23BCE9474_L37+38_Cardio Vascular Disease Prediction Using Multiple Machine Learning Algorithms

YELCHURI J.S.S.MANI DEEPAK 23BCE9474

February 2025

S No	Title	Authors	Year	Algorithms Used	Dataset	Accuracy	Source
1	Evaluation of Machine Learning Models for Predicting Cardiovascular Disease	Suhatril et al	2024	Decision Tree, Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest, Logistic Regression, Neural Network, and Gradient Boosting	Framingham Heart Study dataset	84%.	[15]
2	Machine Learning for Cardiovascular Disease Prediction	V.P Amudhini, T. Santhini, Pavitra Kailash, D. Nivetha	2018	SVM, Bayesian Classifier	UCI Heart Disease dataset (303 records)	92% (SVM), 85% (Naïve Bayes)	[2]
3	A Comprehensive Study on Early Prevention and Detection of Cardiac Health Issues Using Machine Learning and Deep Learning Algorithms	Mangesh Limbitote, Kedar Damkondwar, Dnyaneshwari Mahajan	2024	Random Forest, Linear Model	Public Heart Disease Dataset	88.7%	[7]
4	Effective Prediction of Cardiovascular Diseases	J. Amutha, K. Ruba Soundar, M. Piramu, K. Murugesan	2021	ANN, SVM, Decision Tree	Cleveland Heart Disease dataset	90.12% (SVM), 86.42% (Decision Tree)	[3]
5	A Comparative Study on ML Algorithms for Heart Disease Prediction	Jagdeep Singh, Poornima Singh	2022	KNN, ANN, Decision Tree	Publicly available CVD dataset	88.3% (ANN), 85.2% (KNN)	[13]
6	Cardiovascular Disease Prediction using Neural Networks	Senthil Kumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava	2019	Multi-layer Perceptron (MLP)	Kaggle CVD Dataset	91%	[8]
7	Machine learning Techniques For Heart Disease Prediction	Rifki Wijaya, Tulay Karayilan, Ozkan Kilic	2013	Artificial Neural Network	Custom Hospital Dataset	95%	[11]
8	Prediction of Heart Disease Using Machine Learning	Sanjay Kumar Sen	2017	Naïve Bayes, Bayesian Network, SVM, KNN	Kaggle Heart Disease Dataset	83%	[12]
9	AI-driven Approach for Early Detection of Cardiovascular Disease	Ekundayo, Foluke and Nyavor, Hope	2024	Hybrid Ensemble Model (SVM + Decision Tree)	UCI CVD Dataset	93%	[4]

10	Prediction of Heart Disease Using Machine Learning Algorithms	Prashasti Kanikar	2016	SVM (RBF), Naïve Bayes	UCI Heart Disease dataset (303 records)	57%, 52%	[5]
11	Prediction of Heart Disease Using Machine Learning Algorithms	Sonam Nikhar	2016	Naïve Bayes Classifier, Decision Tree	Cleveland Heart Disease database	Decision tree has better accuracy compared to Naïve Bayes	[9]
12	Prediction Heart Diseases using Associative Classification	Jagdeep Singh et al.	2016	Association and Classification technique	Cleveland Heart Disease database	99.19%	[14]
13	Diag- nosis and Medical Prescription of Heart Disease Using SVM and Feedforward	Shaikh Abdul Han- nan et al.	2010	SVM, Feedforward	Cleveland CVD Dataset	50%-60%	[1]
14	Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers	Kumar, N Komal and Sindhu, G Sarika and Prashanthi, D Krishna and Sulthana, A Shaeen	2020	Random Forest ,Decision Tree,Logistic Regression,Support Vector Machine ,K-Nearest Neighbor	UCI Heart Disease dataset (303 records)	68.57%-85.71%	[6]
15	Machine learning based algorithm for risk prediction of cardio vascular disease (Cvd)	Patil, Prasadgouda B and Shastry, P Mallikarjun and Ashokumar, PS and others	2020	NN(Neural Networks), Gaussian NB,Decision Tree Classifier, Kneighbor Classifier	UCI Heart Disease dataset (303 records)	68%-98%	[10]

References

[1] S. Abdul, V. Bhagile, R. Manza, and R. Ramteke. Diagnosis and medical prescription of heart disease using support vector machine and feedforward backpropagation technique. *International Journal on Computer Science and Engineering*, 1(6):2150–2159, 2010.

[2] V. Amudhini, T. Santhini, P. Kailash, D. Nivetha, and R. Poonguzhali. Survey on machine learning algorithms for prediction of cardiovascular disease. 2018.

[3] M. J. Amutha, K. R. Soundar, M. Piramu, and K. Murugesan. A survey on machine learning algorithms for cardiovascular diseases predic-tion. *IJIRMPS*, 9:45–48, 2021.

[4] F. Ekundayo and H. Nyavor. Ai-driven predictive analytics in cardiovascular diseases: Integrating big data and machine learning for early diagnosis and risk prediction.

[5] P. Kanikar and D. R. Shah. Prediction of cardiovascular diseases using support vector machine and bayesien classification. *International Journal of Computer Applications*, 156(2):9–13, 2016.

[6] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, and A. S. Sulthana. Analysis and prediction of cardio vascular disease using machine learning classifiers. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 15–21. IEEE, 2020.

[7] M. Limbitote, K. Damkondwar, and D. Mahajan. A comprehensive study on early prevention and detection of cardiac health issues using machine learning and deep learning algorithms. *ResearchGate*, 2024. URL https://www.researchgate.net/publication/376648784_A_Comprehensive_Study_on_Early_Prevention_and_Detection_of_Cardiac_Health_Issues_Using_Machine_Learning_and_Deep_Learning_Algorithms.

[8] S. Mohan, C. Thirumalai, and G. Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554, 2019.

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

[10] P. B. Patil, P. M. Shastry, P. Ashokumar, et al. Machine learning based algorithm for risk prediction of cardio vascular disease (cvd). *Journal of critical reviews*, 7(9):836–844, 2020.

[11] K. R. Priya, P. Ramya, S. Pavithra, and M. M. Yaseen. Self-analysis of heart disease using machine learning. 2020.

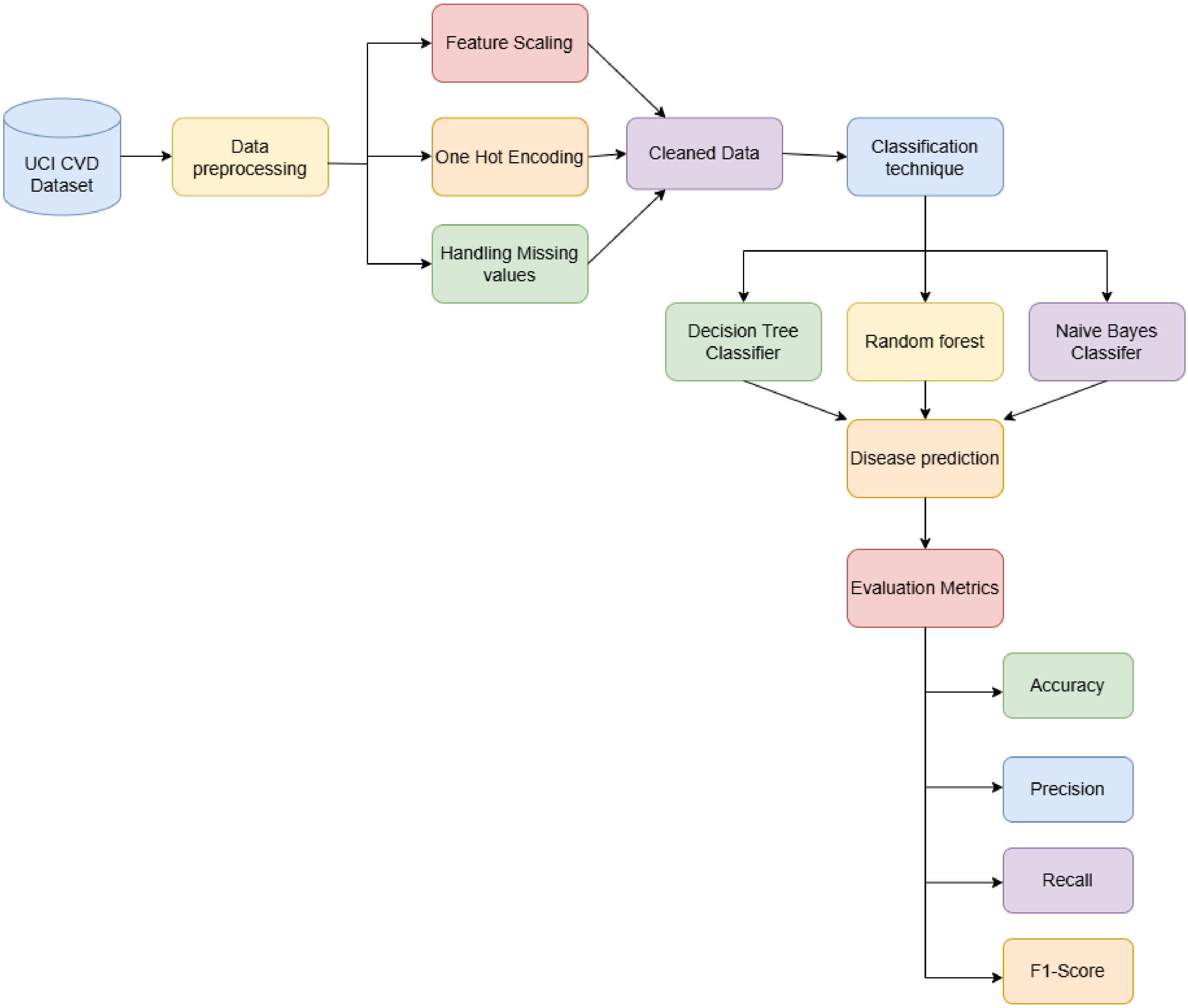
[12] S. K. Sen. Predicting and diagnosing of heart disease using machine learning algorithms. *International Journal of Engineering and Computer Science*, 6(6):21623–21631, 2017.

[13] J. Singh and P. Singh. A comparative study of heart disease prediction using machine learning. *CEUR Workshop Proceedings*, 3635:1–8, 2022. URL `https://ceur-ws.org/Vol-3635/ICCS_CVMLH_01.pdf`.

[14] J. Singh, A. Kamra, and H. Singh. Prediction of heart diseases using associative classification. In *2016 5th International conference on wireless networks and embedded systems (WECON)*, pages 1–7. IEEE, 2016.

[15] Suhatril et al. Evaluation of machine learning models for predicting cardiovascular disease based on framingham heart study data. *ILKOM Jurnal Ilmiah*, 16(1):68–75, 2024. URL `https://www.researchgate.net/publication/381311088_Evaluation_of_Machine_Learning_Models_for_Predicting_Cardiovascular_Disease_Based_on_Framingham_Heart_Study_Data`.

1 Architecture Diagram



Architecture Desription:

1. **Dataset:** The **UCI CVD dataset** is used as input for training the model.
2. **Data Preprocessing:** The raw data is preprocessed to improve model performance. This involves:
    - **Feature Scaling:** Normalizing or standardizing numerical data to bring all features to a similar scale.
    - **One-Hot Encoding:** Converting categorical variables into numerical format.
    - **Handling Missing Values:** Imputing or removing missing data to ensure data consistency.
3. **Cleaned Data:** After preprocessing, the dataset is **cleaned** and ready for model training.
4. **Classification Techniques:** The cleaned data is fed into different classification algorithms:
    - **Decision Tree Classifier**
    - **Random Forest**
    - **Naïve Bayes Classifier**
5. **Disease Prediction:** The selected model predicts whether a patient has cardiovascular disease (CVD) based on input features.
6. **Evaluation Metrics:** The performance of the model is evaluated using various metrics:
    - **Accuracy:** Measures overall correctness of predictions.
    - **Precision:** The proportion of true positive predictions among all positive predictions.
    - **Recall:** The proportion of true positive predictions among all actual positive cases.
    - **F1-Score:** A balance between precision and recall.

2 Problem statement

Cardiovascular diseases (CVDs) are among the leading causes of death globally, emphasizing the need for early detection and timely intervention. Existing diagnostic techniques often rely on extensive clinical evaluation, which can be time-consuming and subject to human error. This project focuses on developing a machine learning-based predictive system that can accurately assess the likelihood of a patient having cardiovascular disease using clinical data. By comparing multiple algorithms—namely Decision Tree, Naïve Bayes, and Random Forest—the model identifies the most effective approach for prediction. After thorough evaluation, the Random Forest Classifier is selected as the optimal model due to its superior accuracy and robustness. The system is built and validated using the UCI Heart Disease dataset to support healthcare professionals in making faster and more reliable diagnostic decisions.

3 Dataset Visualization

The dataset used in this project is UCI Heart Disease dataset,which consists of 920 records with 14 attributes, including demographic data, medical measurements, and the target variable indicating the presence of heart disease. Some key attributes visualized below include age distribution, cholesterol levels, resting blood pressure, and maximum heart rate.

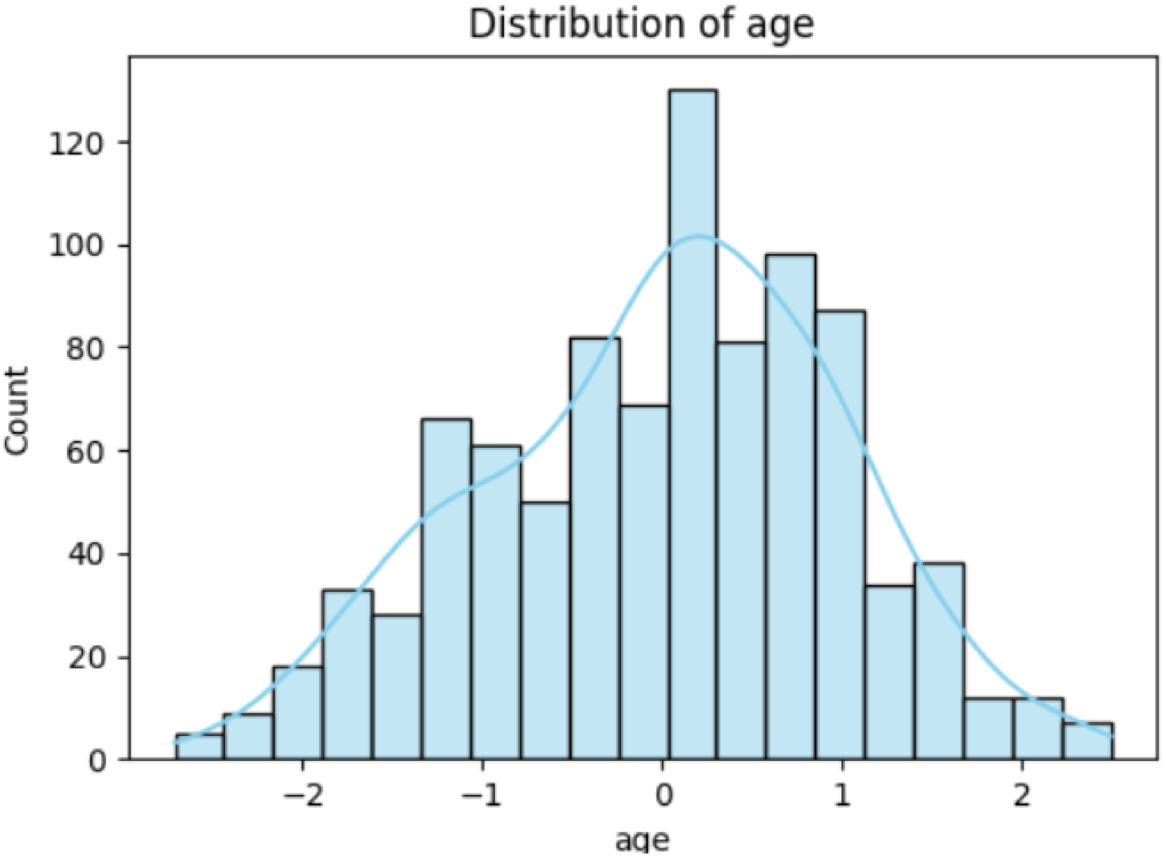


Figure 1: Age Distribution in the Dataset

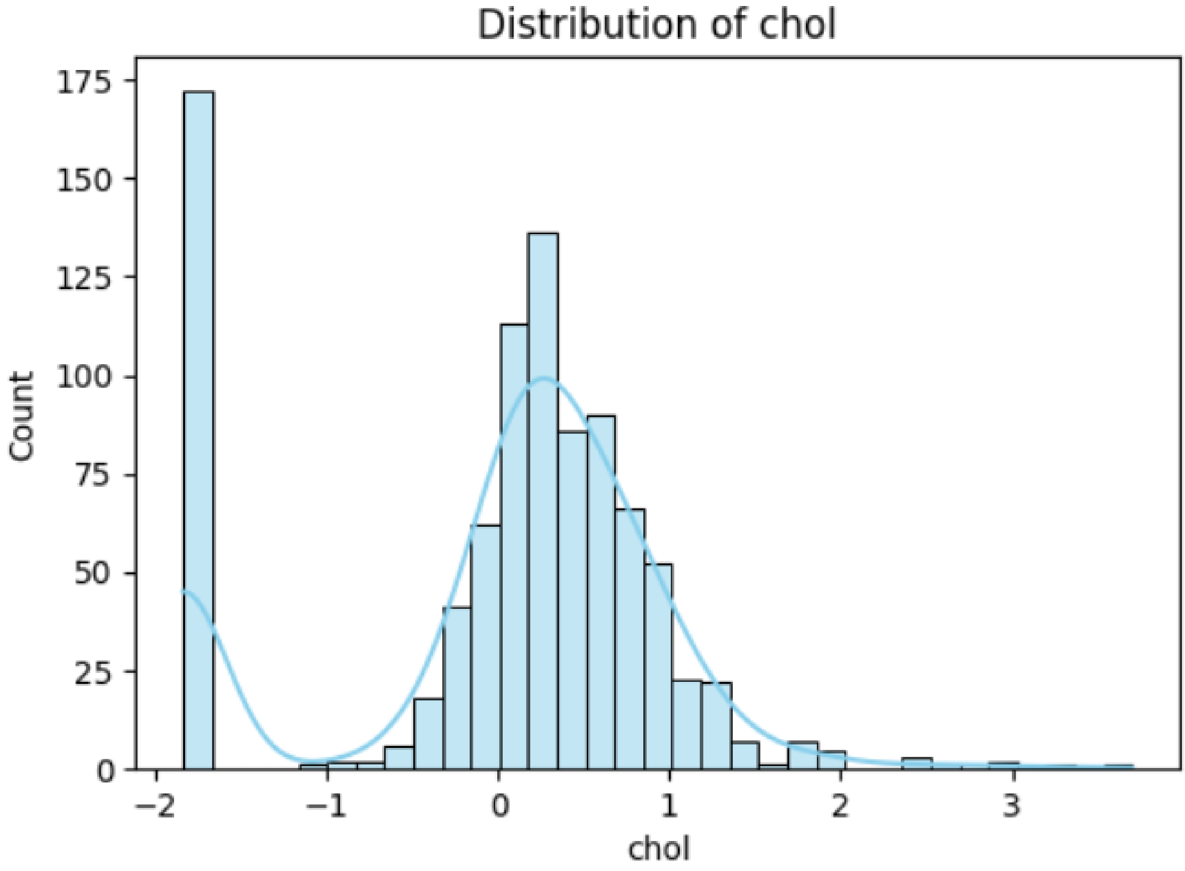


Figure 2: Cholesterol Level Distribution

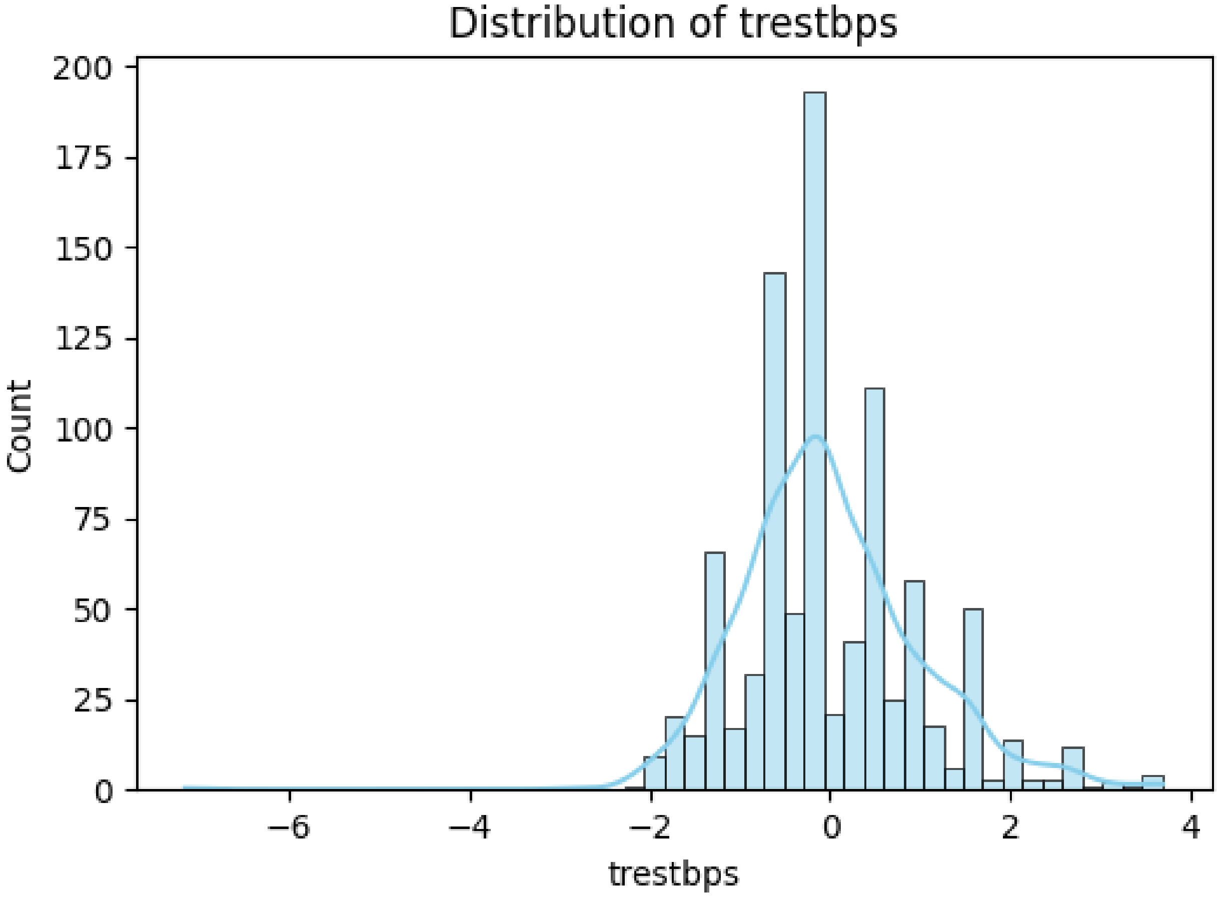


Figure 3: Resting Blood Pressure Distribution

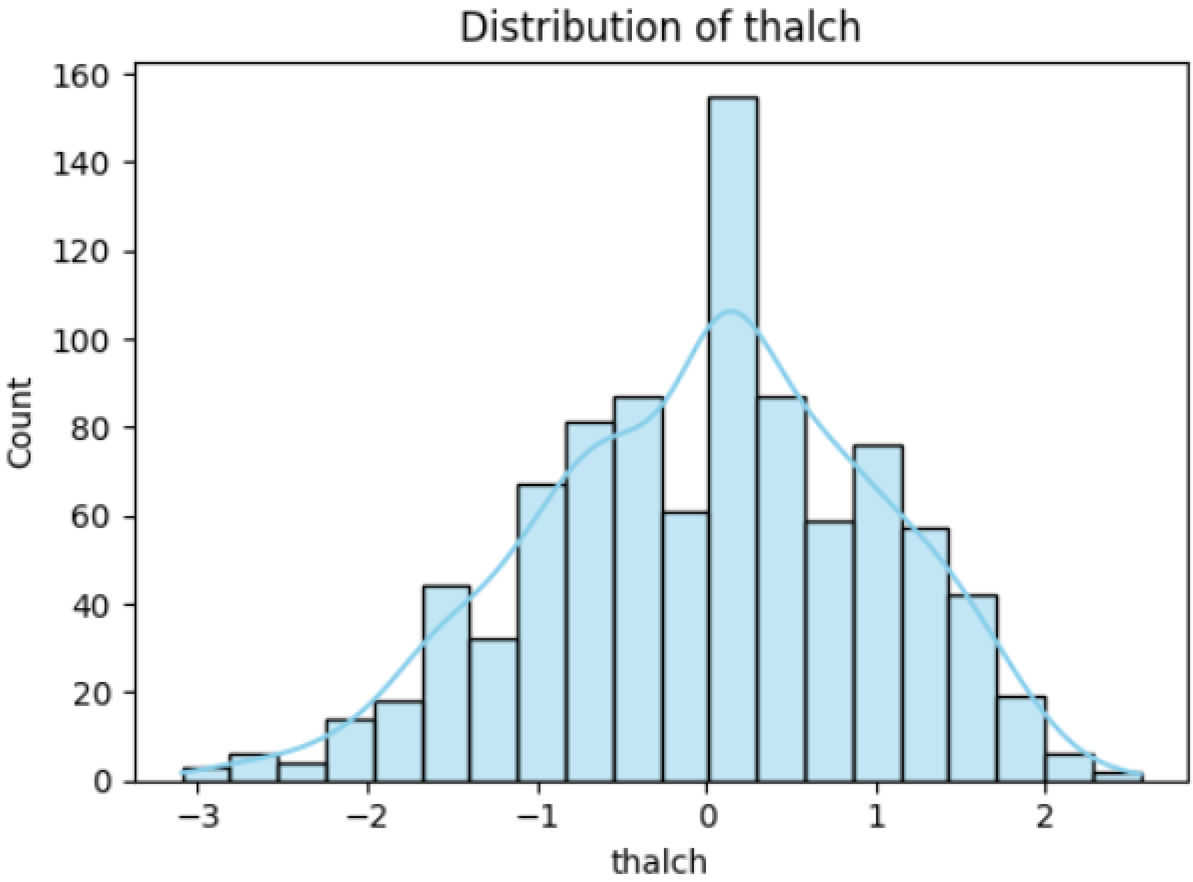


Figure 4: Max Heart Rate Achieved Distribution

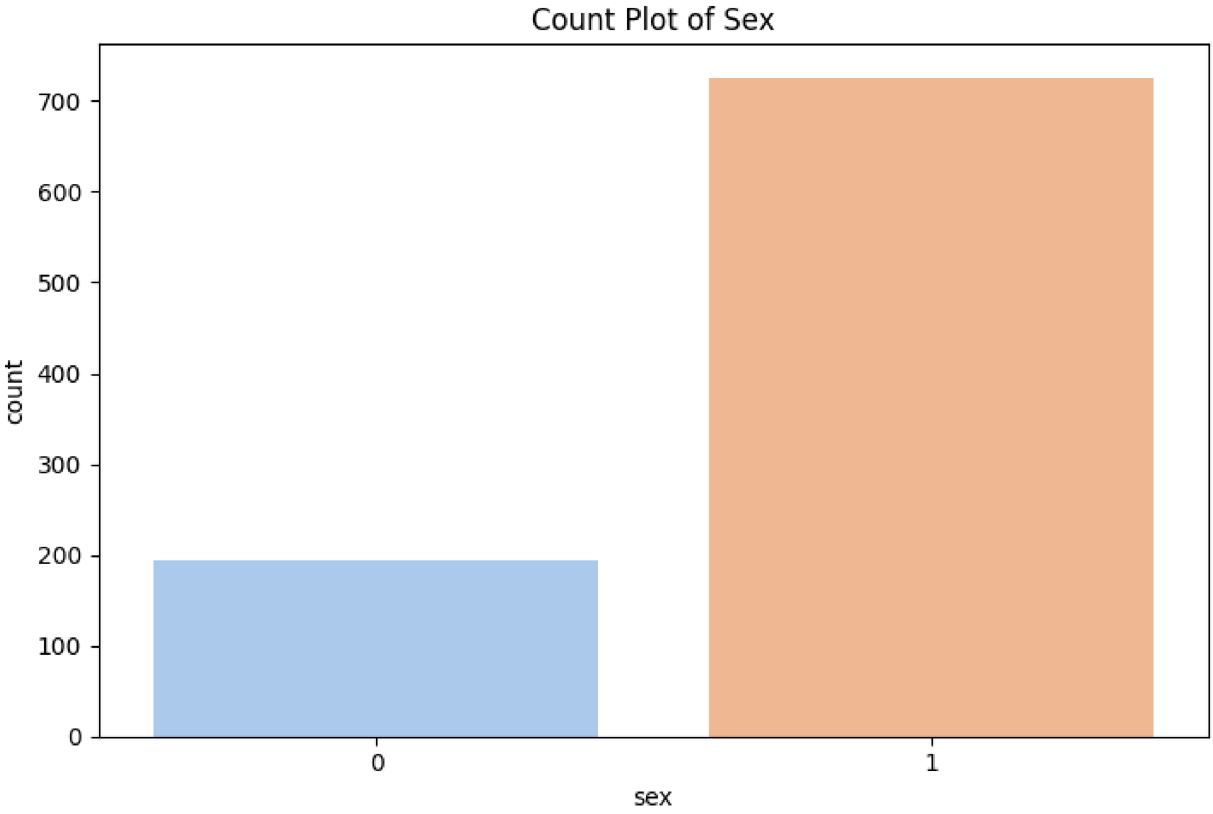


Figure 5: Number of males(1) and females(0)

4 Dataset Description

The dataset used is the UCI Heart Disease dataset, which consists of 920 instances and 14 medical features relevant to heart disease prediction. The attributes are described in the following table:

Feature	**Description**
Age	Age of the patient
Sex	Gender (0 = Female, 1 = Male)
Chest Pain Type (cp)	Type of chest pain (0–3)
Resting Blood Pressure (trestbps)	Resting blood pressure in mm Hg
Cholesterol (chol)	Serum cholesterol in mg/dl
Fasting Blood Sugar (fbs)	> 120 mg/dl (1 = True; 0 = False)
Resting ECG (restecg)	Results of resting electrocardiograph (0–2)
Maximum Heart Rate (thalach)	Achieved maximum heart rate
Exercise-Induced Angina (exang)	Exercise-induced angina (1 = Yes; 0 = No)
Oldpeak	ST depression induced by exercise
Slope	Slope of the peak exercise ST segment
Number of Major Vessels (ca)	Number of major vessels colored by fluoroscopy (0–3)
Thalassemia (thal)	1 = Normal, 2 = Fixed defect, 3 = Reversible defect
Target	Diagnosis of heart disease (0 = No, 1 = Yes)

Table 2: Description of Features in the UCI Heart Disease Dataset

5 Algorithm Description

5.1 Decision Tree Classifier

The Decision Tree algorithm is a supervised learning method used for classification and regression tasks. It splits the dataset into smaller subsets based on feature values using a tree-like structure. The splitting is done recursively until all subsets belong to a single class or a stopping condition is met.

- The algorithm selects the best feature to split the data using measures like Gini impurity or entropy.
- Each split creates branches, dividing the dataset into smaller groups.
- This process continues recursively until the leaves (terminal nodes) are reached.
- A new data point is classified based on the path it follows in the tree.

Working Mechanism:

Formulae: The Gini impurity is given by:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

where p_i is the probability of a particular class at a node.
Entropy, another splitting criterion, is given by:

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2}$$

Parameters:

- **Max depth**: Limits the depth of the tree to prevent overfitting.
- **Min samples per leaf**: The minimum number of samples required at a leaf node.
- **Criterion**: The function used to measure the quality of a split (Gini or entropy).

Decision Trees are easy to interpret and visualize. They can handle both numerical and categorical data, but they may overfit without proper parameter tuning.

5.2 Naïve Bayes Classifier

The Naïve Bayes classifier is a probabilistic machine learning model based on Bayes' Theorem, assuming independence between features. It is particularly useful for classification tasks with categorical data.

Working Mechanism:

- The algorithm calculates prior probabilities for each class.
- It computes the likelihood of each feature given a class using the probability distribution (e.g., Gaussian for continuous data).
- Using Bayes' theorem, the posterior probability of a class is calculated for a given input.
- The class with the highest posterior probability is selected.

Formulae: Bayes' theorem is given by:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{3}$$

where:

- $P(C_k|X)$ is the posterior probability of class C_k given feature X.
- $P(X|C_k)$ is the likelihood of X given class C_k.
- $P(C_k)$ is the prior probability of class C_k.
- $P(X)$ is the probability of feature X.

Parameters:

- **Smoothing parameter**: Adds a small value to avoid zero probabilities.
- **Distribution assumption**: Gaussian, multinomial, or Bernoulli depending on the data.

Despite its simplicity, Naïve Bayes often performs well in real-world situations, particularly in text classification and spam filtering.

5.3 Random Forest Classifier

The Random Forest algorithm is an ensemble learning technique that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting.

Working Mechanism:

- The dataset is randomly sampled with replacement (bootstrap sampling).
- Multiple decision trees are trained on different subsets of the data.
- Each tree makes an independent prediction.
- The final prediction is determined by majority voting (for classification) or averaging (for regression).

Formulae: The prediction in a Random Forest classifier is given by:

$$\hat{y} = \text{mode}(y_1, y_2, ..., y_T) \tag{4}$$

where y_i represents the prediction from each individual tree.
For regression, the final output is:

$$\hat{y} = \frac{1}{T}\sum_{i=1}^{T} y_i \tag{5}$$

where T is the number of trees.

Parameters:

- **Number of trees**: The number of decision trees in the forest.
- **Max features**: The maximum number of features to consider for splitting a node.
- **Min samples per leaf**: Minimum number of samples required at a leaf node.

Random Forests provide better accuracy than single decision trees by reducing variance through averaging. They are robust against overfitting and can handle large datasets with higher dimensionality.

5.4 Conclusion

These three algorithms represent fundamental approaches to machine learning classification. Decision Trees provide interpretability, Naïve Bayes offers computational efficiency, and Random Forests deliver robust performance through ensemble methods. The choice between them depends on the specific requirements of the problem at hand, including dataset characteristics, performance needs, and interpretability requirements.

6 Result Visualization

The performance of the implemented machine learning algorithms—**Decision Tree, Naïve Bayes, and Random Forest**—was evaluated on the dataset. Below is a comparison of their accuracy and other performance metrics.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	79.35%	85.19%	80.70%	82.88%
Naïve Bayes	79.35%	83.93%	82.46%	83.19%
Random Forest	84.78%	86.44%	89.47%	87.93%

Table 3: Performance Comparison of Machine Learning Models

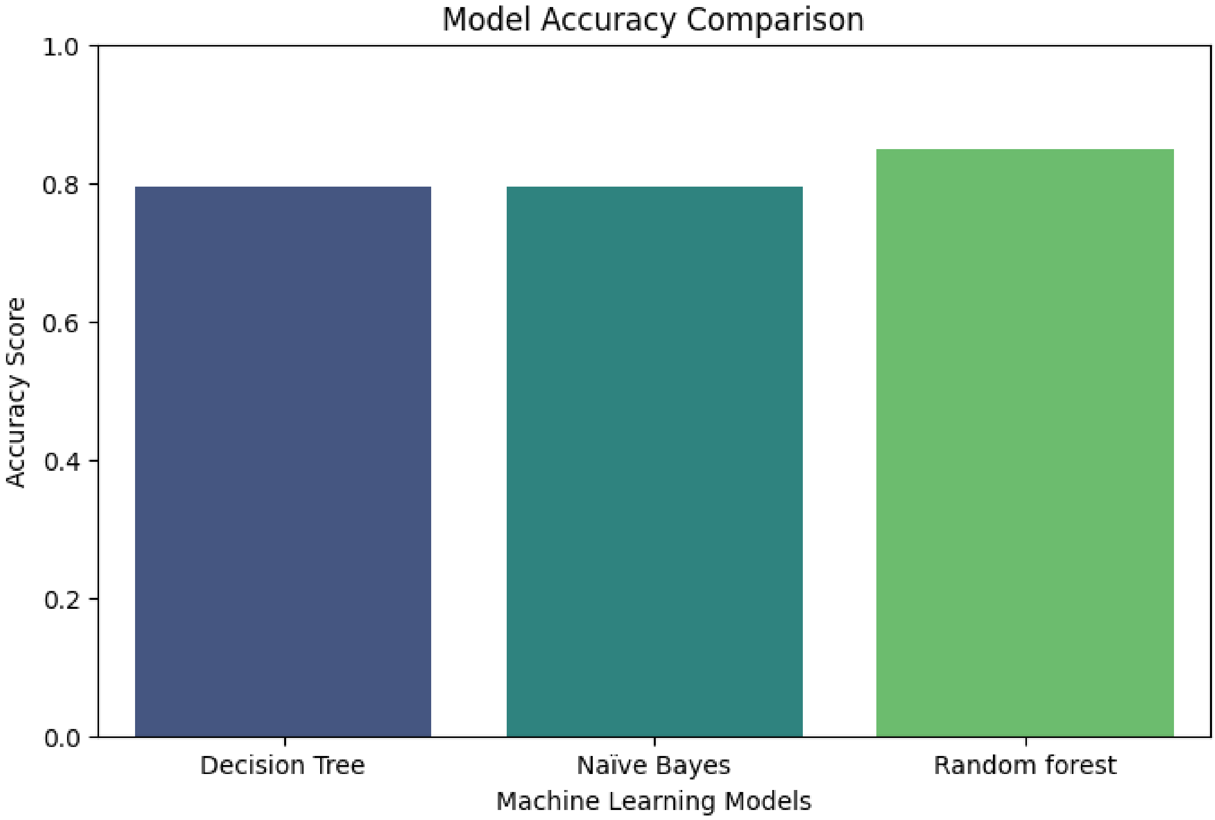


Figure 6: Model Accuracy Comparison

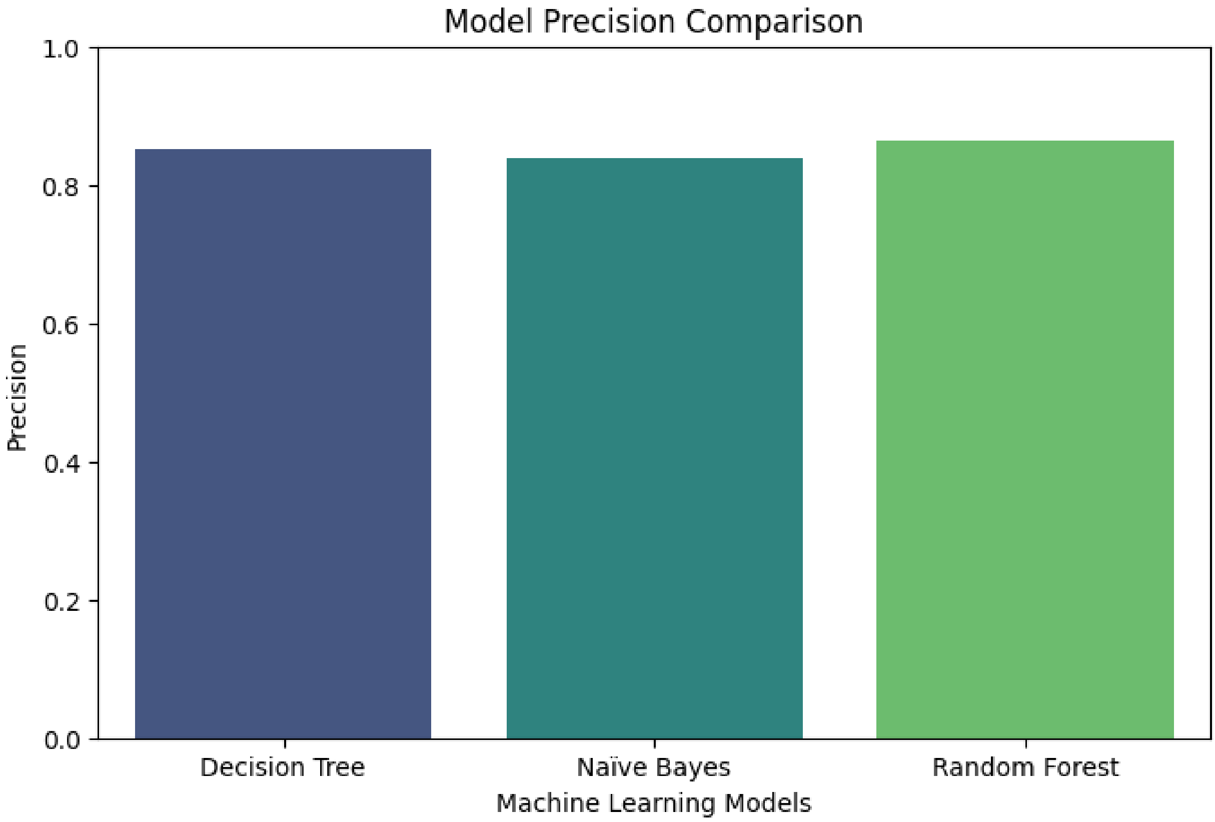


Figure 7: Model Precision Comparison

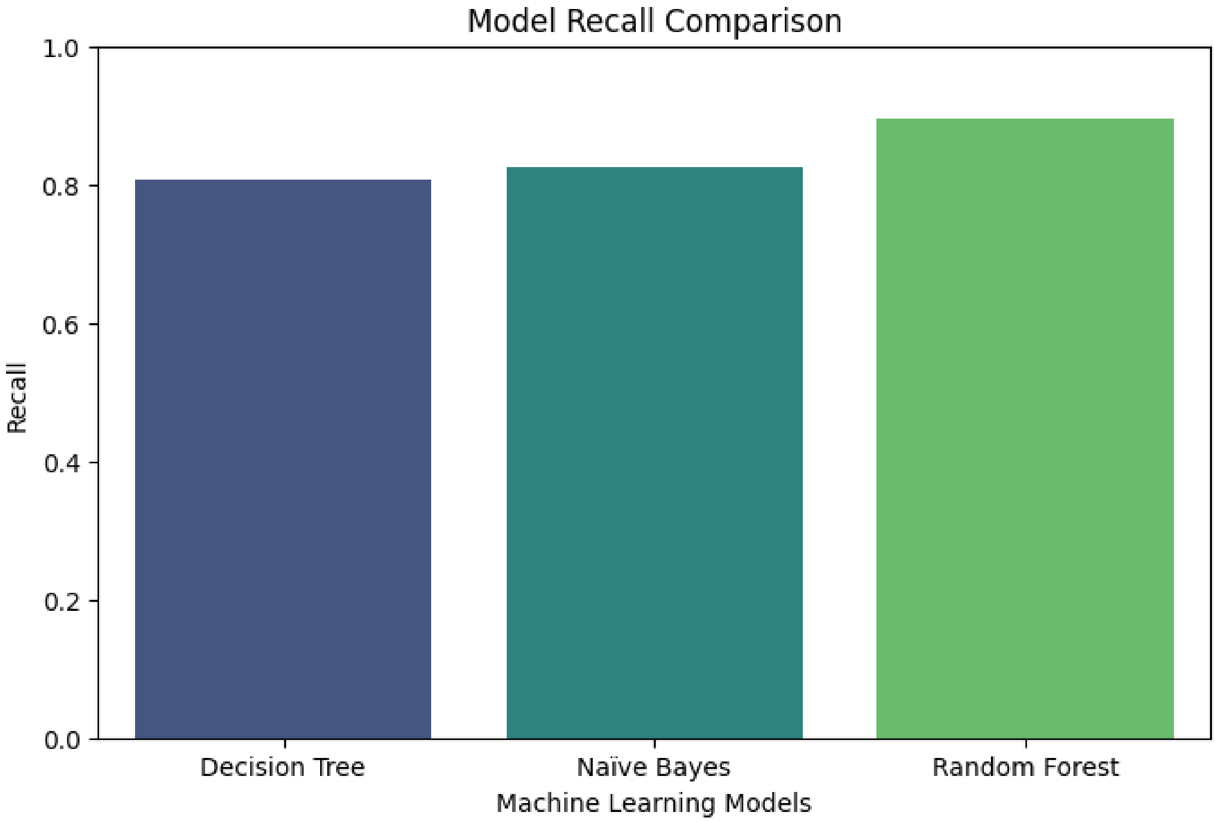


Figure 8: Model Recall Comparison

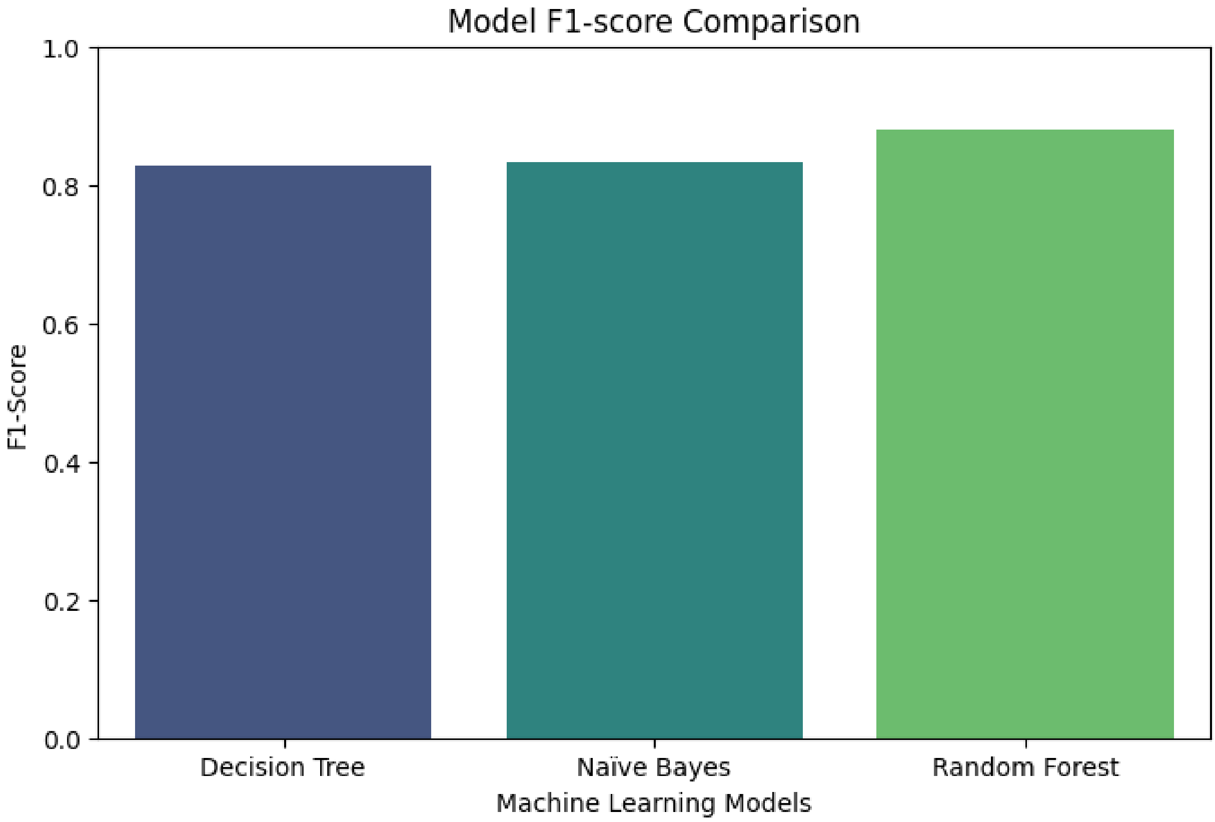


Figure 9: Model F1 score Comparison