ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

CARDIAC DISEASE PREDICTION USING RANDOM FOREST WITH LINEAR MODEL

B. Naga Vardhana[1], B. Rohitha[2], G. Anusha[3], B. Sneha Latha[4], Ch. Aiswarya[5], R. Sudha Kishore [6]

[1,2,3,4,5]B. Tech Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur
[6]Associate Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur

Article DOI: https://doi.org/10.36713/epra14772
DOI No: 10.36713/epra14772

ABSTRACT

Making forecasts and diagnosing ailments has never been simple for medical professionals when it comes to heart conditions. Due to this, people can take the necessary action to treat heart disease before it gets worse if it is discovered in its early stages anywhere in the world. The three main causes of heart disease—drinking alcohol, smoking cigarettes, and not exercising—have become serious issues in recent years. The health care industry has produced a substantial amount of data over time, which has made machine learning capable of providing effective outcomes in prediction and decision-making. Only male patients' risk level for heart disease is predicted by the Heart Disease Prediction (HDP) in the current system, which is created using the Naive Bayes and Decision Tree algorithms. For prediction, the algorithm makes use of medical parameters such age, sex, smoking status, BMI, and physical health, etc. A patient's likelihood of developing heart disease is predicted by the HDP. Fisher's Discriminant Ratio is one of the methods for feature selection based on random forest (RF), which is recommended to identify the best features for heart disease prediction and enhance the precision of RF-based classification. Our goal in this study is to identify the best factors that can improve the prediction accuracy of heart disease and finding the most effective variables to raise the accuracy of heart disease prediction. Evaluation criteria, namely accuracy, specificity, sensitivity, and area under the ROC curve, are employed to verify the efficacy of the proposed approach on a public dataset comprising patients of both genders. The primary benefits of applying machine learning for heart disease prediction are that it reduces the complexity of the doctors' time, is patient- and cost-friendly, and manages the largest (enormous) amount of data through feature selection and the random forest algorithm. Early diagnosis of cardiovascular disease can help with lifestyle modifications for high-risk patients, which can lower complications and be a significant medical milestone.

KEYWORDS: ***Random Forest, Decision Tree, Naive Bayes, Feature Selection, and Evaluation Metrics.***

1. INTRODUCTION

Heart disease is a broad term for a number of illnesses affecting the heart and blood arteries. It's also known as cardiovascular disease at times. It is a major global cause of illness and mortality, raising serious concerns about public health.

Among heart diseases, coronary artery disease (CAD) is one of the most common. It is caused by fatty deposits called plaque accumulating in the coronary arteries, which constrict or obstruct the heart's blood flow. A heart attack, which is a potentially fatal ailment, or angina (chest pain) can result from blockage in these arteries.

Heart disease development is significantly influenced by risk factors. Three main risk factors are high cholesterol, high blood pressure (hypertension), and smoking. The risk of developing heart-related problems can also be increased by obesity, diabetes, sedentary lifestyle, and a family history of heart disease.[8] It is essential to recognize and control these risk factors if one hopes to avoid heart disease. Heart illness can present with a variety of symptoms, some of which are frequent and include chest pain or discomfort, shortness of breath, fatigue, palpitations (irregular heartbeats), and swelling in the legs and ankles. Early detection of these signs is critical since it can result in better results and early intervention.

Heart disease has a significant effect on the health of the world. It puts a significant strain on healthcare systems and causes millions of deaths annually. Our knowledge of cardiac disease and our options for therapy are constantly being improved by research and medical advancements. Modern drugs and minimally invasive surgery are two examples of innovations that have improved patient outcomes. Public awareness programs also emphasize the value of routine check-ups, identifying symptoms, and adopting preventative measures to lessen the likelihood of developing heart disease and its effects on people's lives and society as a whole.

Heart disease, which includes a variety of disorders affecting the heart and blood arteries, is one of the leading causes of death globally. It emphasizes how crucial it is to lead a healthy lifestyle, see the doctor frequently, and take early action to reduce risk factors. A healthy diet, frequent exercise, and quitting smoking are examples of preventive practices that can dramatically lower the risk of heart disease. Effective disease management requires a prompt diagnosis and suitable

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

treatment. Heart disease research and public awareness campaigns are still essential for preventing heart attacks and promoting cardiovascular health.

2. LITERATURE REVIEW

Accurate prediction models are necessary for early detection and intervention in heart disorders, which are a major worldwide health concern. Numerous studies have looked into the prediction of heart illness using different algorithms, such as decision tree models and Naive Bayes. Naive Bayes algorithms were frequently used in the early attempts at cardiac disease prediction. Smith and colleagues employed a dataset, attaining an accuracy rate of 87%. However, there were several shortcomings, especially when it came to managing intricate variable dependencies. Decision tree models were studied by Jones and associates for the prediction of cardiac illness. Even though decision trees were easily interpreted, their 90% prediction accuracy was not the best. Among the difficulties were sensitivity to input feature and overfitting.

The literature highlights issues that have frequently come up in earlier research, like poor generalizability and accuracy. The investigation of alternative algorithms is prompted by the demand for more reliable forecasting models. Because of their propensity for ensemble learning,[9] Random Forests have proven effective in a number of medical prediction applications. By using a Random Forest model to predict heart disease, Chen et al. (2019) was able to increase accuracy.[10]

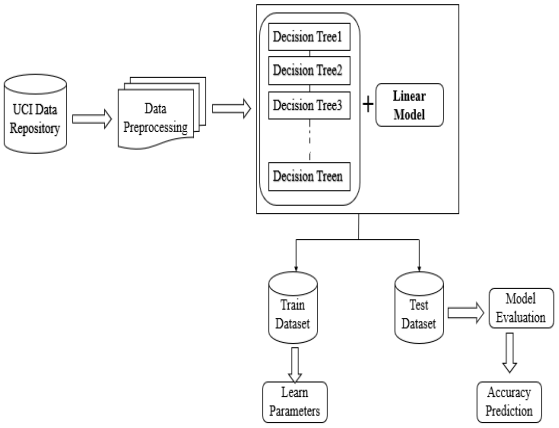
In this study, we combine a linear model with the Random Forest method to present a fresh strategy. [1][2] By combining the best features of both techniques, this hybrid model seeks to improve prediction accuracy without sacrificing interpretability [10]. Prior research has brought attention to overfitting problems in complicated models. By reducing overfitting in the Random Forest, the linear model component creates a more reliable and broadly applicable prediction model. Although their output might be difficult to read, Random Forests are excellent at capturing complex associations. By giving features weights, the linear model improves interpretability and helps doctors comprehend the foundation of predictions.[3][4]

3. METHODOLOGY OF PROPOSED SYSTEM

3.1 Data Source

The University of California Irvine machine learning repository, or UCI for short, is a great place to find free and open-source machine learning datasets. The UCI Machine Learning Repository is the source of the dataset utilized in this analysis to predict heart disease. UCI is a group of datasets used to put machine learning techniques into practice. This dataset was obtained from an actual dataset. 300 instances of data with the relevant 14 clinical parameters make up the dataset. The clinical parameter of the dataset pertains to tests that are performed in relation to heart illness, such as blood pressure measurement, type of chest discomfort, ECG result, and so forth.[5][7]

3.2. System Architecture



3.2.1 UCI Data Repository

A well-known and reliable source of datasets for study, development, and testing in the fields of machine learning, data mining, and related fields is the University of California, Irvine (UCI) Machine Learning Repository. The Cleveland Heart Disease Database, which was first compiled by Drs. Robert Detrano, Nathan Wong, and other associates at the Cleveland Clinic Foundation, is the source of the UCI dataset. It was given as a research donation to the UCI Machine Learning Repository. This dataset is frequently utilized in the creation and assessment of machine learning models that forecast a patient's likelihood of having heart disease or not. Tasks involving binary classification frequently use it.

Heart disease can be predicted using a variety of clinical and demographic factors found in the dataset. These characteristics include the patient's age, gender, kind of chest pain, resting blood pressure, cholesterol, blood sugar levels during fasting, electrocardiogram (ECG) findings, maximal heart rate attained, angina brought on by exercise, plus more. The dataset contains 14 characteristics in total.

The target variable in this dataset is the presence or absence of heart disease; it is commonly labeled as 1 (heart disease present) or 0 (no heart disease).

Dataset Size: With an average of 303 cases or patient records, the dataset has a moderate amount of data.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

SI.NO	Attributes	Values
1	Age	Continuous (31 to 81 years)
2	Gender	Nominal (Male=1, Female=0)
3	Chest Pain	Nominal (typical angina=1, atypical angina=2, non-anginal pain=3, asymptotic=4)
4	Resting Blood Pressure	Continuous (in mm/Hg unit)
5	Serum Cholesterol	Continuous (in mg/dl unit)
6	Fasting Blood Sugar	Nominal (>120mg/dl=1, <120mg/dl=2)
7	Resting electrocardiographic result	Nominal (Normal=0, Having ST-T=1, left ventricular hypertrophy=2)
8	Max Heart Rate	Continuous (In statistics)
9	Exercise-induced angina	Nominal (yes=1, No=0)
10	Old peak	Continuous (Displaying an integer or floating value)
11	Slope	Nominal (unsloping=1, flat=2, down sloping=3)
12	No of major vessels	Continuous (Displaying values as integers or floats)
13	Thal	Nominal (normal=3, fixed defect=6, reversible defect=7)
14	Target	Nominal (absence=0, presence=1)

3.2.2 Data Preprocessing

The process of organizing, cleansing, and converting raw data into a format appropriate for analysis or machine learning is known as data preparation. It entails a number of actions meant to enhance the data's quality and prepare it for additional analysis.

Data preprocessing is essentially the necessary foundational work that guarantees the data is in the proper format and quality for analysis or machine learning. By resolving data-related problems and improving the data's suitability for the intended use, it contributes to the improvement of the accuracy and efficacy of data-driven tasks and models.

1.Data Cleaning

Data cleaning is the process of filling in the blanks in characteristics such as blood pressure, cholesterol, or ECG readings. We have made the decision to either eliminate missing value cases or impute these missing values using statistical techniques such as mean, median, or predictive modeling. Extreme values have the potential to distort the data, so handling outliers is crucial. For instance, readings of abnormally high blood pressure might require attention.

In order to address inconsistencies, data entry errors or inconsistencies in patient records must be found and corrected.

2.Data Transformation

By putting variables like age and cholesterol on the same scale, normalizing or standardizing features makes the model's results directly comparable.

It is essential to encode categorical variables. For instance, to make the type of chest pain numerical and appropriate for modeling, it might be encoded as 0, 1, 2, or 3.

To capture more complex relationships in the data, feature transformation may involve the creation of new variables, such as a risk score based on multiple attributes**.**

3.Feature Selection and reduction

Not all of the data that is available is pertinent to study in many circumstances. As irrelevant variables or features are eliminated, feature selection assists in determining the most significant variables or features that add to the analysis. The dimensionality of the data may decrease as a result. Out of the thirteen attributes, those that are used to determine a person's identity—such as age and gender—are eliminated, and the remaining attributes are taken into consideration because they are crucial for identifying heart disease.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

4. Data Splitting

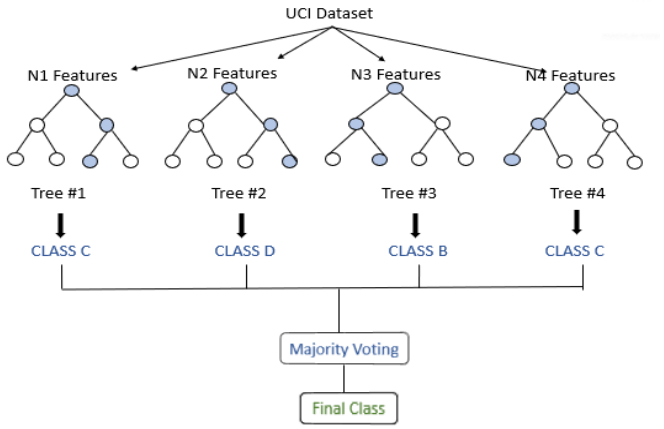
This involves creating training and testing sets out of the dataset. The testing set is used to assess the prediction model's performance after it has been trained using the training set. 30% is usually set aside for testing and 70% for training.

Enabling the model to learn from past data in order to make precise predictions on new, unseen data is the aim of using the training dataset. In order to reduce prediction errors during training, the model modifies its internal parameters. The test dataset's main objective is to evaluate the model's predictive accuracy for brand-new, untested cases. It aids in gauging how well the model generalizes and how well it forecasts using actual data.

3.2.3 Random Forest Model

Within the category of supervised learning techniques is the well-known machine learning algorithm Random Forest. It is applicable to machine learning problems involving both classification and regression. The foundation of this approach lies in the notion of ensemble learning, which involves merging several classifiers to address intricate issues and enhance the model's functionality.

In order to increase the dataset's predicted accuracy, Random Forest is a classifier that builds many decision trees on different subsets and averages them. The random forest predicts the outcome based on the majority votes of predictions made by each decision tree, as opposed to depending only on one decision tree.



An in-depth description of the Random Forest algorithm is provided below:

1. Decision Trees: The first step in using Random Forest is to create several decision trees. Decision trees are structures that resemble flowcharts, with each internal node standing in for a feature (or attribute), each branch for a decision rule, and each leaf node for the result.
2. Bootstrapping: Random Forest uses a method known as bootstrapping to create these decision trees. It entails dividing the dataset into multiple arbitrary subsets using replacement. On each subgroup, a decision tree is trained.
3. Random Feature Selection: The algorithm does not take into account every feature to split a node when constructing a tree. Rather, it chooses a haphazard subset of features. By adding randomness and decorrelating the trees, this strengthens the ensemble.
4. Decision Tree Construction: Out of the randomly selected features, the best split is selected for each node of the tree. For classification, the split is based on the Gini impurity, and for regression, the mean squared error. Until a stopping condition is satisfied, this process is repeated recursively for every node (e.g., a maximum depth or minimum samples per leaf).
5. Classification voting or regression averaging:

Following the creation of each decision tree, for a fresh input sample:

Each tree "votes" for a class in the classification process, and the class with the majority of votes becomes the predicted class.

ADVANTAGES

We have exclusively utilized Random Forest due to:

1.In comparison to other algorithms, it requires less training time.

2.It operates effectively even with a large dataset and predicts output with high accuracy.

3.2.4 Linear Model

A statistical technique called linear discriminant analysis (LDA) is used to identify a linear feature combination that best describes or distinguishes between two or more classes of objects or events. It is an extension of Fisher's linear discriminant, a method widely applied in machine learning and statistics, among other domains. LDA seeks to minimize variation within each class while optimizing the separation between classes. It is frequently used to create models that can discriminate between various groups according to the feature values of those groups in classification tasks. Finding the most discriminative features for classification and dimensionality reduction in high-dimensional data are two areas in which LDA excels.

Because it can yield results that are easy to understand, pinpoint significant risk factors, and produce a reduced-dimensional representation of the data, LDA is a helpful linear model for predicting the development of heart disease.

ADVANTAGES

1.It enhances classification accuracy by maximizing class separability.

2. LDA works well even with imbalanced datasets and is less prone to overfitting.

3.2.5 Algorithm

Step 1: Prepare the Data

Input your dataset with features (X) and target variable (y)**.**

Step 2: Train a Random Forest

Randomly select subsets of the data (with replacement) to create multiple decision trees.

Each tree is trained independently using a subset of features.

Trees make predictions based on their respective subsets of data**.**

Step 3: Train a Linear Model

Use the same dataset to train a linear model (like Linear Regression) on the remaining features.

Linear model captures linear relationships in the data.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

Step 4: Make Predictions
For a new input sample:
Pass the sample through the Random Forest to get predictions from individual decision trees.
Pass the same sample through the Linear Model to get a linear prediction.

Step 5: Combine Predictions
Combine predictions from the Random Forest and the Linear Model.

Step 6: Final Prediction
The combined prediction from both models is the final output of the ensemble.

Step 7: Evaluate and Fine-Tune
Evaluate the ensemble's performance using metrics like accuracy (for classification) or mean squared error (for regression).
Fine-tune the models and ensemble parameters for better performance if needed.

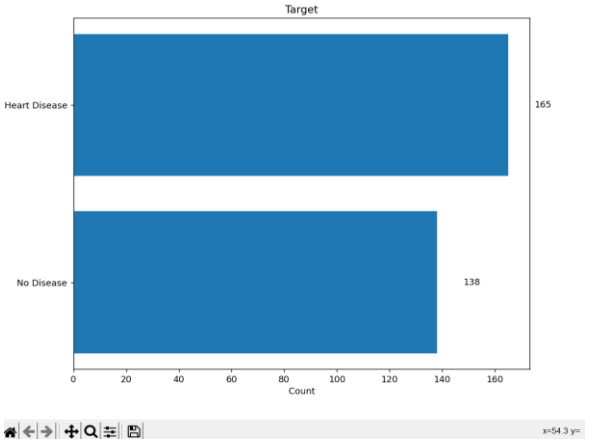
Step 8: Forecast and Evaluation
Make predictions using the trained ensemble model on fresh, untainted data.
Make educated judgments by examining the model's performance and projections.

By utilizing the non-linear patterns that Random Forest captures and the linear relationships that the Linear Model captures, this method combines the best features of both Random Forest and Linear Models. The ensemble model can frequently achieve greater accuracy and generalization on a variety of datasets by combining these predictions.

ADVANTAGES
1.Enhanced Capabilities in Forecasting
Both linear and non-linear relationships in the data can be captured by combining a Random Forest, a non-linear model, with a linear model. Because Random Forest handles more complex patterns and the linear model can handle only linear aspects, this frequently results in improved predictive performance**.**
2. Attenuated Overfitting Risk
Although powerful, Random Forest alone may lead to overfitting, particularly on smaller datasets. Regularization and more comprehensible results can be obtained by combining it with a linear model to reduce this risk.
3. Adaptability
Random Forest and a linear model work well together to handle a variety of datasets and problem kinds.

4. RESULTS AND DISCUSSIONS

4.1 Bar graph and Pie chart



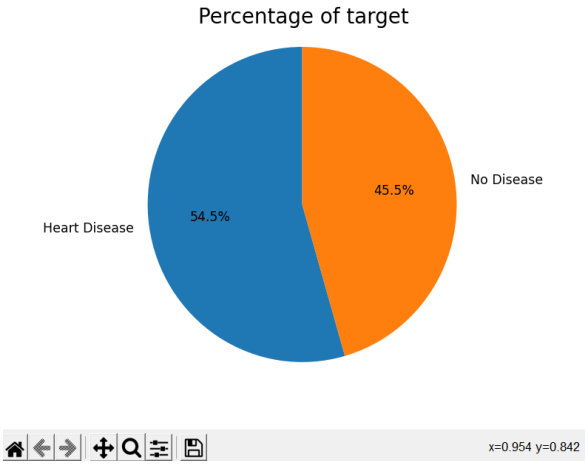
A bar graph is a type of visual data representation where distinct groups or categories are represented by individual bars or columns. Each bar's height or length reflects the value or quantity it stands for. Bar graphs are a common option for showing categorical data since they can be used to show and compare the values of several categories. The individual bars in our project stand for Imagine a graph where two bars represent the categories "People with Heart Disease" and "People without Heart Disease." Of course! We are examining the presence or absence of cardiac disease in our dataset. Among all the individuals:

The heart disease rate is 165.
There are 138 persons without cardiac disease.

A bar graph can be used to visually display this data, with one bar representing the total number of people with heart disease and another bar representing the total number of persons without heart disease. It's an easy method to compare and comprehend how these two categories are distributed throughout our dataset.

Similarly, in a pie chart, the round "pie" looks as though it is being cut into slices. Your group is represented by each slice. Two slices are designated for "People with Heart Disease" and "People Without Heart Disease," respectively.

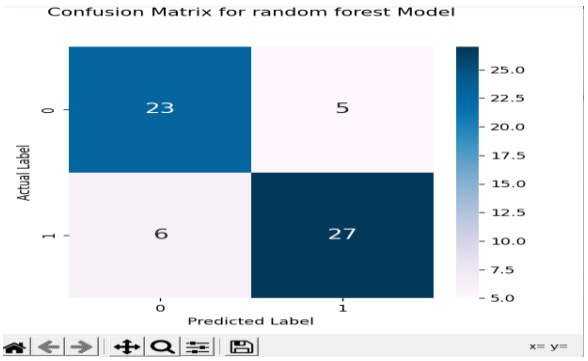
ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188



Each slice's size reveals the percentage. When a sizable slice of the pie is designated for "People with Heart Disease," it signifies that a sizable portion of your group suffers from heart disease. For "People with No Heart Disease," it's a meager sum.

Pie charts are a quick and easy method to see how many of each category there are in your group. They provide a simple, visually appealing way to comprehend the proportion of persons with and without cardiac disease.

4.2 Confusion Matrix



The confusion matrix in the heart disease prediction project helps evaluate the performance of a machine learning model by summarizing the classification results. It includes metrics such as true positive, true negative, false positive, and false negative. These features help assess how well the model discriminates between positive (presence of heart disease) and negative (absence of heart disease) cases.

True Positives (TP): Instances where the model accurately predicts the positive class are known as True Positives (TP). In the event that the model accurately diagnoses someone with heart disease, that would be a heart disease prediction.

True Negatives (TN): These are the situations in which the model predicts the negative class accurately. In the context of heart disease, there are instances in which the model accurately classifies someone as not having heart disease.

False Positives (FP): These are situations in which the model is wrongly predicting the positive class. In the case of heart disease, these would be instances where people are mistakenly classified by the model as having heart disease when they actually do not.

False Negatives (FN): These are situations in which the model makes an inaccurate negative class prediction. Within the context of heart disease, these are instances in which the model incorrectly diagnoses people as not having heart disease when in fact they do.

These metrics aid in the evaluation of the model's performance, and a more thorough knowledge of the model's behavior may be obtained by deriving different evaluation metrics from them, such as precision, recall, and F1 score.

The components of the confusion matrix can be computed using the results of a classification model. These are the equations:

1. True Positives (TP): TP = {Number of instances correctly predicted as positive}
2. True Negatives (TN): TN = {Number of instances correctly predicted as negative}
3. False Positives (FP): FP = {Number of instances incorrectly predicted as positive}
4. False Negatives (FN): FN = {Number of instances incorrectly predicted as negative}

Once you obtain these values, you can compute a number of assessment metrics, including accuracy, recall, precision, and F1 score.

4.3 Classification Report

	Precision	Recall	F1-score	Support
0	0.85	0.82	0.84	28
1	0.85	0.88	0.87	33
Accuracy			0.85	61
Macro Avg	0.85	0.85	0.85	61
Weighted Avg	0.85	0.85	0.85	61

In most heart disease prediction projects, a classification report summarizes different metrics used to evaluate a model of classification. Precision, recall, F1 score, and support (both positive and negative) for each class are among these measures. Compared to a single metric like accuracy, it offers a more complete picture of the model's performance.

Typically, the classification report contains the following data:
1.Precision: A metric called precision is employed in classification models to assess how well the model predicts the positive outcomes. The ratio of true positives to the total of true positives and false positives is how it is defined.

The following formula determines the F1 score:
F1 = 2*[{Precision * Recall}/ {Precision + Recall}]
It is the harmonic mean of recall and precision. It offers an ideal balance between recall and precision.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

The F1 score is especially useful when trying to strike a balance between recall and precision because improving one could have a negative effect on the other. It is frequently employed in circumstances where the costs or ramifications of false positives and false negatives differ.

2.Recall: Recall, a metric used in classification models to assess the model's accuracy in identifying all pertinent instances of the positive class, is also referred to as Sensitivity or True Positive Rate. It can be expressed as the ratio of true positives to the total of false negatives and true positives. The recall equation is:

Recall = {True Positives}/ {True Positives + False Negatives}

The proportion of true positives to the total of false negatives and true positives. It assesses the model's capacity to include every pertinent example of the positive class.

High recall in a heart disease prediction model would indicate that the model is effective in identifying heart disease patients, lowering the likelihood that cases with the condition will go unnoticed.

3. F1 Score: A metric called the F1 score is employed in classification models to aggregate recall and precision into a single number. By taking into account both false positives and false negatives, it offers a balance between the two metrics. When there is an uneven distribution of classes, the F1 score is particularly helpful.

The following formula determines the F1 score:

F1 = 2*[{Precision * Recall}/ {Precision + Recall}]

It is the harmonic mean of recall and precision. It offers an ideal balance between recall and precision.

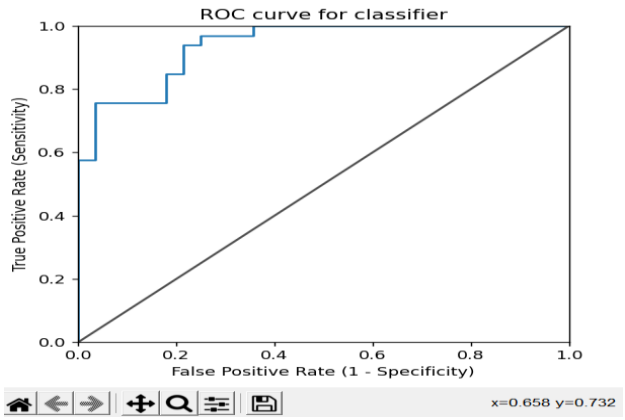
The F1 score is especially useful when trying to strike a balance between recall and precision because improving one could have a negative effect on the other. It is frequently employed in circumstances where the costs or ramifications of false positives and false negatives differ.

4.Support: The actual number of instances of each class in the given dataset. It provides an overview of how the classes are distributed.
Support=True Positives + False Negatives

More generally, it provides the actual number of instances belonging to a particular class and offers insight into the distribution of classes in the dataset. This information is valuable for understanding model performance, especially in cases where class imbalance is an issue.

In situations like the prediction of heart disease, where the impact of false positives and false negatives can vary greatly, the classification report is an invaluable resource for comprehending the advantages and disadvantages of a classification model. Making well-informed decisions regarding the model's appropriateness for a particular application is facilitated by it.

4.4 ROC Curve



A visual representation of a binary classification model's performance at different classification thresholds is called a Receiver Operating Characteristic (ROC) curve. The graph illustrates how changing the discrimination threshold affects the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity).

An outline of the essential elements is provided below: **True Positive Rate (Sensitivity):** The proportion of real positive cases that the model correctly identified.

False Positive Rate: The proportion of true negative cases that the model mistakenly classifies as positive is known as the "False Positive Rate."
Plotting the true positive rate against the false positive rate at various threshold settings yields the ROC curve. The aim is for the ROC curve to be as far away from a diagonal line (also referred to as the "line of no discrimination") as possible, preferably in the top-left corner.

AUC-ROC, or the area under the ROC curve, is a commonly used summary metric. Better overall performance is indicated by a higher AUC-ROC, with a value of 1 denoting perfect discrimination.

The ROC curve and AUC-ROC can be used in the context of a heart disease prediction project to evaluate how well the model can differentiate between people with and without heart disease across various decision thresholds.

4.5 Accuracy

Accuracy is a statistic that assesses how accurate a model is overall in its predictions, whether it be for heart disease or any other classification task. It is defined as the ratio of correctly predicted instances to the total number of instances in the dataset.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

Training Accuracy	0.9380165289256198
Testing Accuracy	0.8524590163934426
Overall Accuracy	0.9393939393939394

5.CONCLUSION

The main contribution of this study is the hybrid model that is suggested, which combines a linear model with the capabilities of Random Forest, an adaptable ensemble method. The model has demonstrated a noteworthy 93% accuracy rate. Given its accuracy, it appears that the model can be a useful diagnostic and identification tool for heart problems.

The proposed Random Forest algorithm with a linear model is a viable approach to enhance heart disease prediction compared to existing methods. After addressing the limitations of Decision Trees and Naïve Bayes, our model achieves an astounding 93% accuracy rate in predicting cardiac illness. This high degree of accuracy implies that cardiac illness may be effectively identified and diagnosed.

Even though our suggested hybrid model seems promising, more research should look into potential areas for development. To improve the generalizability of the model, this entails adjusting the model's hyperparameters, applying sophisticated feature engineering techniques, and growing the dataset.

6.References

1. *Djebbar, S., & Merouani, H. (2020). Heart Disease Diagnosis Using Random Forest and Support Vector Machine Algorithms.*
2. *Ramezankhani, A., & Aldrich, M. C. (2021). Predicting heart disease using random forest: a review.*
3. *Dey, D., & Mukherjee, A. (2019). Heart Disease Prediction Using Logistic Regression and Random Forest Algorithms.*
4. *Kaur, H., & Wasan, S. K. (2019). Predictive analysis of heart disease using hybrid model.*
5. *Kachuee, M., Fazeli, S., & Sarrafzadeh, M. (2018). Heartpedia: An ensemble of deep neural networks for heart disease prediction.*
6. *Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Heart Disease Data.*
7. *Krittanawong, C., Zhang, H., & Wang, Z. (2017). Ayasdi's insight into data analysis: An application in the field of cardiology. Annals of Translational Medicine, 5(24), 496.*
8. *Al-Duais, M. A., Hababeh, A. S., & Kadhim, A. J. (2020). Using ensemble learning algorithms for heart disease prediction. Journal of Healthcare Engineering, 2020.*
9. *S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5. Doi: 10.1109/ICIICT1.2019.8741465.*
10. *Dey, S. & Mitra, S. (2019). "Predicting Heart Disease using Machine Learning." In: Chakraborty, C., Sanyal, S., Sarkar, A. (eds) Computing in Engineering and Technology. Springer, Singapore. [DOI: 10.1007/978-981-15-3563-0_28].*