International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Machine Learning Based Heart Disease Prediction Using Random Forest

S. Swarna[1], G. Anil[2], P. Shivamani[3], K. Madhu Babu[4]

1 UG student, Dept. of Electronics and Computers Engineering, Sreenidhi Institute of Science and Technologies, Telangana, India

2 UG student, Dept. of Electronics and Computers Engineering, Sreenidhi Institute of Science and Technologies, Telangana, India

3 UG student, Dept. of Electronics and Computers Engineering, Sreenidhi Institute of Science and Technologies, Telangana, India

4 Asst. Professor, Dept. of Electronics and Computers Engineering, Sreenidhi Institute of Science and Technologies, Telangana, India

Abstract - A recent study by the World Health Organization sheds light on the alarming increase in cardiovascular diseases, contributing to approximately 17.9 million deaths annually. This study delves into the effectiveness of employing the Random Forest algorithm, a robust machine learning approach, to forecast the likelihood of heart disease based on diverse risk factors. By leveraging a dataset encompassing demographic, clinical, and lifestyle attributes, the Random Forest model underwent training to categorize individuals into two groups: those with or without heart disease. Through meticulous feature selection and ensemble learning, the algorithm adeptly captures intricate relationships among predictors, thereby augmenting prediction accuracy. Evaluation metrics including accuracy and AUC-ROC curve were employed in order to determine model's effectiveness. Impressively, our model achieves a prediction accuracy of 97%. Moreover, a comparative analysis with other prominent machine learning models such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression (LR), XGBoost, Decision Tree revealed that the Random Forest approach outperforms others in terms of accuracy and efficiency in prediction tasks.

Keywords: Random Forest (RF), Machine Learning (ML), Accuracy,

Classification.

1.INTRODUCTION

Early detection and prediction of heart disease risk can lead to better management and prevention strategies, ultimately improving patient outcomes, ML techniques have shown promise in accurately predicting heart disease risk by analysing various patient attributes and medical data. In this project [2]. By leveraging data-driven techniques, we can create a predictive model that offers valuable insights into an individual's cardiovascular health, allowing for early intervention and personalized treatment strategies. This endeavor holds the potential to enhance healthcare consequences and alleviating the impact of heart disease on both individuals and community at large.[4]

1.1 Reason behind exclusively selecting Random Forests.

- It exhibits a shorter processing duration in contrast to alternative algorithms.[4]
- Demonstrates remarkable precision in forecasting results, efficiently handling even extensive datasets.
- It possesses the capability to manage absent data entries within the dataset without necessitating imputation procedures.

- Reducing the risk of overfitting.[2]

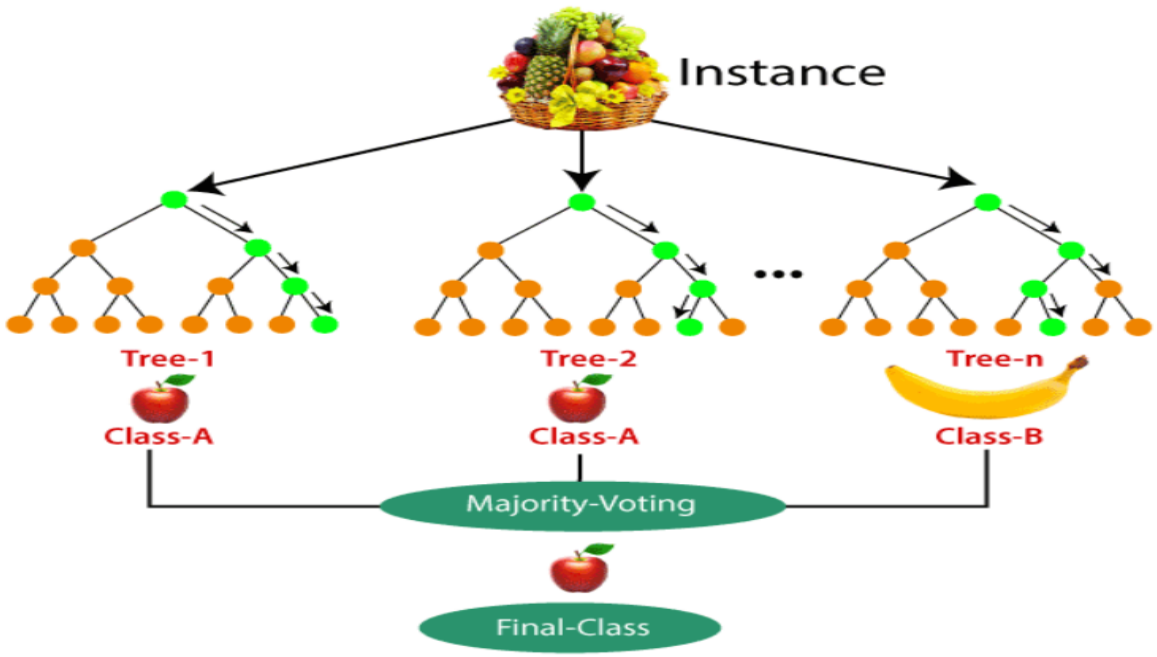


Fig 1: Random Forest Algorithm

2. LITERATURE SURVEY

In their 2020 IEEE paper titled "Heart Disease Prediction using Machine Learning Techniques," Vijeta Sharma, Shrinkhala Yadav, and Manjari Gupta aimed to uncover the interconnections among various attributes within the dataset to effectively forecast the likelihood of heart disease. Their study compared several supervised machine learning algorithms, to ascertain the most accurate predictor. Among these methods, Random Forest emerged as the most promising, offering enhanced accuracy in predicting heart disease risk. The findings suggest that integrating RF into clinical settings could empower medical practitioners with a reliable decision support system, potentially improving patient care and outcomes.[2]

M. Snehith Raja, M. Anurag, Ch. Prachetan Reddy (2021 IEEE) "Machine Learning Based Heart Disease Prediction System" found that RF model attaining exceptional performance and accuracy levels due to its adaptability, thus achieving high success rates. [1]

Muntasir Mamun, Md. Milon Uddin, Vivek Kumar (2022 IEEE) "MLHEARTDIS: Can Machine Learning Techniques Enable to Predict Heart Diseases?" proposed six ML models using survey data of over 400k US residents from the year 2020.Hence, they reached an elevated level of performance, boasting an accuracy rate of 91.57% with the Logistic Regression model.[4]

Mohammed Ali Shaik, Radhandi Sreeja, Safa Zainab (2023 IEEE) "Improving Accuracy of Heart Disease Prediction through Machine Learning Algorithms" incorporated four diverse components, each addressing distinct factors pertinent to heart diseases and results better prediction with Random Forest (RF) and K-Nearest Neighbour (KNN).[3]

3.EXISTING METHOD

The existing system adopts a Hybrid algorithm approach, which amalgamates the strengths of multiple algorithms to enhance overall performance. This technique integrates various methodologies such as ensemble methods, meta-learning, or blending diverse model types like neural networks and decision trees[3]. By leveraging the unique advantages of each constituent algorithm, the goal is to achieve superior results compared to any individual algorithm in isolation.

The dataset utilized in this system comprises real-time data collected from individuals, undergoing preprocessing steps such as elimination of duplicate and null records through feature selection[4]. The Hybrid algorithm is then combined with user input to train the model, enabling it to recognize patterns and iteratively learn from the data by adjusting internal parameters and refining outputs. This iterative learning process continues until the algorithm attains satisfactory performance on the training data.

Once trained, model gains the ability to forecast outcomes on unfamiliar data, thereby providing predicted outputs based on learned patterns and insights gleaned from the training phase.

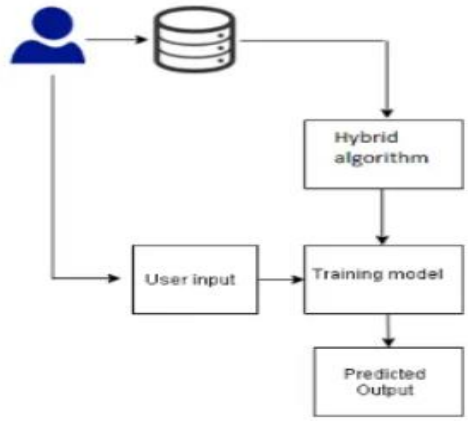


Fig 2: Block Diagram

3.1 Drawbacks:

- Interpretation of Result
- Data Acquisition

4. PROPOSED METHOD

The proposed uses the Random Forests (RF) algorithm**,** this classifier comprises multiple decision trees trained on different portions of the dataset, leveraging their collective predictions to enhance overall accuracy. Rather than depending on a single tree, the random forest algorithm combines the predictions from each tree, ultimately determining the final output based on the majority consensus[2]. This approach reduces overfitting peril and enhances the model's resilience to noisy data. Furthermore, the method effectively uses categorical and numerical data, and it can manage missing data[3]. By analysing a dataset containing various features related to a person's health, such as age, blood pressure, cholesterol levels and more, can learn patterns and make predictions about the likelihood of someone having heart disease.

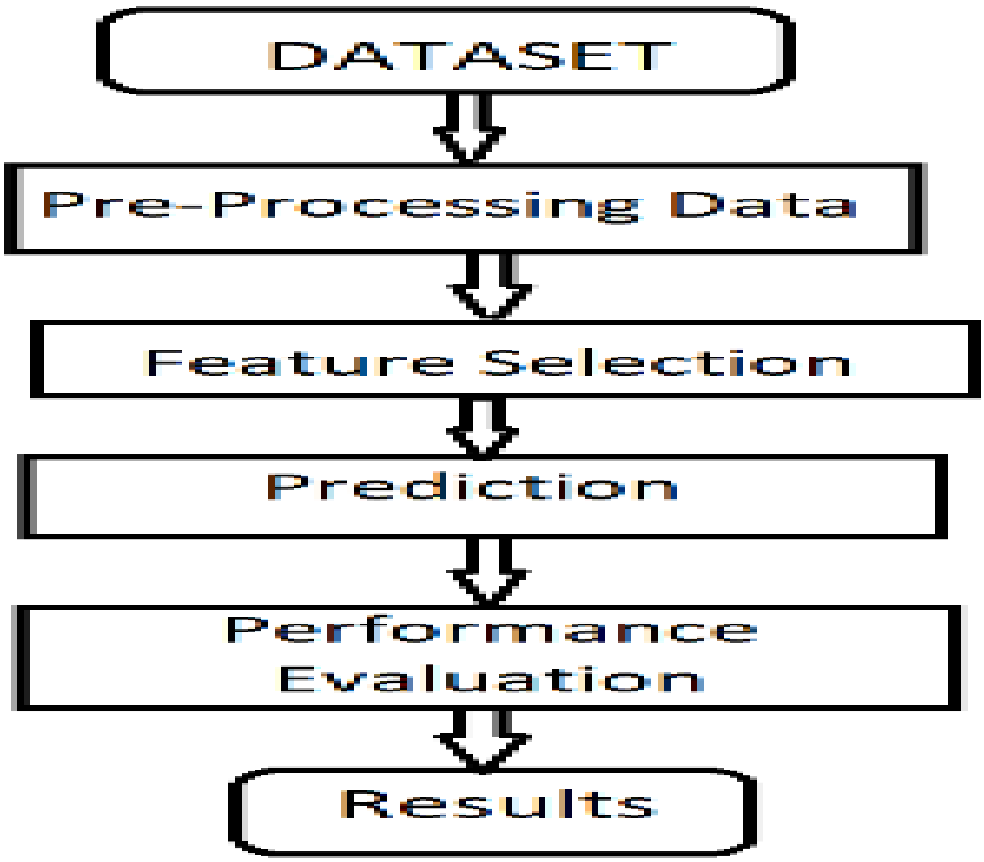


Fig 3: A framework for heart disease prediction

5. IMPLEMENTATION

- The dataset, sourced from Kaggle, comprises essential attributes like age, gender, cholesterol level, blood pressure, among others, alongside the target variable signifying the existence or non-existence of cardiac element.

- Data preparation for model training encompasses activities such as handling missing data entries and distinguishing between continuous and categorical attributes, which are part of the data pre-processing phase.

- To prevent overfitting, prioritizing feature selection is crucial. This involves picking out the most relevant features from the original dataset to enhance model performance. Essentially, it's about pinpointing the influential features that strongly influence predicting the target variable
- With the dataset nearly prepared, the next step is to divide it into training and testing sets. Utilizing the Random Forest algorithm, the model is trained on training data This method, belonging to the ensemble learning category, builds numerous decision trees during training and produces the most common class among the trees' predictions

- Following that, the trained model's effectiveness is assessed on testing set with suitable metrics like accuracy, precision, recall, F1-score, and ROC-AUC score. These metrics help gauge model's predictive capabilities and overall performance.

6. OBSERVATIONS

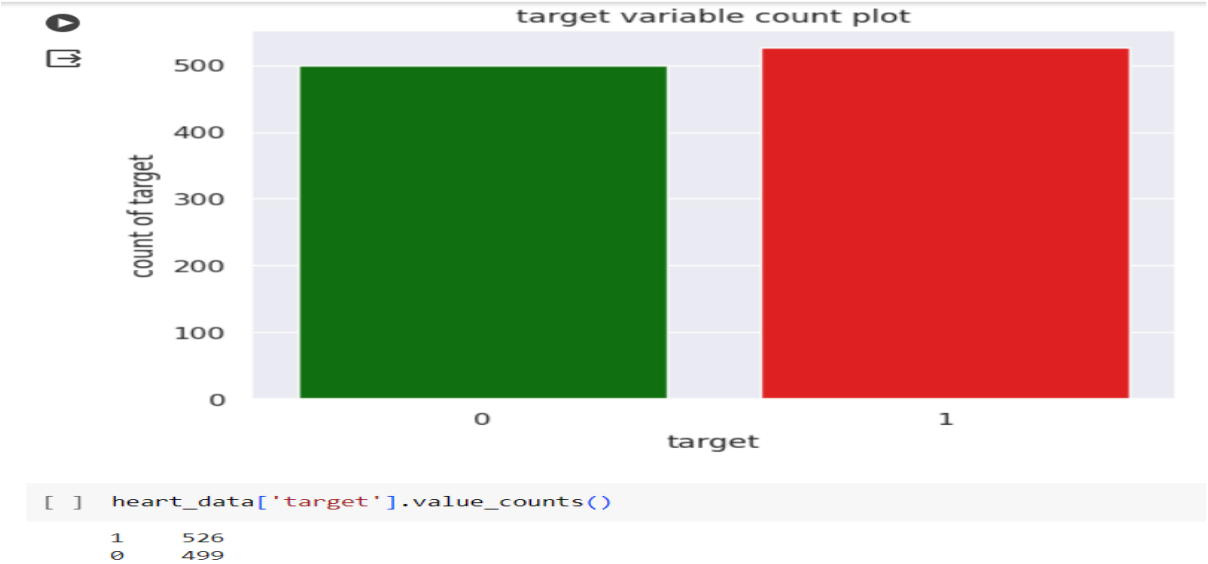


Fig 4: Target variable count plot

In Figure 4, the green plot indicates a count of 526, representing individuals without heart disease, while the red plot, with a count of 499, represents individuals with heart disease

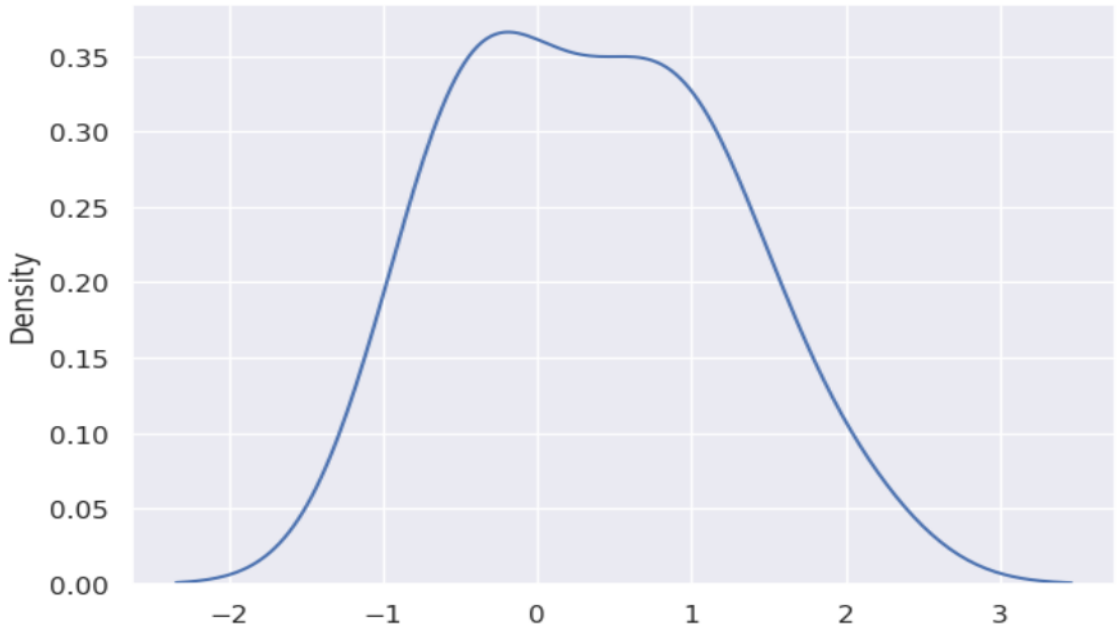


Fig 5: Skewness of data

Skewness refers to the degree of distortion in a dataset. When the curve shifts to the left or right, it indicates skewness. Excessive skewness can disrupt the functioning of statistical models. However, in Figure 5, the dataset exhibits very minimal skewness..

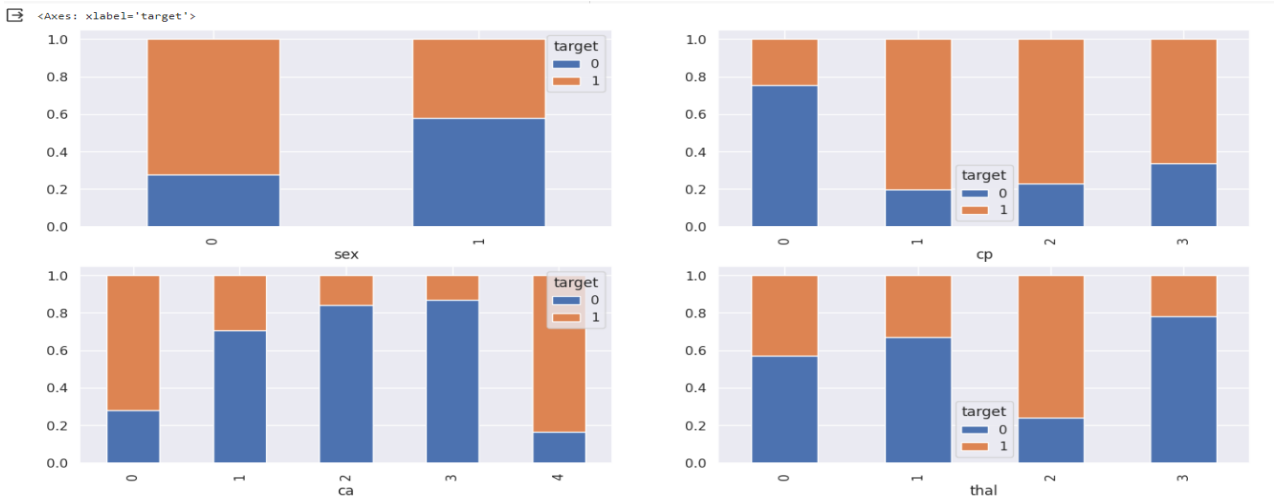


Fig 6: Independent variable vs Target variable

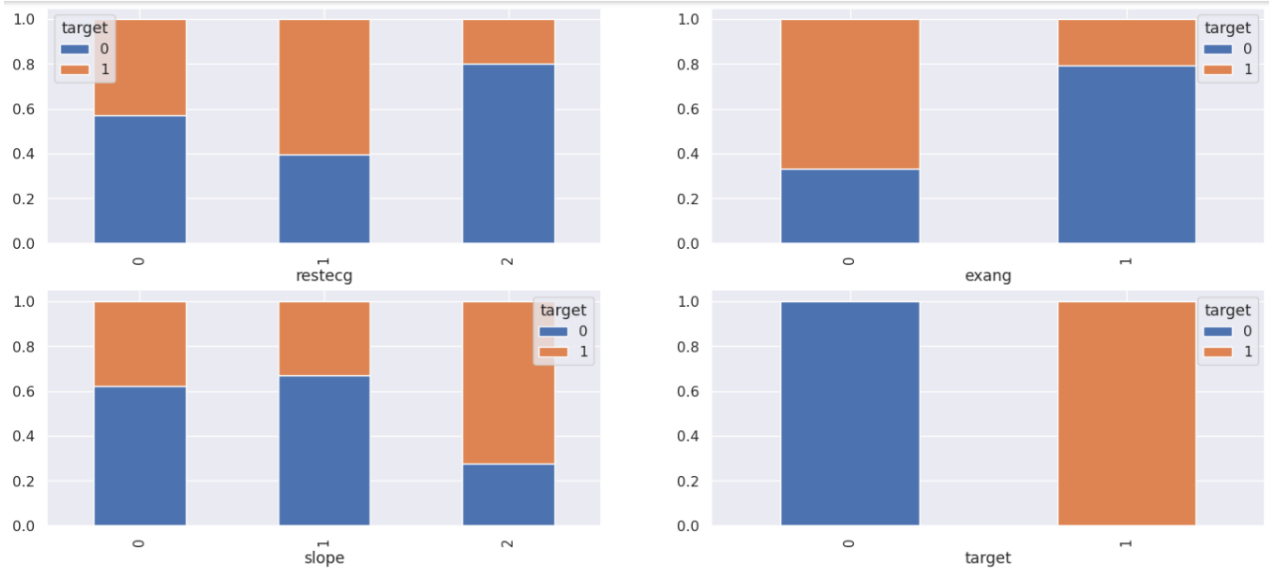


Fig 7: Independent variable vs Target variable

Figure 6 and 7 illustrate the independent attributes present in the dataset and their relationship with the target attribute for patients.

7. RESULTS

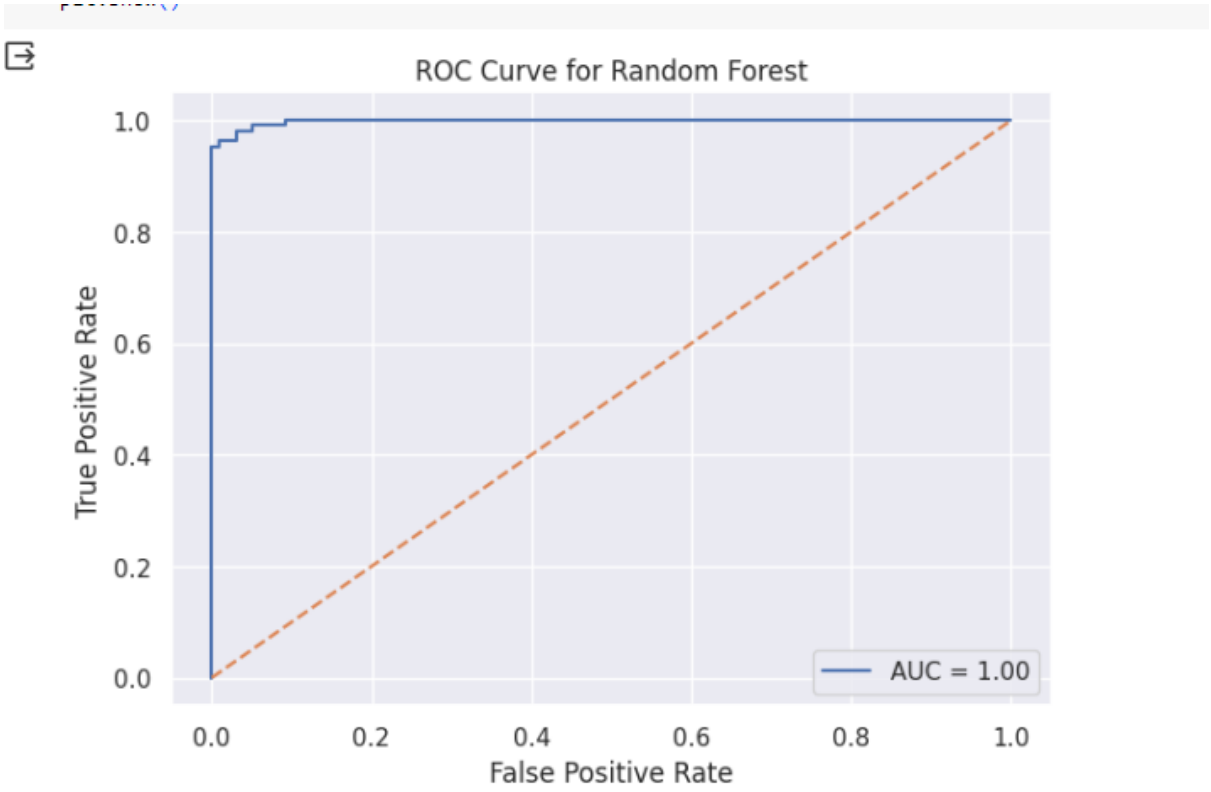
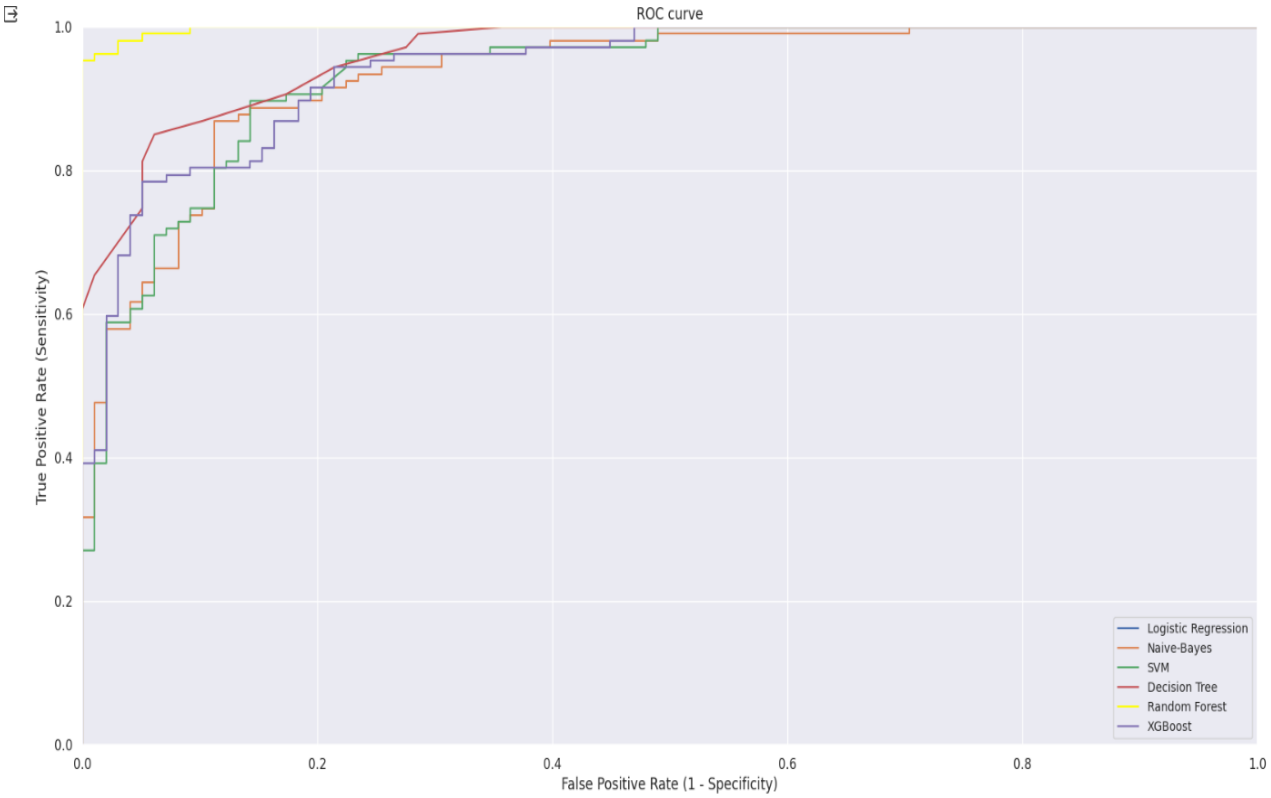


Fig 8: ROC curve for Random Forest Classifier



```
Area under the curve for Logistic Regression : 0.94
Area under the curve for Naive-Bayes         : 0.93
Area under the curve for SVM                 : 0.93
Area under the curve for Decision Tree       : 0.96
Area under the curve for Random Forest       : 1.0
Area under the curve for XGBoost             : 0.94
```

Fig 9: ROC curve for alternative algorithms

The ROC curve serves as a robust assessment tool due to its ability to provide a holistic perspective on prediction accuracy across different cutoff points. It visualizes the True Positive Rate (TPR) versus the False Positive Rates (FPR) across different thresholds, providing understanding into spectrum of the model's performance. ROC curve positional nearer to the upper left corner signifies a model that is more optimal in its performance. Figure 8 depicts the ROC curve for the Random Forest Classifier algorithm, aligning perfectly with its performance. Additionally, Figure 9 displays the ROC curves for alternative algorithms.

8. CONCLUSION

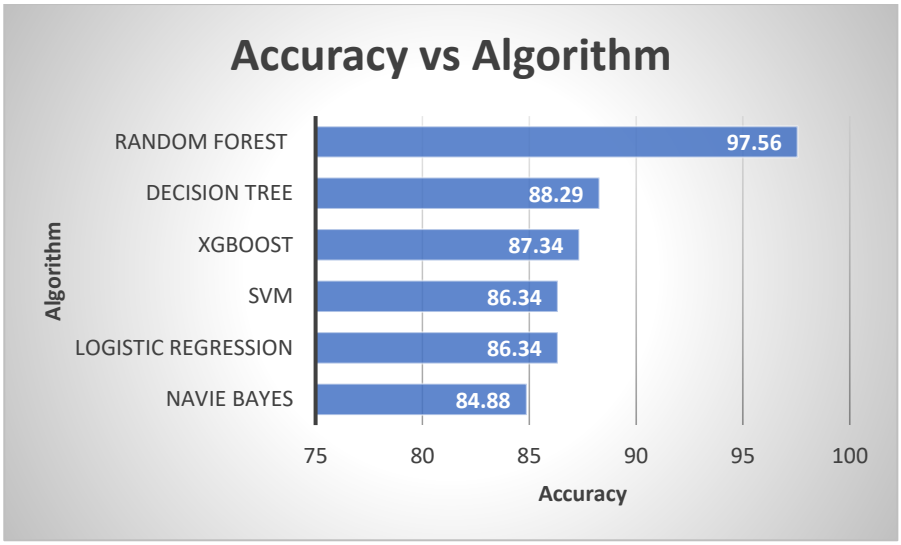


Fig 10: Accuracy vs Algorithm

Fig 10 displays the accuracy of various algorithms, which serves as a prevent metric for assessing the effectiveness of models.

Complete accuracy is a commonly employed measure in the field of machine learning, it's not always the most reliable indicator of model performance. It's important to assess various metrics and consider factors beyond just accuracy when determining the best model for a particular task.

```
print("1)  Logistic Regression    :",round(f1_score(y_pred,y_test), 2))
print("3)  Naive-Bayes            :",round(f1_score(pred_gnb,y_test), 2))
print("4)  SVM                    :",round(f1_score(pred_svc,y_test), 2))
print("5)  Decision Tree          :",round(f1_score(pred_dt,y_test), 2))
print("7)  Random Forest          :",round(f1_score(pred_rf,y_test), 2))
print("9)  XGBoost                :",round(f1_score(predxg,y_test), 2))
```

```
1)  Logistic Regression    : 0.88
3)  Naive-Bayes            : 0.86
4)  SVM                    : 0.88
5)  Decision Tree          : 0.89
7)  Random Forest          : 0.98
9)  XGBoost                : 0.88
```

Fig 11: F1-Score

Precision signifies the fraction of positive instances predicted correctly out of all instances classified as positive by the classifier. Recall, sometimes referred to as sensitivity, quantifies the ratio of correctly predicted positive instances among all actual positive instances. The F1 Score, derived from the harmonic mean of precision and recall, provides a balanced measure of a classifier's accuracy and completeness as shown in fig 11. A greater F1 score suggests superior performance, indicating both recall and precision are high, meaning the classifier is effective at classifying positive instances correctly while reducing the occurrence of false positives and false negatives.

Here, Random Forest has the highest f1_score. Hence, based on the f1_score, Random Forest is the best fit model.

```
print("1)  Logistic Regression    :",cmlr[0][1])
print("3)  Naive-Bayes            :",cm_gnb[0][1])
print("4)  SVM                    :",cm_svc[0][1])
print("5)  Decision Tree          :",cm_dt[0][1])
print("7)  Random Forest          :",cm_rf[0][1])
print("9)  XGBoost                :",cm_xg[0][1])
```

```
1)  Logistic Regression    : 7
3)  Naive-Bayes            : 11
4)  SVM                    : 6
5)  Decision Tree          : 10
7)  Random Forest          : 2
9)  XGBoost                : 7
```

Fig 12: Type I Error

False Positives, also known as (Type I Error) happen when we erroneously reject a true hypotensis. The lower the count of False positives, the more effective the model and it is clear in fig 9 Random Forest Classifier as lower value as compared to other algorithms. This is because, while predicting, if we predict that a person has a heart disease, but later he/she does not actually have any heart disease, such erroneous predictions can escalate the risk factor to a concerning level.

Random Forest algorithm has the least number of False Positives (Type I Error) as shown in fig 12. Hence, based on the False Positives (Type I Error), Random Forest is the best fit model.

After assessing the performance using five distinct metrics, It becomes clear that the Random Forests Classifier stands out as the premier option for forecasting the likelihood of heart disease in individuals.

9. REFERENCES

[1] M. Snehith Raja , M. Anurag, Ch. Prachetan Reddy, "Machine Learning Based Heart Disease Prediction System" in Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI).

[2] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, "Heart Disease Prediction using Machine Learning Techniques" in Proceedings of the 2020 IEEE 2nd International Conference on Computer Communication and Informatics (ICCCI).

[3] Muntasir Mamun, Md. Milon Uddin, Vivek kumar, Asm Mohalmenul Islam, "MLHeartDis: Can Machine Learning Techniques Enable to Predict Heart Diseases?" in proceedings of 2020 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)

[4]https://r.search.yahoo.com/_ylt=Awr1TXhPBhhm4n4Rg1G7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1712879311/RO=10/RU=https%3a%2f%2fwww.javatpoint.com%2fheart-disease-prediction-using-machine-learning/RK=2/RS=QXs18v03zSj1Bnj498w.iyAIbw0-

[5]https://r.search.yahoo.com/_ylt=Awr1TXhPBhhm4n4RlVG7HAx.;_ylu=Y29sbwNzZzMEcG9zAzQEdnRpZAMEc2VjA3Ny/RV=2/RE=1712879311/RO=10/RU=https%3a%2f%2ftowardsdatascience.com%2fproject-predicting-heart-disease-with-classification-machine-learning-algorithms-fd69e6fdc9d6/RK=2/RS=zBzfD2cRW8GK7X1jtIGvUtSH5jg-