HEART DISEASE PREDICTION USING MACHINE LEARNING

Dr. Sireesha Moturi[1], B. Anusha[2], S. Jhansi[3],V.Sowmya Sree[4]
[1] Assoc. Professor, [2, 3 & 4] Student
[1]sireeshamoturi@gmail.com, [2] anuanusha5741@gmail.com, [3] jhansi.sid2003@gmail.com, [4] sowmyavendatam3@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

Abstract— **The primary cause of death globally is heart disease. This is a serious health problem that affects millions of people worldwide. By using machine learning to predict heart disease, we can work to improve early detection and prevention strategies. This study involved developing a model to predict heart disease risk using a variety of factors such as age, blood pressure, cholesterol levels and other medical data. We use machine learning algorithms including decision trees, random forests, support vector machines, k-nearest neighbors, and logistic regression. Custom selection uses a process to analyze reported data when performing a sample evaluation using competitive and performance metrics such as sensitivity, specificity, accuracy, and F1 scores. Our study aims to develop a predictive model to detect heart disease. According to a comparative study of these models, random forest and decision tree are more accurate in predicting heart disease, with 100% accuracy. This article comes after carefully reviewing the algorithms useful in predictive models to achieve the best results.**

Keywords— **Classification models like Random Forest Classification, KNN, Logistic Regression, Decision tree, NaïveBayes, Support Vector Machine.**

I.INTRODUCTION

Heart disease remains one of the biggest global health issues, accounting for a large portion of global morbidity and death rates. Early detection and accurate risk assessment of cardiovascular illnesses remain critical for successful preventative and intervention methods, despite advances in medical research and technology. In order to analyze complicated medical data and make highly accurate predictions about the course of diseases, machine learning techniques have emerged as potential tools in the healthcare industry [1].We propose a comprehensive framework for heart disease prediction utilizing machine learning algorithms. The main goal of this is to create and evaluate predictive models that can reliably estimate an individual's risk of developing heart disease based on a range of clinical, demographic, and lifestyle factors.Leveraging a diverse dataset comprising electronic health records, patient demographics and laboratory test results [2].

This study aims to capture multifaceted aspects of cardiovascular health and uncover hidden patterns and relationships within the data. By employing a variety of machine learning algorithms, we seek to build robust predictive models capable of generalizing across diverse populations and healthcare settings [3].Firstly, we ensure consistency and compatibility across the dataset by addressing missing values, encoding categorical variables, and standardizing numerical features. Additionally, feature selection techniques will be applied to identify the most informative predictors contributing to heart disease risk. Next, we assess several Machine Learning methods, such as K-Nearest Neighbors, Decision Trees, Random Forest, Logistic Regression, Gaussian NaiveBayes and Support Vector Machine. To assess an algorithm's performance and capacity to generalize to new data, it is initially developed on the preprocessed dataset and then subjected to cross-validation techniques [4].

To the extent that predictive models may be used in clinical practice, there are still issues that need to be handled. These include data quality, sample analysis and other issues. In order to reduce differences in cardiovascular therapy, it is essential to work on removing medical data differences and ensuring equal access to technology [5].Furthermore, in the context of clinical decision-making, this study will highlight the significance of model understanding and clarity. By employing feature selection strategies and visualization tools, we seek to improve comprehension of the underlying mechanisms of the prediction models and build acceptability and tolerance among clients and healthcare professionals. All things considered, heart disease prediction through machine learning algorithms is an advanced approach in cardiovascular medicine that has the potential to completely transform clinical decision-making, risk assessment, and patient treatment. Healthcare systems may proactively identify individuals at high risk of heart disease, implement targeted therapies, and eventually improve population-level health outcomes by leveraging the power of data-driven insights and predictive analytics.

II.LITERATURE SURVEY

Marimuthu M.et al. [1] aims to compare the performance of different machine learning methods in predicting heart disease. It explored the learning process and use of technology to determine the best course of action using data that included a wide range of patient characteristics and clinical characteristics. The most important thing to consider is heart disease. This findings inform the use of machine learning in healthcare and help develop more powerful predictive models for cardiovascular disease diagnosis and risk assessment.

Rajesh, N. et al. [2] focused on leveraging machine learning algorithms for heart disease prediction. It aims to develop predictive models capable of accurately identifying individuals at risk of heart disease based on various medical and demographic features. By employing a range of machine learning techniques, including traditional algorithms and possibly deep learning approaches, the project sought to enhance early identification and diagnosis of cardiac disease can ultimately lead to better patient outcomes and more informed healthcare decisions.

R. S. Singh et al. [3]. identified optimal machine learning classifiers for heart disease prediction, leveraging data from the International Journal of Control and Automation. Their study guides healthcare professionals in selecting effective machine learning techniques for improved cardiovascular risk assessment, advancing predictive modeling in cardiovascular medicine for better patient outcomes and healthcare delivery.

Asri, H. et al. [4]. gathered and preprocessed clinical data for thorough analysis in order to use machine learning to forecast cardiovascular illnesses. Through training, assessing, and fine-tuning models, they investigated logistic regression, decision trees, and neural networks with the goal of maximizing performance. In order to potentially improve public health outcomes, they hope to implement an accurate prediction model for early diagnosis and risk assessment of CVDs.

Swain et al. [5] uses various learning techniques including logistic regression, support vector machine, k-nearest neighbor, Gaussian Naive Bayes, decision tree classifier, and random forest classifier for accurate disease prediction, aiming to streamline risk assessment and improve outcomes.

Mhamdi, M. et al. [6] gather and preprocess a variety of datasets in order to use machine learning to create predictive models for determining the risk of heart disease. In order to improve the management of cardiovascular health, rigorous validation and fine-tuning are intended to give medical practitioners a highly advanced predictive tool for individualized risk assessment and preventative treatments.

III.PROPOSED WORK

The data analysis process begins with the collection of a dataset. Data extraction involves collecting data from a variety of sources. Once collected, outlier detection techniques are applied to identify and potentially remove anomalous data points that could impact results. Feature selection follows, where the most informative attributes are chosen to feed into the model. With a refined dataset, model training commences, utilizing machine learning algorithms to learn patterns and relationships within the data. After training, the model's performance is evaluated using appropriate metrics to assess its accuracy, reliability, and generalization capabilities. Finally, changes to the model or the analytic procedure that may be made in response to the evaluation results, resulting in continuous improvements.

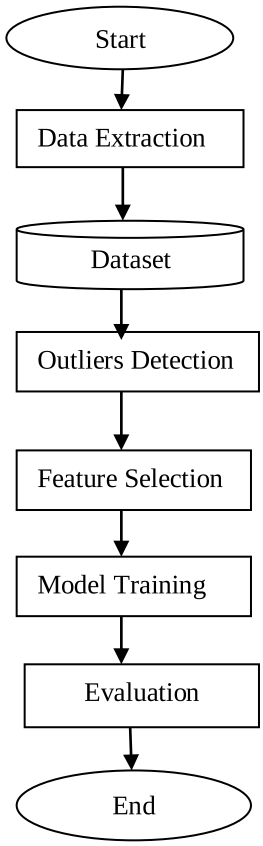


Fig.1.Block Diagram

A. Dataset Description:

The dataset is choosen from kaggle website[14].A selection of 14 variables including age, sex, blood pressure, cholesterol levels, fasting blood sugar level, resting electrocardiographic results, thalach, exercise induced angina, oldpeak, slope, ca, thalium, and target are used in the majority of research, out of the 76 attributes in the dataset. This dataset consists of 1025 rows and 14 columns, one of which is the aim as shown in fig.2.The target column will function as the label, while the thirteen attributes will function as the data. The dataset aims to forecast a patient's likelihood of having heart disease or not. The absence of null values, or empty values, is evident in the dataset.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
0	52	1	0	125	212	0	1	168	0	1.0
1	53	1	0	140	203	1	0	155	1	3.1
2	70	1	0	145	174	0	1	125	1	2.6
3	61	1	0	148	203	0	1	161	0	0.0
4	62	0	0	138	294	1	1	106	0	1.9
...	...	...	..	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0
1021	60	1	0	125	258	0	0	141	1	2.8
1022	47	1	0	110	275	0	0	118	1	1.0
1023	50	0	0	110	254	0	0	159	0	0.0
1024	54	1	0	120	188	0	1	113	0	1.4

	slope	ca	thal	target
0	2	2	3	0
1	0	0	3	0
2	0	0	3	0
3	2	1	3	0
4	1	3	2	0
...	...	..	...	...
1020	2	0	2	1
1021	1	1	3	0
1022	1	1	2	0
1023	2	0	2	1
1024	1	1	3	0

[1025 rows x 14 columns]

Fig.2.Dataset

Here the table1 represents attributes taken from the dataset. These attributes are used to predict heart disease presence or not.

Table.1.Description of the attributes

No	Features	Description	Scale
1	Age	Age in years	29 - 77
2	GD	Gender	Female (0), Male (1)
3	CP	Chest pain type	Typical angina (1), Atypical angina (2), Non-angina pain (3), Asymptomatic (4)
4	trestbps	Resting blood pressure on admission to the hospital (mm/Hg)	94 - 200
5	chol	Serum cholesterol (mg/dl)	126 - 564
6	Fbs	Fasting blood sugar is greater than 120 mg/dl	No (0), Yes (1)
7	Restecg	Resting electrocardiographic results	Normal (0), Having ST-T wave abnormality (1), Showing probable or definite left ventricular hypertrophy by Estes' criteria (2)
8	Thalach	Maximum heart rate achieved (ppm)	71 - 202
9	Exang	Exercise induced angina	No (0), Yes (1)
10	Oldpeak	ST depression induced by exercise relative to rest	0 - 6,2
11	slope	The slope of the peak exercise ST segment	Up sloping (0), Flat (1), Down sloping (2)
12	ca	Number of major vessels colored by fluoroscopy	0-3
13	Thal	The heart status	Normal (3), Fixed defect (6), Reversible defect (7)
14	num	Diagnosis of heart disease	Healthy (0), Patient has heart disease (1)

B. Preprocessing Techniques:

Preprocessing techniques refer to a set of procedures or methods used to prepare and clean data before it is analyzed or used for machine learning tasks. These techniques are crucial to ensure that the data is in a suitable format, free from errors, inconsistencies, and irrelevant information.

There is a class imbalance, as shown in the below fig.3. When a dataset has significantly more samples of one class than others, it can lead to class imbalance, a difficulty in machine learning. We apply SMOTE which is a technique used to address class imbalance in machine learning.The class is balanced after applying SMOTE, as shown in the figure below fig.4.

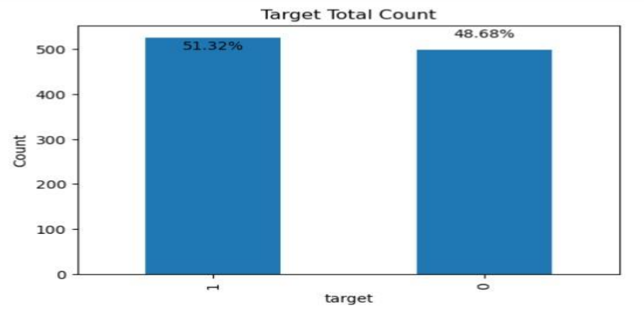


Fig.3.Class Imbalance

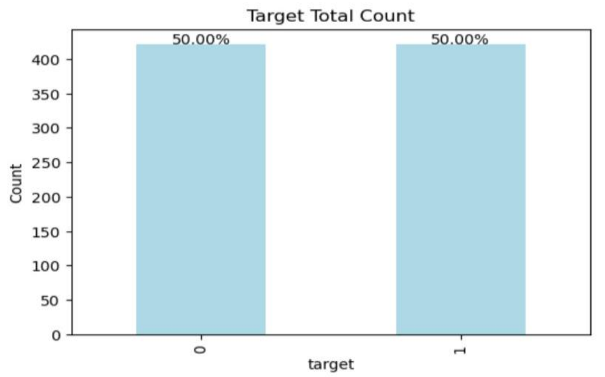


Fig.4.Class Balance

The dataset is loaded into memory. Initial cleaning steps are performed, such as handling missing values and dropping irrelevant columns. Our dataset has no missing values as shown in fig.5.Outliers are detected and removed from the dataset as shown in fig.6. and fig.7.

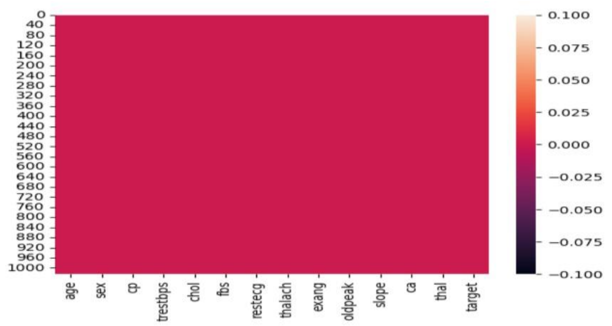


Fig.5.Clean Dataset

Outliers Detection & Removal:

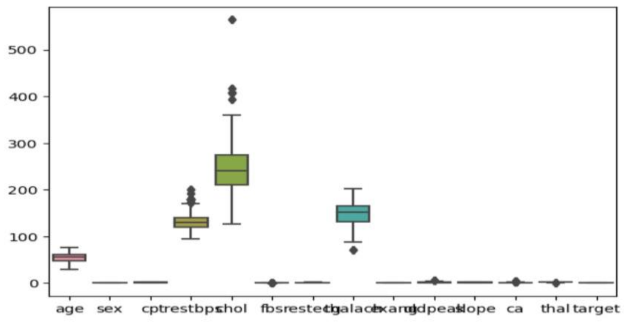


Fig.6.Outliers Detection

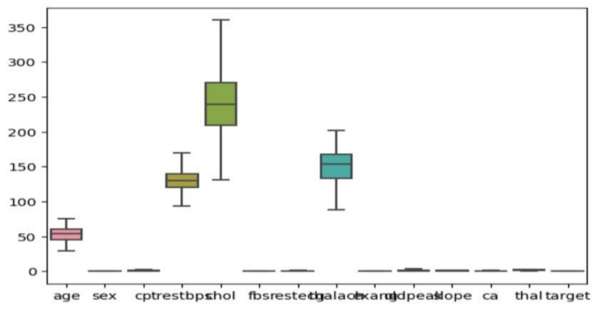


Fig.7. Outliers Removal

Correlation Matrix

The correlation matrix in Fig.8. can be used as a simple way to summarize the correlations between each variable in a dataset. It provides the dataset's variable relationships. By determining whether there are positive or negative correlations between dependent and independent features, it helps in feature selection. Weakly correlated traits might be eliminated to reduce overfitting. Redundancy and multi collinearity are reduced when only one of the associated variables is taken into account when correlation values are greater than 1. This matrix helps optimize models for better interpretability and performance.

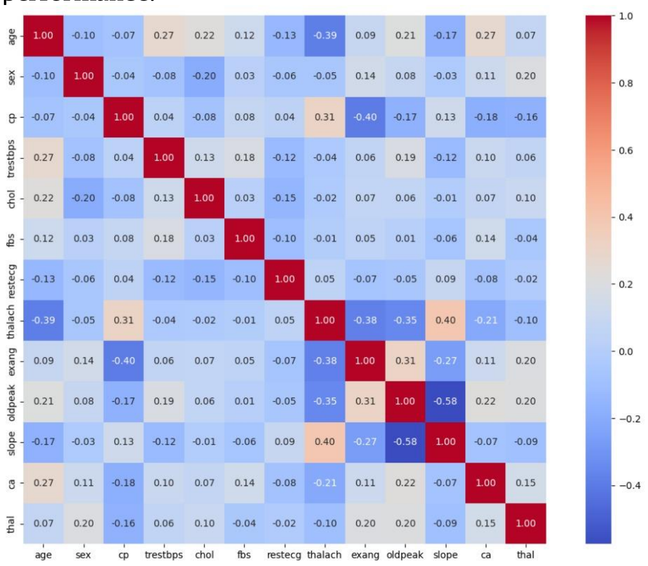


Fig.8.Correlation Matrix

C. Feature Selection:

Feature selection is a method that reduces sample variation by using key data and removing noise in the data.

D. Model Selection:

Here we use multiple classification algorithms for model selection. Here are the models that have been trained and evaluated. They are Random Forest Classifier, Logistic Regression, NaiveBayes, K-Nearest Neighbours, Decision Tree Classifier and Support Vector Machine.

E. Model Selection:

A supervised machine learning method is called logistic regression. For classification issues, it is employed. In which the objective is to estimate the likelihood that a given instance will belong to a class or not.

A decision tree is a technique for decision-making that is similar to a flowchart. It has the structure of a tree, where the leaves stand for results and the branches for options.

Random Forest is a popular machine learning algorithm that is applied to supervised learning methods. Its basis is the concept of ensemble learning, which is the act of combining multiple classifiers to improve the functionality of the model and solve a difficult problem.

K-Nearest Neighbors, is a popular machine learning algorithm used for classification and prediction. It works by finding the closest data points (neighbors) to a new data point and arranging a class in order of most of its neighbors.

A Support Vector Machine is a supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane (decision boundary) that best separates data points belonging to different classes in a high-dimensional space.

In Gaussian Naive Bayes continuous attributes are considered, and the data features are Guassian distributed throughout the dataset. A type of classification technique based on the Naive Bayes algorithm that works with continuous normally distributed features is called Gaussian Naive Bayes, and it is utilized in the Sklearn package.

IV.RESULTS AND ANALYSIS

We found that the models that were used are Random Forest, Decision Trees, Logistic Regression, Support Vector Machine and K-Nearest Neighbors performed differently as shown below:

I. Logistic Regression

With an accuracy of 86%, the logistic regression model proved to be a reliable classifier of data points. The ROC curve in the below image fig.9. shows that the model is performing well, with an AUC of 0.94.

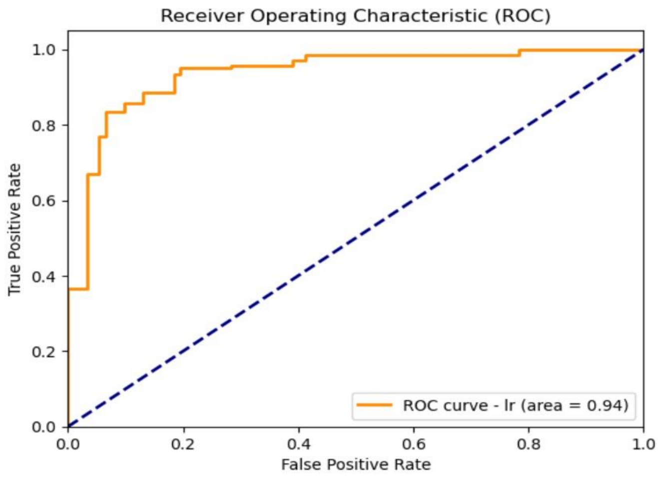


Fig.9. ROC Curve-Logistic Regression

A confusion matrix is a table that allows you to visualize the performance of an algorithm. In this case, it shows how many data points were correctly classified 123 and incorrectly classified 32 by the logistic regression model. Class 0 shines with a slightly higher recall of 0.92, effectively identifying 92% of true rejections. While its precision is a close 0.88, suggesting a good balance between correctly labeled rejections and avoiding false positives.

	precision	recall	f1-score	support
0	0.83	0.83	0.83	92
1	0.88	0.88	0.88	139
accuracy			0.86	231
macro avg	0.86	0.86	0.86	231
weighted avg	0.86	0.86	0.86	231

Fig.10 Metrics for Logistic Regression

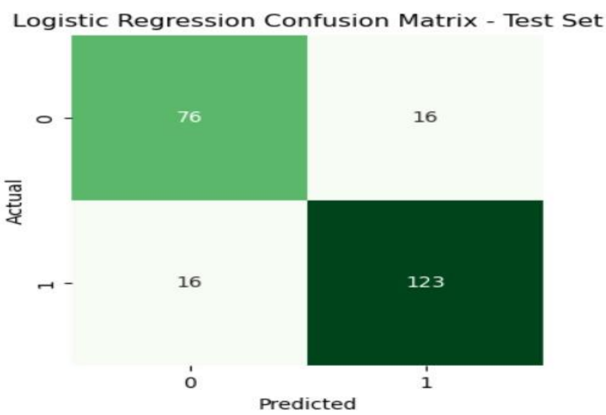


Fig.11.Confusion matrix of Logistic Regression

II. Decision Tree

The Decision Tree model demonstrated its efficiency in classifying tests with an accuracy of 100%. The ideal ROC curve fig.12. would be a straight line in the upper left corner of the graph. This would indicate that the model is able to correctly classify all positive cases without any false positive cases.

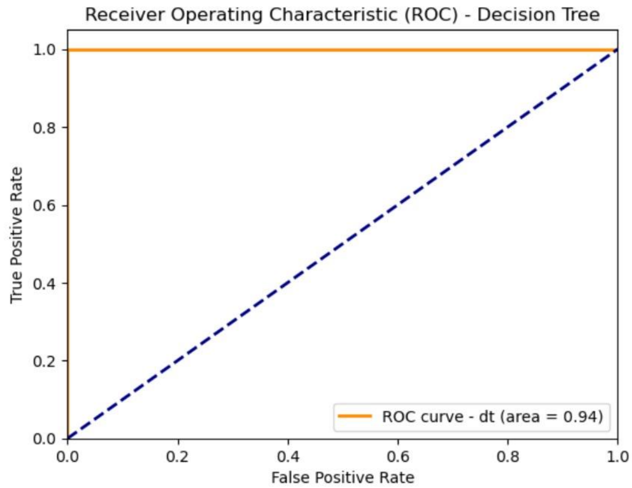


Fig.12. ROC Curve-Decision Tree

In a confusion matrix,this model made 139 correct predictions and 0 incorrect predictions.Class 0 has a recall of 1.00, which means it correctly identified all the actual positive cases.The precision of Class 0 is also 1.00, indicating that all the cases predicted as positive by the model were actually positive.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	92
1	1.00	1.00	1.00	139
accuracy			1.00	231
macro avg	1.00	1.00	1.00	231
weighted avg	1.00	1.00	1.00	231

Fig.13. Metrics for Decision Tree

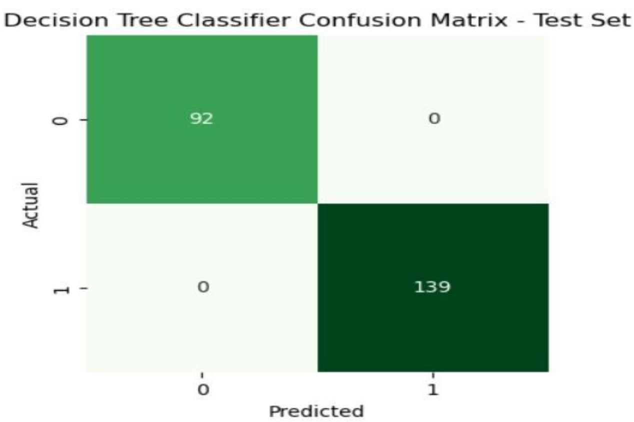


Fig.14.Confusion matrix of Decision Tree

III. Random Forest

Random Forest gives with an accuracy of 100%. A ROC curve, which shows in below fig.15. that the model can accurately categorize positive situations without producing any false positives, would look like a straight line in the graph's upper left corner.

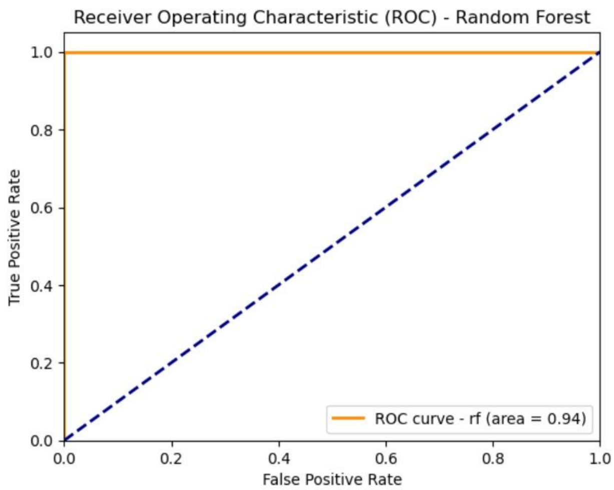


Fig.15.ROC Curve-Random Forest

The matrix shows the performance of a random forest classifier on a test set, and it indeed indicates that the model made 139 correct predictions and 0 incorrect predictions. Both precision and recall are 1.00, signifying the model correctly classified all Class 0 instances.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	92
1	1.00	1.00	1.00	139
accuracy			1.00	231
macro avg	1.00	1.00	1.00	231
weighted avg	1.00	1.00	1.00	231

Fig.16. Metrics for Random Forest

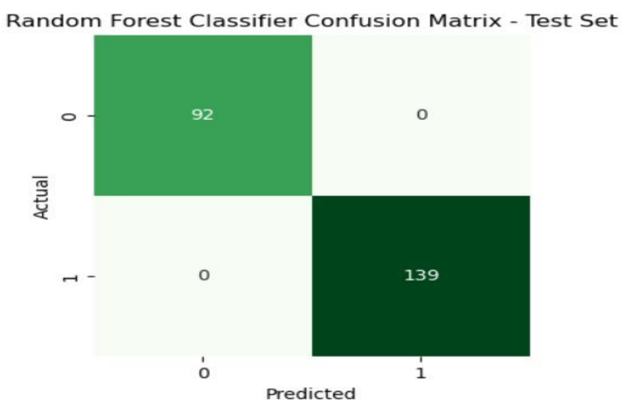


Fig.17.Confusion matrix of Random Forest

IV. KNN

The K-Nearest Neighbours model is good at identifying data items according to their closest neighbours,it reached its maximum accuracy of 82.25%. The ROC curve fig.18. indicates that there is potential for improvement in this instance for the KNN model because of its low AUC.

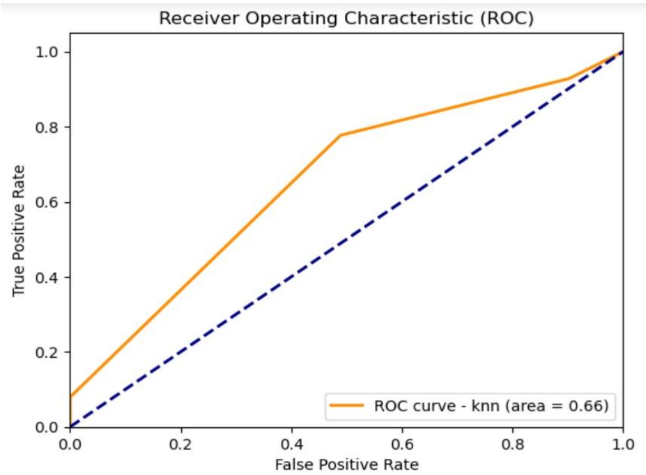


Fig.18. ROC Curve-KNN

The value in the top left cell, 9, represents the number of negative data points that the model correctly classified as negative. The precision and recall for Class 1 are not as high as those for Class 0, and the overall accuracy is also moderate at 0.54.

	precision	recall	f1-score	support
0	0.47	0.10	0.16	92
1	0.61	0.93	0.74	139
accuracy			0.60	231
macro avg	0.54	0.51	0.45	231
weighted avg	0.55	0.60	0.51	231

Fig.19. Metrics for KNN

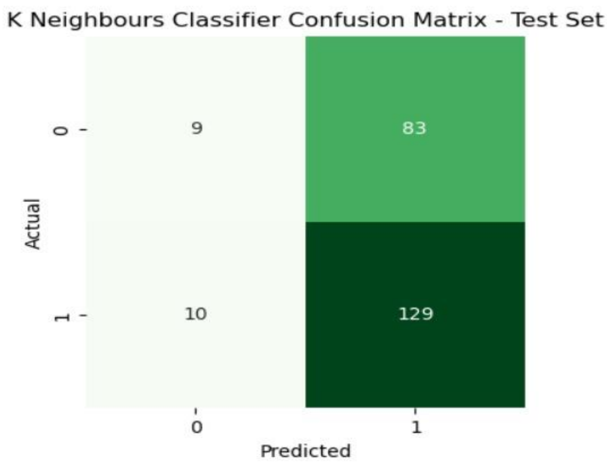


Fig.20.Confusion matrix of KNN

V. SVM

Possesing a respectable accuracy of 88.74%, the Support Vector Machine demonstrated its efficiency in classifying tests. The AUC of the curve fig.21. seems to be moderate. This indicates that although there is still potential for improvement, the model has exceeded random prediction to some measure.

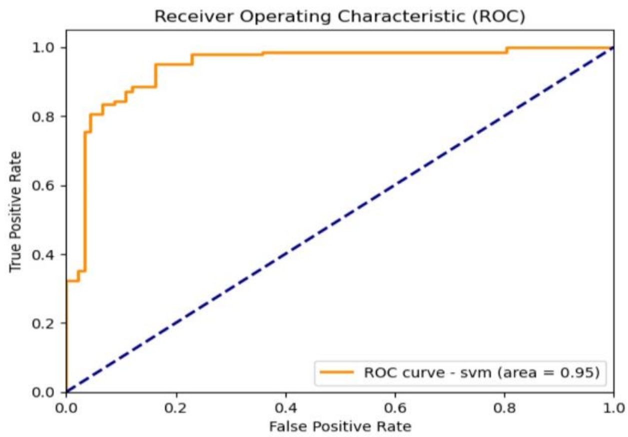


Fig.21. ROC Curve-KNN

This can be found in the bottom right cell of the matrix, which shows 128 True Negatives (TN) and 11 True Positives (TP).That the recall is also high means that the model is also good at identifying true positives. The metric known as recall quantifies the percentage of real positive cases that the model properly detected.

	precision	recall	f1-score	support
0	0.88	0.84	0.86	92
1	0.90	0.92	0.91	139
accuracy			0.89	231
macro avg	0.89	0.88	0.88	231
weighted avg	0.89	0.89	0.89	231

Fig.22. Metrics for SVM

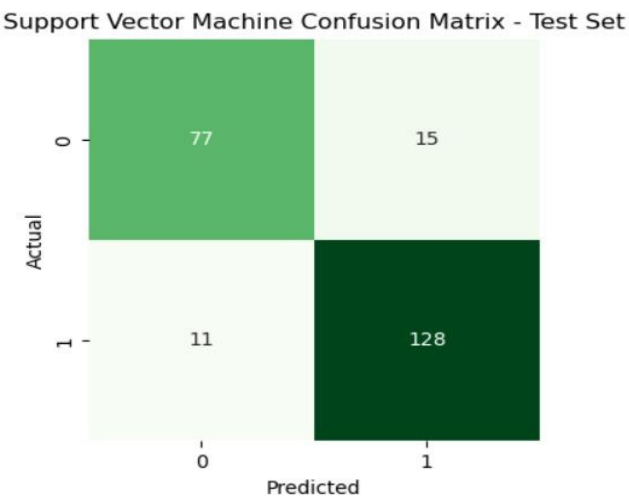


Fig.23.Confusion matrix of SVM

VI. Gaussian NB

The Gaussian NB model demonstrated its efficiency in heart disease prediction with an accuracy of 84.44%. With an area under the curve of 0.94, the ROC curve fig.24. in the below image shows good results. The model performs a good task of differentiating between positive and negative occurrences in this instance.

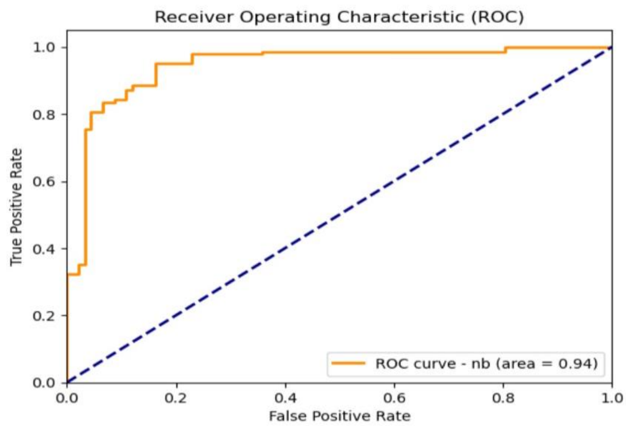


Fig.24.ROC Curve-Naïve Bayes

The confusion matrix you sent shows that the Naive Bayes model made 139 correct predictions and 0 incorrect predictions on the test set. This means that the model correctly predicted 77 instances that actually belong to class 0.This means the model did not make any incorrect predictions of class 0.

	precision	recall	f1-score	support
0	0.79	0.84	0.81	92
1	0.89	0.85	0.87	139
accuracy			0.84	231
macro avg	0.84	0.84	0.84	231
weighted avg	0.85	0.84	0.84	231

Fig.25. Metrics for Gaussian NB

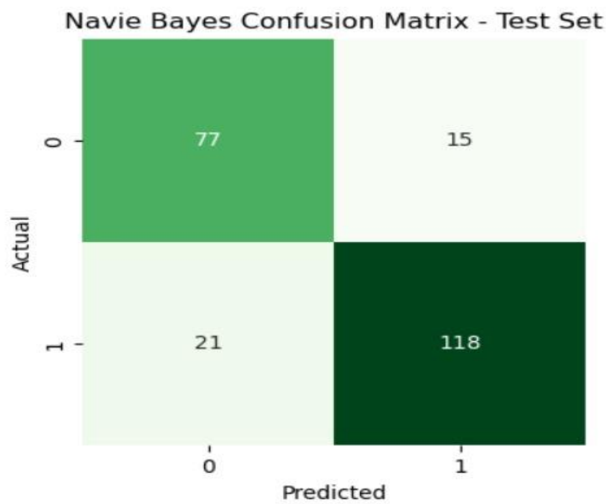


Fig.26.Confusion matrix of Gaussian NB

The below fig.25. and Table.2. shows accuracies of before and after applying SMOTE.

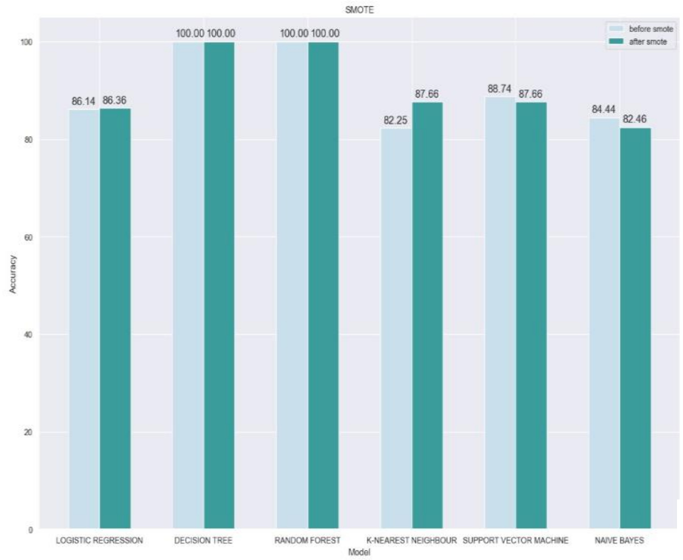


Fig.25.Accuracies of before and after SMOTE

Table.2.Accuracies of before and after SMOTE

ML Models	Accuracy before SMOTE	Accuracy after SMOTE
Logistic Regression	86.14	86.36
Decision Tree	100	100
Random Forest	100	100
KNN	82.25	87.66
SVM	88.74	87.66
Gaussian NB	84.44	82.46

Figure 26 and Table 3 below provide information on cross validation.

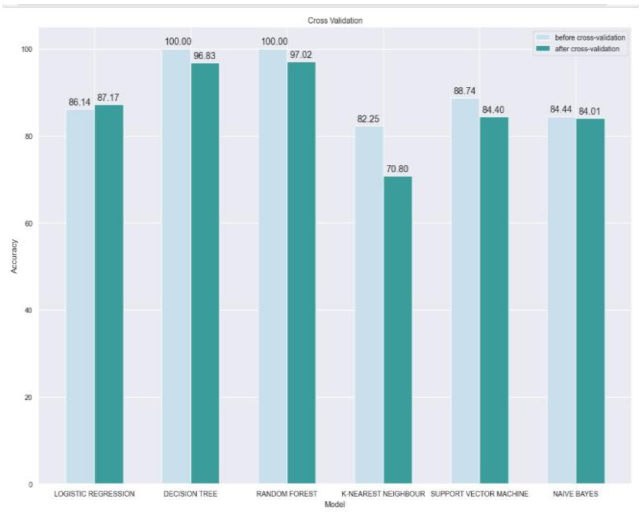


Fig.26.Accuracies of Cross Validation

Table.3.Accuracies of before and after Cross Validation

ML Models	Accuracy before Cross Validation	Accuracy after Cross Validation
Logistic Regression	86.14	87.17
Decision Tree	100	96.83
Random Forest	100	97.02
KNN	82.25	70.83
SVM	88.74	84.40
Gaussian NB	84.44	84.01

Fig.27.Accuracies of existing and proposed model

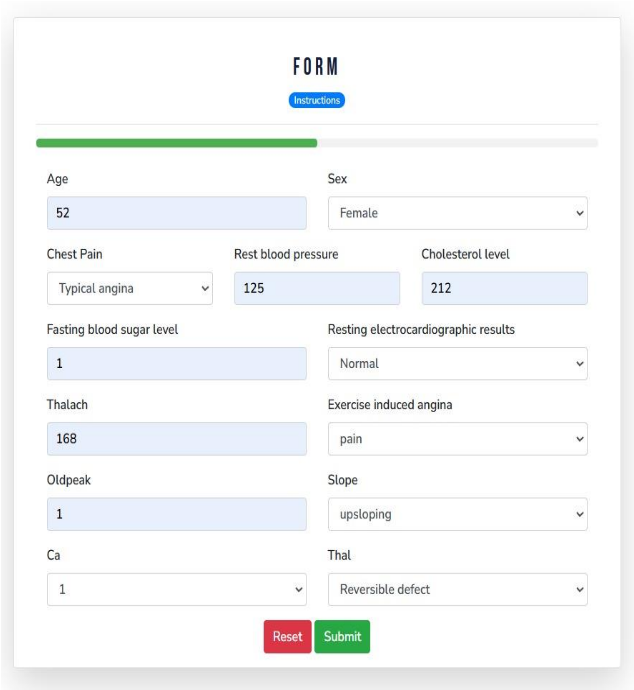
Table.3.Performance comparision of proposed work with existing work

Accuracies of the existed model	
Models used	**Accuracies**
SVM	85.18
Random Forest	92.15
KNN	74.07
Accuracies of the proposed model	
Models used	**Accuracies**
Logistic Regression	86.74
Decision Tree	100
Random Forest	100
KNN	82.25
SVM	88.74
Gaussian NB	84.44

Performance of existing work have been compared with the proposed work in table 4. M.S.GuruPrasad, J.kiran, D.K.SanthoshKumar, M.S.Pratap, S.Chandrappa and

Arnav Kotiyal predict heart disease with an accuracy value of 85.18,92.25,74.07 using KNN, Decision tree, SVM machine learning classifiers respectively. Where proposed model having higher accuracy rate compared to existed model.

V.USER INTERFACE:



NO DISEASE

VII.CONCLUSION

Predicting heart disease involves a number of procedures, including sample analysis, blood testing, and image processing. With machine learning, all of this is possible. Every machine learning model offers to produce positive results. Everything varies depending on the dataset used, and well-maintained, high-quality datasets yield accurate findings. In order to reduce human error and help medical professionals examine biomedical data more thoroughly, machine learning algorithms are now used in decision-making. The machine learning algorithms like decision trees, random forests, knn, logistic regression, gaussian Nb and support vector machines are compared in this work. The results obtained by each technique are 100,100,82.25,86.14,84.44 and 88.77 respectively. Comparing Decision Tree and Random Forest to other algorithms, that they perform better in the classification of heart disease. We use cross validation in this while random forests are more accurate than decision trees. We conclude that our analysis shows that the random forest works effectively.

REFERENCES

[1] Marimuthu, M., Deivarani, S., and Gayathri, R. Analysis of heart disease prediction using various machine learning techniques.In Advances in Computerized Analysis in Clinical and Medical Imaging (pp. 157-168). Chapman and Hall/CRC,2019

[2] Rajesh, N., Maneesha, T., Hafeez, S., and Krishna, H. Prediction of heart disease using machine learning algorithms. International Journal of Engineering and Technology (UAE) ,7 (2.32 SpecialIssue32) ,363-366,2018

[3] R. S. Singh, B. S. Saini, and R. K. Sunkaria, “Detection of coronary artery disease by reduced features and extreme learning machine,” Medicine and Pharmacy Reports, vol. 91, no. 2, pp. 166–175, 2018.

[4] Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. A Survey on Machine Learning Techniques in Cardiovascular Disease Prediction. Journal of King Saud University - Computer and Information Sciences, 30(4), 431–448. [DOI: 10.1016/j.jksuci.2016.03.011],2016

[5] Swain, Debabrata, Santosh Pani, and Debabala Swain.An efficient system for the prediction of Coronary artery disease using dense neural network with hyper parameter tuning. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8(6S).ISSN: 2278–3075,2019

[6] M'hamdi, M., "Heart Disease Prediction Using Machine Learning Algorithms: A Comparative

Study."Proceedings of the 1st International Conference on Artificial Intelligence and Machine Learning (AIML). ACM, New York, NY, USA, 2020. (DOI: 10.1145/3402942.3403076).

[7] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning Techniques.SN Computer Science. 2020 Oct 16; 1(6): 1–6. doi: https://doi.org/10.1007/s42979-020-00365-y.

[8] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," Second International Conference on Electronics,Communication and Aerospace Technology(ICECA),Coimbatore,India, pp.12751278,doi:10.1109/ICECA.2018.8474922.,2018

[9] Mezzatesta, S, Torino, C, De Meo, P, Fiumara, G, and Vilasi, A. A machine learning based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Comput Methods Prog Biomed. (2019) 177:9–15. doi: 10.1016/j.cmpb.2019.05.005

[10] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., and Mareeswari, V.(2018) “Prediction of cardiovascular disease using machine learning algorithms.” in International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 1–7,2018

[11] Padmanabhan, M, Yuan, P, Chada, G, and Nguyen, HV. Physician-friendly machine learning: a case study with cardiovascular disease risk prediction. J Clin Med. 8:1050. doi: 10.3390/jcm8071050,2019

[12] Akkaya, B., Sener, E. and Gursu, C.A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques. 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara,9-11 June 2022, 1-8.https://doi.org/10.1109/HORA55278.2022.9799978, 2019

[13] Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ. Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mob Inf Syst. 2020;2020:1–11https://doi.org/10.1155/2020/8843115.

[14]Kaggle website link https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[15] Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. Circulation. 1991;83(1):356–62. https://doi.org/10.1161/01.cir.83.1.356.

[16]Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Med Eng Phys. 2022;103825.