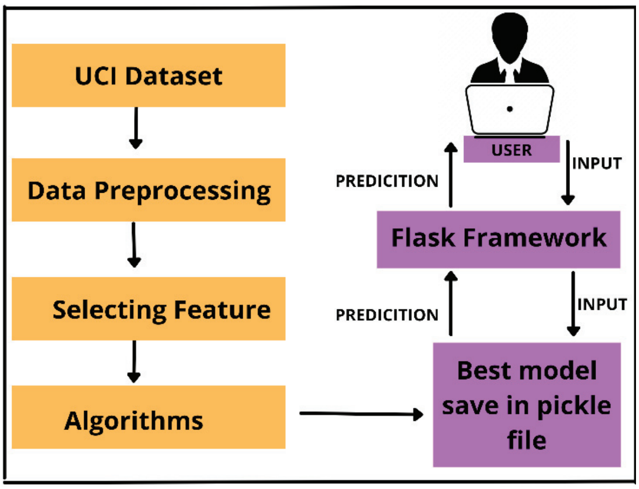


Fig. 1. Block diagram for proposed model

There are 5 algorithms used in our project:

- Support Vector Machines (SVMs)
- Random Forest (RF)
- Gradient Boosting Classifier
- Logistic Regression
- Decision Tree

A. Affected and Normal Heart disease comparison:

The ratio of people with and without heart disease was calculated, and it was discovered that 44 percent of the people in the sample have heart disease which is slightly lower in terms of people who do not have heart disease that is 56 percent which can be seen at Figure.2. As a result, the dataset is fairly balanced.

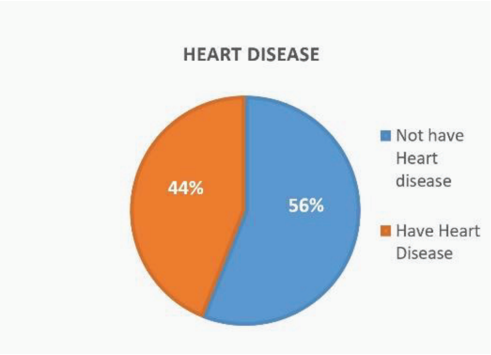


Fig. 2. Pie Chart of people with and without heart disease in the dataset

B. Heat map:

The heat map is a useful tool for visualizing how values are concentrated in two dimensions of a matrix. This enhances pattern detection and gives the impression of depth. To utilize a heat map, the data must be in a matrix format. In order for the data we enter into the cells to be relevant, the index name and column name must be comparable in some way. The Heat Map helps see correlations and offers information on how other variables (columns) affect the diagnosis column which has shown in Figure.3.

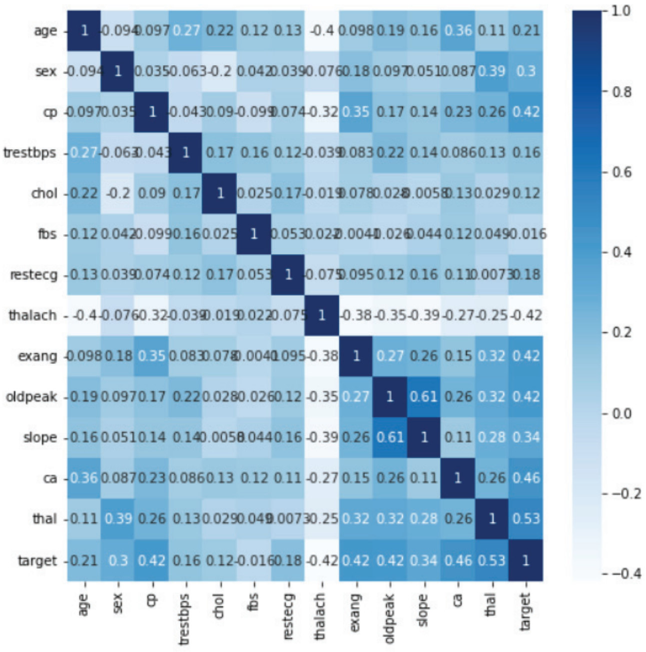


Fig. 3. Heat map

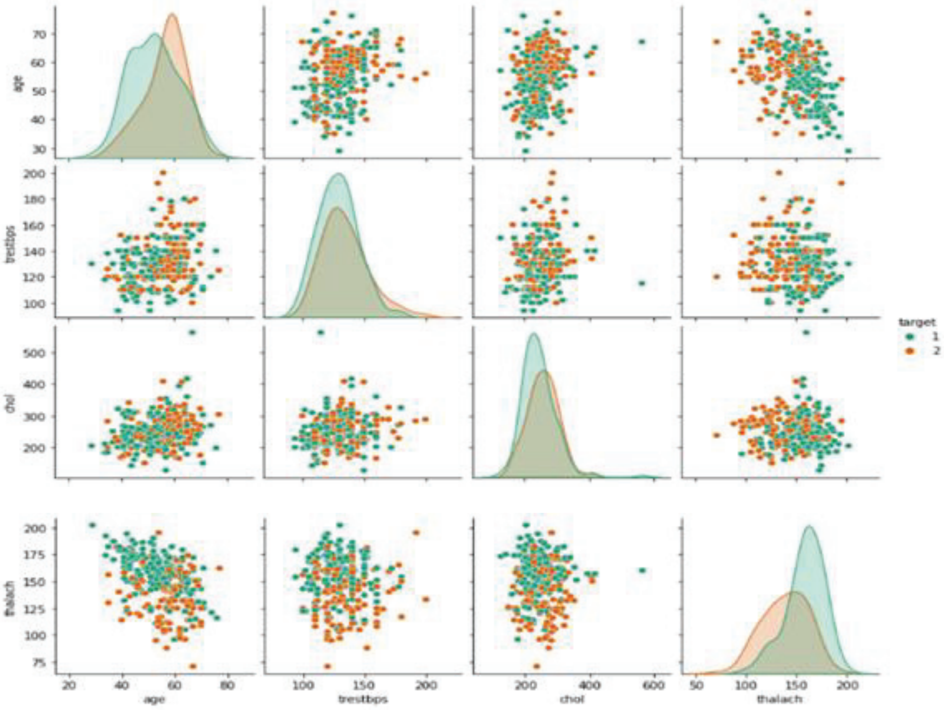


Fig. 4. Pair Plot Diagram

C. Pair Plot of diagnosis with other features:

The pair plot clearly depicts the number of healthy and diseased cells associated with the attributes. This allows us to see which characteristics to consider when predicting outcomes. There are 270 records, 150 of which do not have heart disease (Target = 1) and 120 of which have heart disease (Target = 2). In our data set, the healthy and afflicted disease comparison is shown in the Figure.4.

D. Correlation with other features:

Correlation is a measure of how two variables vary over time. A correlation matrix can be plotted to illustrate whether variable has a high or low correlation with another one. Correlation is a measure of the stiffness of a continuous relation between two categorical variables. Correlation is a normalized version of covariance that describes the relationship between two variables. Correlation coefficients are always in the range of -1 to 1. Pearson's correlation coefficient is another name for the correlation coefficient.

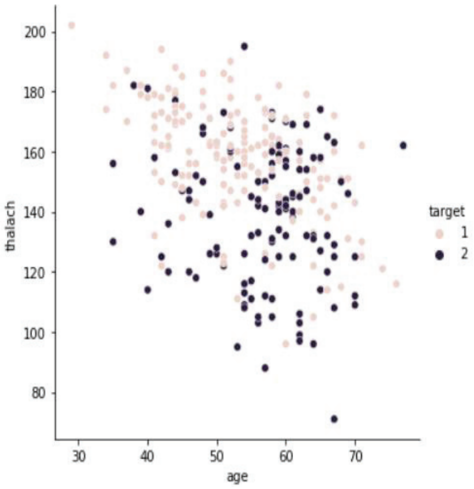


Fig. 5. Correlation between age and heart rate

In Figure.5. We can see Heart disease is more common in the elderly, and the maximum heart rates are lower in the elderly with heart disease.

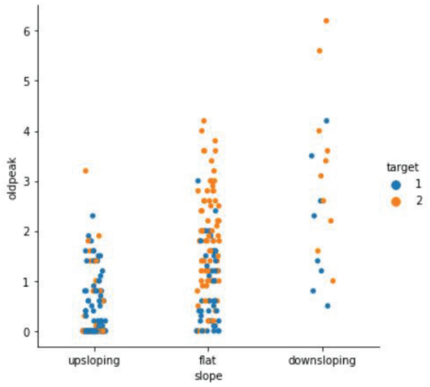


Fig. 6. Correlation between oldpeak and slope

In Figure.6. People with a downsloping ST segment have higher levels of ST depression and are more likely to develop heart disease. The bigger the ST depression, the more likely you are to have an illness.

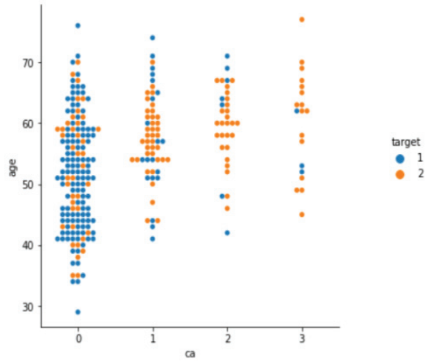
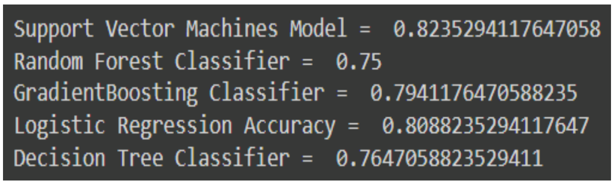


Fig. 7. Correlation between number of major vessels colored by fluoroscopy and age

From the Figure.7. We can conclude the majority of heart disease patients are elderly and have Fluoroscopy in one or more main vessels.

E. Accuracies:

These five algorithms (Support Vector Machine, Decision Tree Classifier, Random Forest, Logistic Regression and Gradient Boosting) are compared because they were all predicted to generate high-accuracy predictions, but we needed one that was slightly more accurate than the others, so the Support Vector Machine was chosen because it had slightly more accuracy which is 82.35% than the Random Forest, Gradient Boosting, Logistic Regression, and Decision Tree Classifier. The following Figure.8. depicts the accuracy of all of the models' predictions and help us in deciding which one to utilize.



```
Support Vector Machines Model =  0.8235294117647058
Random Forest Classifier =  0.75
GradientBoosting Classifier =  0.7941176470588235
Logistic Regression Accuracy =  0.8088235294117647
Decision Tree Classifier =  0.7647058823529411
```

Fig. 8. Accuracies of model

By using the Flask web application framework to deploy and construct the online web application. Firstly, we saved the SVMs machine learning model in a pickle file that we train and test on the complete dataset, obtain the user input from the HTML template, perform the prediction, and return the outcome in web application. In addition, a front-end HTML template that allows the user to input the patient's heart disease symptoms shown in Figure.9 and identify if the victim exhibits heart disease or not.

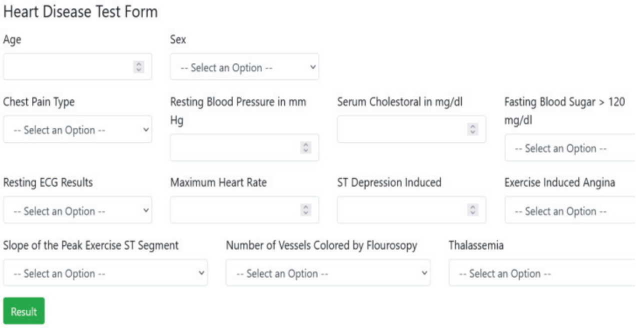
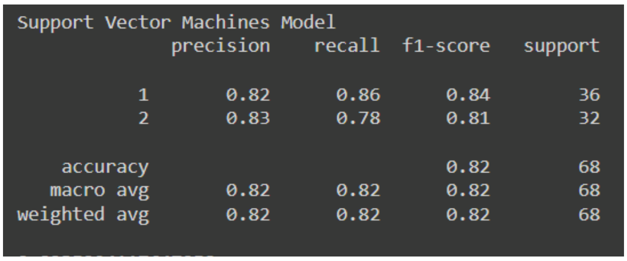


Fig. 9. User Input form for Heart Disease prediction

IV. Result

The model's efficiency is evaluated using the Confusion Matrix, Precision, Recall, and F1 scores, shown in Figure.10. The following is the mathematical formula for all metrics:



```
Support Vector Machines Model
              precision    recall  f1-score   support

           1       0.82      0.86      0.84        36
           2       0.83      0.78      0.81        32

    accuracy                           0.82        68
   macro avg       0.82      0.82      0.82        68
weighted avg       0.82      0.82      0.82        68
```

Fig. 10. Performance Metrics

A. Confusion Matrix:

It is among the most frequent matrices for measuring the classification model's effectiveness. It gives a detailed account of the actual and anticipated outcome. The frequency of correct and wrong predictions is represented by n×n matrix.

B. Accuracy:

It indicates the model's precision. It's a proportion of total true prediction to total prediction.

$$\text{Accuracy} \;=\frac{\text{TP}+TN}{TP+FN+FP+FN}$$

C. Precision:

Precision depicts the proportion of the relevant result that is relevant. It's a proportion of genuine positives to total predicted positives.

$$\text{Precision} \;=\frac{TP}{TP+FP}$$

D. Recall:

The percentage of accurate results that are correctly classified is shown by recall. It's a proportion of genuine positives to overall positives.

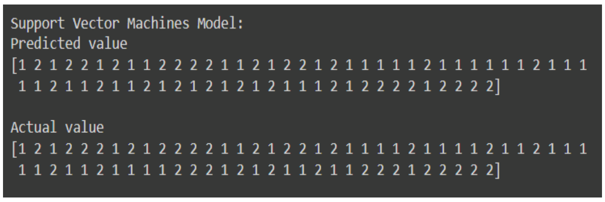
$$\text{Recall} \;=\frac{TP}{TP+FN}$$

E. F1 Score:

It's a combination of accuracy and recall that's been weighted. It is much more advantageous if the dataset is unequal.

$$\text{F1 Score} \;=2*\frac{\text{Precision}\,^{*}\,\text{Recall}}{\text{Precision}\;+\;\text{Recall}}$$

To see if the model is behaving as expected, compare the actual test value to the value predicted by Support Vector Machines (SVMs) in Figure 11.

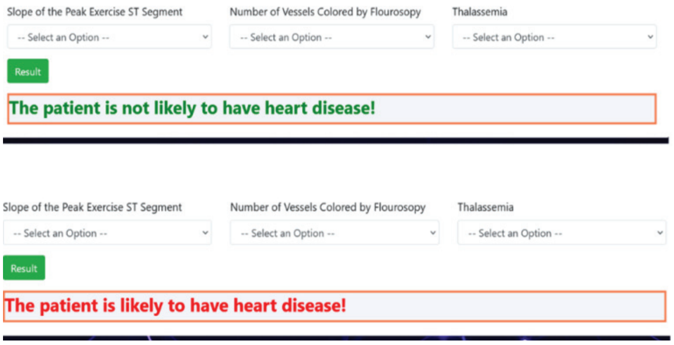


```
Support Vector Machines Model:
Predicted value
[1 2 1 2 2 1 2 1 1 2 2 2 2 1 1 2 1 2 2 1 2 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1
 1 1 2 1 1 2 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 2 2 2 2 1 2 2 2 2]

Actual value
[1 2 1 2 2 2 1 2 1 2 2 2 2 1 1 2 1 2 2 1 2 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 1
 1 1 2 1 1 2 1 1 1 1 2 2 2 1 2 1 2 1 1 2 1 1 2 2 2 1 2 2 2 2 2]
```

Fig. 11. Test values and Predicted values

In the process of learning, we were able to build a web application from the ground up to forecast heart disease. We began by defining the problem and gathering information. Then we worked on data preparation, exploration, modelling, and evaluation of the models. Finally, we used a flask to deploy the model. So that a person can enter their heart disease feature information into a web application and receive a predicted result about whether or not they have heart disease shown in Figure.12.



The patient is not likely to have heart disease!

The patient is likely to have heart disease!

Fig. 12. Predicted output on Web application

V. CONCLUSION

In this Research paper, to forecast heart disease, five machine learning algorithms are applied (Support Vector Machine, Random Forest, Gradient Boosting, Logistic Regression, and Decision Tree Classifier).The prediction of each algorithm is compared to determine which one is best suited for the prediction. The best algorithm for prediction is the Support Vector Machine, which has a pinpoint prediction accuracy [on "UCI Machine learning repository for Statlog (Heart) Data Set"]. As a result, combining our Support Vector Machine method and the attributes of this dataset, heart disease may be predicted with near-perfect accuracy. Hence, we have built a web application that predict whether a person is having a heart disease or not using Support Vector Machine Algorithm.

ACKNOWLEDGMENT

The authors appreciate the conference's financial support from the GHRCE's Research and Development unit.

REFERENCES

[1] World Health Organization (2017) Cardiovascular diseases (CVDs). Available: https://www.who.int/news-room/fact-sheets/detail/cardiovasculardiseases- (cvds).

[2] Alqudah, A. M. (2017). Fuzzy expert system for coronary heart disease diagnosis in Jordan. Health Technol, 7: 215–222.

[3] McClellan, M., Brown, N., Califf, R. M., Warner, J. J. (2019) Call to Action: Urgent Challenges in Cardiovascular Disease: A Presidential Advisory From the American Heart Association. Circulation, 139: e44– e54.

[4] Archana Singh, Rakesh Kumar. Heart Disease Prediction Using Machine Learning Algorithms. 2020 International Conference on Electrical and Electronics Engineering (ICE3). IEEE. 2020.

[5] Pouriyeh S., Vahid S., Sannino G., Pietro G. D., 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). Heraklion. pp. 204-207.

[6] R. Kannan and V. Vasanthi. Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease in Soft Computing and Medical Bioinformatics. Springer Singapore, June 2018, pp. 63–72.

[7] Ahmed M. AlaaI, Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, Mihaela van der Schaar. Cardiovascular disease risk predict ion using automated machine learning: A prospective study of 423,604 UK Biobank participants. PloS One 14 (5): e0213653, May 2019.

[8] Sabrina Mezzatesta, Claudia Torino, Pasquale DeMeo, Giacomo Fiumara, Antonio Vilasi. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Computer Methods and Programs in Biomedicine, Elsevier, vol. 177, pp. 9-15, August 2019.

[9] Amin UlHaq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir and Ruinan Sun. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Hindawi, Mobile Information Systems, vol. 2018, pp. 1 -15, December 2018.

[10] L. Ali et al. An Optimized Stacked Support Vector Machines Based Expert System for the Effective Predict ion of Heart Failure. IEEE Access, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[11] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain,T. Dawson, P Fergus and M. Al-Jumaily. Predicting the Likelihood of Heart Failure with a Multi-Level Risk Assessment Using Decision Tree. 2015 Third International Conference on Technological 60 Advances in Electrical, Electronics and Computer Engineering, IEEE, pp. 101 -106, June 2015.

[12] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid,S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, vol. 64, no. 5, pp. 304–310, 1989.

[13] B. Edmonds. Using localised 'gossip' to structure distributed learning. 2005.

[14] Fsd Bayu Adhi Tama, 1 Afriyan Firdaus, 2 Rodiyatul FS. Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine. Vol. 11, issue 3, pp. 12-23, 2008.

[15] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak. Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT 2019.

[16] Santhana Krishnan J and Geetha S. Prediction of Heart Disease using Machine Learning Algorithms. ICIICT, 2019.

[17] P. P. Sengar, M. J. Gaikwad and A. S. Nagdive. Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE. 2020.