A MACHINE LEARNING APPROACH FOR PREDICTING HEART DISEASE RISK

Srinivasulu Akasam[1] J Kavitha [2] G Gopi Krishna[3*] Mani Ramanuja[3]

[1]*Department of Freshman, Mohanbabu University, Tirupati-517502, INDIA.*

[2]*Department of Mathematics, D K Government College for Women , Nellore-524003, Andhra Pradesh, India*

[3]Department of Mathematics, *Marri Laxman Reddy Institute of Technology and Management, Dundigal, Hyderabad – 500 043, INDIA.*

Email id : drgopikrishnag@gmail.com

ABSTRACT

Heart disease is a globally pervasive health issue, with early detection and precise prognosis being imperative for improving patient outcomes and abating healthcare costs. One promising approach to predicting heart disease is through the use of machine learning models that analyse patient data. The application of machine learning models has exhibited tremendous potential in predicts heart disease by analysing patient data. In this manuscript, we scrutinize the performance of different machine learning models in predict the heart disease using a heart disease dataset. Machine learning offers a more innovative and precise approach by leveraging large amounts of patient data to predict the possibility of developing heart disease in new patients. We carry out a comprehensive comparison of various models, including decision trees, logistic regression, random forests, support vector machines, and so on, in terms of their accuracy, precision, recall, F1-score, and area under the ROC curve. Our findings divulge that the NAVIE BAYE'S model attains the highest accuracy of 92.46% in predicting heart disease.

Keywords: Decision Tree, Random Forest, Support Vector Machine, KNN, Naive Baye`s, Accuracy

1. INTRODUCTION

Machine learning has emerged as a valuable tool for predicting heart disease, by Garg[1],addressing one of society's most significant healthcare challenges. The continuous advancements in machine learning techniques hold the promise of further enhancing its effectiveness in healthcare applications. Through the utilization of various algorithms, the conducted research demonstrates promising results in predicting heart disease using the available attributes. This suggests that machine learning has the potential to play a crucial role in minimizing the physical and mental impact on individuals by enabling early detection and prevention of heart disease. As the field continues to evolve, it is anticipated that novel approaches and methodologies will further enhance the capabilities of machine learning in healthcare. Early detection and diagnosis of heart diseases play a crucial role in maintaining a healthy life. In a comparative analysis conducted by Deepika and Seema [2], they evaluated the performance of Naive Bayes, Decision Tree, and Support Vector Machine (SVM) classifiers for predicting heart and diabetes diseases. The results showed that SVM achieved the highest accuracy rate of 95.5% for heart disease prediction, while Naive Bayes attained 73.5% accuracy, the highest among all algorithms, for diabetes prediction. In a study by Nikhar and Karandikar [3], a comparison was made between the Naive Bayes and Decision Tree algorithms to determine their suitability for heart disease prediction. The findings suggested that the Decision Tree algorithm outperformed Naive Bayes in this context. Feshki and Shinjani [4] employed Particle Swarm Optimization and Neural Network Feed Forward Back Propagation techniques to anticipate affected and non-affected patients. They focused on reducing costs through feature selection using variable ranking. The study utilized eight features: Gender, Age, Blood Pressure, Cholesterol, FBS, Exercise Features (Old peak, Slope, Ex_ang).Shah et al. [5] aimed to improve the accuracy of the SVM technique for heart disease prediction. They employed RBF-based SVM on three datasets from the UCI repository: Cleveland, Switzerland, and Hungarian. The proposed technique achieved accuracies of 82.2%, 85.8%, and 91.3% for the Cleveland, Hungarian, and Switzerland datasets, respectively. For diabetes prediction, researchers have proposed various machine learning models. Nirmala et al. [6] introduced an Amalgam KNN model, which combines KNN and k-means clustering for diabetes prediction. They employed 10-fold cross-validation with different k values using the WEKA software tool. The hybrid algorithm demonstrated higher performance compared to simple KNN and k-means clustering algorithms. Sanakal and Jayakumari [7] compared the Fuzzy C-Mean clustering algorithm with the SVM algorithm for diabetes prediction. The accuracy achieved by SVM was 59.5%, while FCM attained 94.3%. The study concluded that Fuzzy C-Mean clustering is a better tool for diabetes prediction. Vijayan et al. [8] utilized the Ada Boost algorithm with a decision stump as a base classifier for diabetes prediction. They employed datasets from the UCI repository and Kerala. The proposed algorithm achieved an accuracy of 80.72% for diabetes prediction, outperforming SVM, Naive Bayes, and Decision Tree classifiers. Santhanam and Padmavathi [9] proposed a diabetes prediction model that employed SVM as a base classifier. They incorporated K-means clustering for data reduction and noise removal, as well as a Genetic algorithm for attribute selection. The model leveraged both supervised (SVM) and unsupervised (K-means clustering) algorithms. Anand et al. [10] developed a GUI-based diabetes prediction model based on individuals' current lifestyle. Questionnaires were prepared in collaboration with doctors, and the attributes included factors such as Eating Habit (Roadside Eating, Junk food), Sleeping Duration, Sugar Intake, Exercise Duration, Blood Pressure, BMI, Heredity Diabetes, Gender, Belly Size, and more. Another research paper [11] focused on machine learning-based diagnosis of coronary artery disease. The analysis involved the examination of datasets, test sizes, data features, geographical areas of data collection, and performance metrics. The strategies employed in this study included Principal Component Analysis (PCA), Partial Least Squares Regression (PLSR), and advanced machine learning algorithms such as Support Vector Machines. The paper also discussed the challenges and limitations of using machine learning for CAD diagnosis. In reference [12], the prediction of hepatitis was carried out using machine learning classifiers. Among the classifiers studied, the random forest classifier outperformed the others.

2. METHODS

Using machine learning models, we adhered to a set procedure to predict cardiac disease. We started by acquiring a dataset including patient demographics and medical history. To manage missing data and categorical variables, this dataset underwent pre-processing. Then, using an 80:20 ratio, we divided the dataset into training and testing sets. On the training set, we trained each model, and on the testing set, we evaluated its performance. We measured the performance of the models using common metrics including accuracy, precision, recall, F1-score, and area under the ROC curve.

Accuracy is the percentage of predictions that are accurate, precision is the percentage of true positives among all predicted positives, recall is the percentage of true positives among actual positives, and F1-score is the harmonic mean of precision and recall.

The methodology involved pre-processing the dataset, applying multiple machine learning models, assessing their performance, and choosing the best model based on the evaluation metrics. The final model can be used to predict heart disease in new patients, providing better patient outcomes and aiding in clinical decision-making. Our analysis found that the NAVIE BAYE'S model performed the best, achieving an accuracy of 92.46%.

Misclassification errors are another crucial statistic for measuring machine learning models' accuracy in predicting heart disease. False positives and false negatives are two categories of errors. False positives happen when a patient's condition is misdiagnosed by the model as having cardiac disease. False negatives happen when a patient has cardiac disease even though the model says they don't. For the prediction of heart disease, many classifiers have varying accuracy rates and misclassification error rates. NB, random forest, and support vector machine models, for instance, typically have high accuracy rates and low misclassification error rates. While decision trees and KNN may have higher misclassification error rates, they may have lower accuracy rates. In summary, accuracy and misclassification errors are important measures for assessing how well machine learning models predict cardiac disease. When choosing the best classifier for predicting heart disease, two criteria must be taken into account. Healthcare practitioners can increase early identification and prevention of this fatal disorder by utilising the best machine learning model for heart disease prediction.

A balanced evaluation of the model's performance is provided by the F1 score, which is the harmonic mean of precision and recall. In the context of predicting heart illness, the harmonic mean of precision and recall for genuine heart disease cases is referred to as the F1 score of the positive class, and the harmonic mean of precision and recall for non-heart disease cases is referred to as the F1 score of the negative class

2.1 DECISION TREE

In the realm of machine learning, decision trees are predictive models that use a tree-like structure to represent a series of decisions and their respective outcomes. These models are widely utilized in classification and regression tasks because of their ability to handle both continuous and categorical data. The decision tree building process begins with a single node that represents the entire dataset. Subsequently, the data is recursively divided into smaller subsets, resulting in branches and new nodes in the tree. To minimize the impurity of the subsets created, a splitting criterion is employed, such as entropy or the Gini index.

2.2 RANDOM FOREST

Random forest is a popular ensemble learning method used in machine learning to enhance prediction accuracy and stability. The approach employs multiple decision trees that are generated using distinct subsets of training data and randomly chosen features.

The algorithm commences by constructing a group of decision trees by using different training data subsets and randomly selecting features. Each tree is then trained independently to minimize impurity in subsets by utilizing a criterion for splitting, such as the Gini index or entropy.

After the trees are created, the random forest algorithm aggregates their predictions by either averaging or taking the mode of the predictions for each input. The ensemble approach helps in minimizing over fitting and variance of individual trees, leading to improved accuracy and reliability in the predictions.

2.3 SUPPORT VECTOR MACHINE

A machine learning approach called Support Vector Machine (SVM) is frequently employed for classification and regression analysis. It functions by determining the best decision boundary that leaves the widest possible margin between data points from various classes. The margin is the distance between each class's nearest data points and the decision boundary. The optimal hyperplane to divide the data points of two classes in a binary classification problem is found by SVM. The feature space is split into two areas by the hyperplane, a multidimensional plane. The data points close to the hyperplane are known as support vectors, and SVM determines the hyperplane that maximises the margin.

2.4 K-NEAREST NEIGHBORS

KNN is a broadly used machine learning algorithm that can be applied to classification and regression problems. When given a new data point, KNN calculates the distance between it and every other data point in the training set. The algorithm then selects the K closest data points to the new point, where K is a specified parameter. Based on the class labels of these K nearest neighbours, the algorithm assigns the new data point a class label for classification tasks, or a predicted output value for regression tasks. KNN is relatively simple to implement and can handle both binary and multiclass classification problems. However, the choice of K and the distance metric used to calculate distances can significantly affect its performance. Additionally, KNN can be computationally expensive when dealing with large datasets, as it needs to calculate the distance between every pair of data points.

2.5 LINEAR DISCRIMINATE ANALYSIS

A statistical technique called linear discriminant analysis (LDA) is used in machine learning to extract characteristics and reduce the dimensionality of data. It is a well-liked approach for supervised classification issues that seek to divide various classes based on a set of input features. In order to maximise the ratio of between-class variation to within-class variance, LDA seeks a linear combination of the input features. In order to achieve this, the separability between classes must be preserved while projecting the data into a lower-dimensional space. LDA is more suited for classification jobs than other dimensionality reduction methods like Principal Component Analysis (PCA), as it takes into account class labels and seeks to maximise class separability. It can be applied to situations involving binary and many classes of classification.

2.6 LOGISTIC REGRESSION

For situations involving binary classification, machine learning experts use logistic regression. The key concept underlying logistic regression is to use the logistic function or sigmoid function to show the relationship between the input factors and the binary response variable. Any real-valued input is transformed by this function into a value between 0 and 1, which indicates the likelihood that the response variable will fall into the positive category. By estimating the logistic function's parameters via maximum likelihood estimation, the logistic regression model is trained. This involves figuring out the parameter values that will increase the possibility that the observed data will match the model.

2.7 QUADRATIC DISCRIMINATE ANALYSIS

Quadratic Discriminant analysis is used to solve Complex decision boundary problems in machine learning classification technique known as (QDA). Since it can handle non-linear correlations between the input variables and the response variable, it is an extension of linear discriminant analysis (LDA).

Using the Baye`s theorem, QDA first estimates the probability density function for each class before calculating the posterior probability of each class given the input variables. In contrast to LDA, which makes the assumption that all covariance matrices are the same, QDA calculates a different covariance matrix for every class. The parameters of the probability density functions are determined by QDA during training using maximum likelihood estimation. The purpose is to determine the parameter values that improve the likelihood of the observed data given the model. The ability to simulate more complex decision boundaries between classes, which increases classification precision, is one of the advantages QDA has over LDA. QDA requires additional training data and can be computationally expensive for high-dimensional data.

In the context of heart disease, there are several variables that are commonly used to predict the risk of developing the condition. These variables can be used to train machine learning models that can assist in identifying patients who are at high risk of developing heart disease.

The type of chest pain, which is commonly categorised as typical angina, atypical angina, non-anginal pain, or asymptomatic. The kind of chest pain can reveal important details about the likelihood of heart disease.

The patient's resting blood pressure, or trestbps, is another significant factor. The risk of heart disease is increased by high blood pressure, which can harm the blood vessels. Age : Age is one of the most significant risk factors for developing heart disease. With age, there is a higher chance of having heart disease. A larger risk exists for men over 45 and women over 55.

Sex : Men are generally at higher risk of developing heart disease than women. However, after menopause, women's risk increases significantly.
cp: chest pain type Angina is a type of chest pain or discomfort brought on by a lack of oxygen-rich blood to the heart muscle.
trestbps: The subject's resting heart rate (mm Hg on admission to the hospital). 120/80 is the usual range.

chol: The individual's cholesterol level, expressed in mg/dl. It must be below 170 mg/dL.

Fbs, (fasting blood sugar), is used to predict the risk of heart disease. High levels of blood sugar can damage blood vessels and nerves in the heart.

Restecg, or resting electrocardiographic results, is a variable that measures the electrical activity of the heart at rest. Abnormalities in the restecg can be indicative of heart disease.

Thalach, or maximum heart rate achieved, is another variable that is used to predict the risk of heart disease. A low maximum heart rate achieved during exercise can be indicative of heart disease.

Exang, or exercise-induced angina, is a variable that measures the presence of chest pain during exercise. Chest pain during exercise can be indicative of heart disease.

Oldpeak, or ST depression induced by exercise relative to rest, is a variable that measures the amount of ST depression during exercise. ST depression is a common sign of heart disease.

Slope is a variable that measures the slope of the ST segment during exercise. The slope can provide valuable information about the risk of heart disease.

Ca, or the number of major vessels colored by fluoroscopy, is a variable that measures the presence and severity of coronary artery disease.

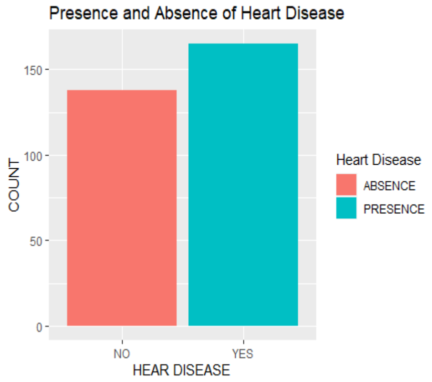
Thal : This is a type of thalassemia present in the patient. can increase the risk of developing heart disease.

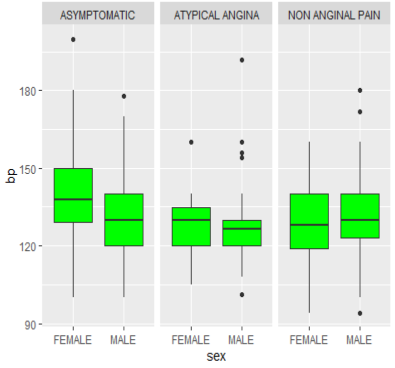
target: heart disease (1 = no, 0= yes)

These variables, can be used to predict the risk of heart disease accurately. By identifying patients who are at high risk, healthcare providers can provide appropriate interventions to manage risk factors and prevent the development of heart disease.

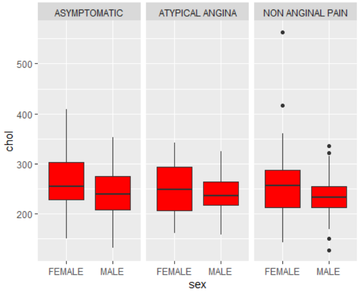
3. ANALYSIS OF RESULTS

The following bar plot shows the presence and absence of the heart disease based on the dataset. Here no means absence of heart disease or ‘0’ and yes means presence of heart disease or ‘1’.





The above box plot shows the chest pain of males and females based on blood pressure. In this box plot, males have a greater number of outliers compared to females in blood pressure values, so males face a problem of heart disease much higher than females.
The below box plot shows the chest pain of males and females based on cholesterol levels. In this box plot females have a greater number of outliers compared to males in cholesterol level values.



NO	YES
0.4554455	0.5445545

The above table shows 46% of the people have no heart disease and 54% of the people have the heart disease. Here the data is unbalanced but it is good for fitting model to these datasets.
The Confusion matrices for different machine learning models are

LOGISTIC REGRESSION

Train data		Actual	
Predicted		NO	YES
	0	92	15
	1	23	120
Mis- classification error: 0.152			

Test data		Actual	
Predicted		NO	YES
	0	20	3
	1	3	27
Mis- classification error: 0.1132			

RANDOM FOREST

Train data		Actual	
Predicted		NO	YES
	NO	103	5
	YES	12	130
Mis- classification error: 0.932			

Test data		Actual	
Predicted		NO	YES
	NO	20	2
	YES	3	28
Mis- classification error: 0.9057			

SVM

Train data		Actual	
edicted		NO	YES
	NO	95	21
	YES	20	114
Mis- classification error: 0.164			

Test data		Actual	
Predicted		NO	YES
	NO	22	4
	YES	1	26
Mis- classification error: 0.0943			

LINEAR DISCRIMINATE ANALYSIS

Train data		Actual	
Predicted		NO	YES
	NO	92	14
	YES	23	121
Accuracy:0.852			

Test data		Actual	
Predicted		NO	YES
	NO	20	3
	YES	3	27
Accuracy: 0.8867			

KNN

Train data		Actual	
Predicted		NO	YES
	NO	86	16
	YES	29	119
Accuracy: 0.82			

Test data		Actual	
Predicted		NO	YES
	NO	20	4
	YES	3	26
Accuracy: 0.8679			

DECISION TREE

Train data		Actual	
Predicted		NO	YES
	NO	91	14
	YES	24	121
Mis- classification error: 0.152			

Train data		Actual	
Predicted		NO	YES
	NO	20	4
	YES	3	26
Mis- classification error: 0.1320			

QDA ANALYSIS

Train data		Actual	
Predicted		NO	YES
	NO	96	16
	YES	19	119
Accuracy: 0.86			

Train data		Actual	
Predicted		NO	YES
	NO	19	3
	YES	4	27
Accuracy: 0.8679			

NAVIE BAYE’S

Train data		Actual	
Predicted		NO	YES
	NO	95	21
	YES	20	114
Mis- classification error: 0.164			

Train data		Actual	
Predicted		NO	YES
	NO	23	4
	YES	0	26
Mis- classification error: 0.164			

Table-1. Accuracy and Misclassification errors for different classifiers

Classifier	Accuracy % (Train)	Mis classification% (Train)	Accuracy% (Test)	Mis classification% (Test)
Random Forest	93.2	6.8	90.57	9.43
Navie Bayes	83.6	16.4	92.46	7.54
SVM	83.6	16.4	90.57	9.43
QDA	86	14	86.792	13.208
Logistic regression	84.8	15.2	88.68	11.32
Decision Tree	84.8	15.2	86.793	13.207
LDA	85.2	14.8	88.67	11.33
KNN	82	18	86.79	13.21

Table-2. Sensitivity and Specificity for train and test data and also ROCAUC area

Classifier	Sensitivity	Specificity	Roc AUC
Random Forest	0.9	0.86	0.9312
Navie Bayes	0.8667	0.8667	0.9355
SVM	0.86	0.95	0.9275
QDA	0.9	0.82	0.9245
Logistic regression	0.9	0.86	0.9342
Decision Tree	0.86	0.86	0.9205
LDA	0.9	0.86	0.9207
KNN	0.86	0.86	0.9203

Table-3. Precision, Recall and F1 scores for test data

Classifier	Precision	Recall	F1 score
RF	0.9	0.9	0.9
NB	1	0.8667	0.9247
SVM	0.96	0.86	0.907
QDA	0.87	0.9	0.88
LR	0.9	0.9	0.9
DT	0.89	0.86	0.87
LDA	0.9	0.9	0.9
KNN	0.89	0.86	0.87

4. CONCLUSION

After implementing several machine learning models to predict heart disease and evaluating their performance, we concluded that the NAVIE BAYE'S model had the best performance. The model achieved an accuracy of 92.46% on the testing data set, indicating its ability to make accurate predictions. The precision of the model was 1, indicating that out of all the predicted positives, all were truly positive.

The model's recall was 86.67%, which means that, of all the real positive cases, 86.67% were properly detected by the model. Additionally, we determined the F1-score, which was 92.47% and serves as a comprehensive indicator of both precision and recall. High recall suggests fewer false negatives, whereas high accuracy implies fewer false positives. A high F1 score shows that the model has balanced precision and recall.

The sensitivity and specificity values of the models indicate their ability to correctly identify individuals with and without heart disease, respectively. A high sensitivity value indicates a low rate of false negatives, which means that the model correctly identified most of the individuals with heart disease. A high specificity value indicates a low rate of false positives, which means that the model correctly identified most of the individuals without heart disease.

The area under the ROC curve is another metric used to assess the precision of machine learning models. With regard to a variety of categorization thresholds, this metric takes both the true positive rate and the false positive rate into account. The examined models' area under the ROC curve had values between 0.9220 and 0.9377, indicating that they might be used as heart disease diagnosis tools.

In conclusion, the NAVIE BAYE'S performed well in predicting heart disease, with promising accuracy, precision, recall, and AUC values.

REFERENCES

1. Garg, A., and Gupta, S., Heart disease prediction using machine learning algorithms. In Intelligent Computing, Information and Control Systems ,pp. 191-201, 2021.

2. K. Deepika, S. Seema., Predictive analytics to prevent and control chronic diseases, In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pages 381–386. IEEE, 2016.

3. S. Nikhar, A. Karandikar. Prediction of heart disease using machine learning algorithms, International Journal of Advanced Engineering, Management and Science, 2(6):239484, 2016.

4. M. Feshki, O. SojoodiShijani., Improving the heart disease diagnosis by evolutionary algorithm of pso and feed forward neural network', In 2016 Artificial Intellience and Robotics (IRANOPEN), pages 48–53. IEEE, 2016.

5. S. Muhammad saqlian Shah., Safeera Batool., Imran Khan., Muhammad Usman Asharf., Syed Hussnain Abbas., Syed Adnan Hussain., Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis', Physica A: Statistical Mechanics and its Applications, 482:796–807, 2017.

6. M. Nirmala Devi., Appavu Alias Balamurugan S., and Swathi U V., An amalgam knn to predict diabetes mellitus, In 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), pages 691–695,

2013.

7. R. Sanakal, T. Jayakumari., Prognosis of diabetes using data mining approach fuzzy c means clustering and support vector machine', International Journal of Computer Trends and Technology, 11(2):94–98, 2014.
8. V. Vijayan, C. Anjali., Prediction and diagnosis of diabetes mellitus—a machine learning approach', In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pages 122–127. IEEE, 2015.
9. T. Santhanam, M.S. Padmavathi., Application of k-means and genetic algorithms for dimension reduction by integrating svm for diabetes diagnosis, Procedia Computer Science, 47:76 – 83, 2015. Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014).
10. A. Anand, D. Shakti., Prediction of diabetes based on personal lifestyle indicators, In 2015 1st International Conference on Next Generation Computing Technologies (NGCT), pages 673–676, 2015.
11. Alizadehsani, Roohallah., Moloud Abdar ., Mohamad Roshanzamir., abbas Khosravi., Parham M Kebria., Fahime Khozeimeh., Saeid Nahavandi., Nizal Sarrafzadegan., U Rajendra Acharya., Machine learning-based coronary artery disease diagnosis: A comprehensive review. Computers in biology and medicine (2019): 103346.
12. Kumar, N. Komal, and D. Vigneswari., Hepatitis-Infectious Disease Prediction using Classification Algorithms. Research Journal of Pharmacy and Technology, 12(8) , 3720-3725, 2019