Heart Disease Prediction Using Machine Learning Models: A Comparative Study

Mantasha Mohammed Sadiq
Computer Science and Engineering, Sardar Patel Institute of Technology, Mumbai, India
Email: mohammed.mantasha24@spit.ac.in

Heart Disease Prediction Using Machine Learning Models: A Comparative Study

Abstract

This study evaluates four machine learning models-Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest-for predicting heart disease using a clinical dataset of 1025 patient records. After preprocessing and feature engineering, models were trained and tested, with Logistic Regression achieving the highest accuracy of 82.46% and F1 score of 84.38%. The results highlight that simpler linear models can outperform more complex algorithms for this task, providing a reliable tool for early cardiovascular risk assessment. This work contributes to the development of clinical decision support systems leveraging machine learning for preventive cardiology.

***Keywords*:**
Heart Disease Prediction, Machine Learning, Logistic Regression, Clinical Decision Support, Accuracy, Feature Engineering, Data Preprocessing.

Introduction

Cardiovascular diseases (CVD) remain the leading cause of mortality worldwide, necessitating early and accurate prediction methods to improve patient outcomes. Traditional clinical risk assessments often fail to capture complex interactions among risk factors. Machine learning (ML) offers promising capabilities to analyze large datasets and uncover hidden patterns for disease prediction.

This project aims to compare the performance of four ML algorithms-Logistic Regression, SVM, KNN, and Random Forest-in predicting heart disease presence based on patient clinical data. The objective is to identify the most effective model to support early diagnosis and intervention.

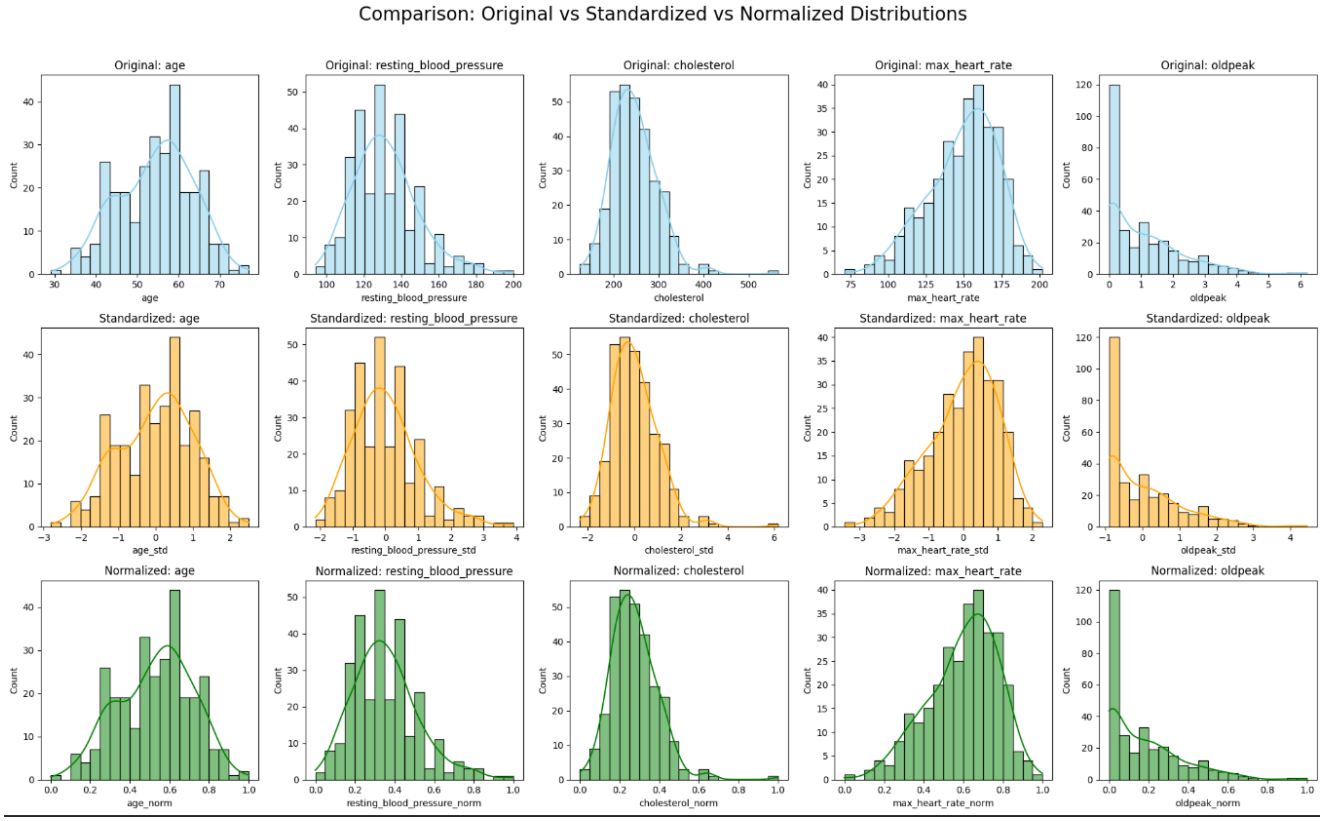
Methodology

Data Collection and Description

The dataset comprises 1025 patient records with 14 features including demographic data (age, sex), clinical measurements (blood pressure, cholesterol), and diagnostic test results (ECG, chest pain type). The target variable is binary, indicating presence or absence of heart disease.

Data Preprocessing

- Conversion of categorical variables (e.g., sex, chest pain type) into numerical formats.
- Handling missing values and ensuring correct data types.
- Standardization of numerical features to normalize scales.

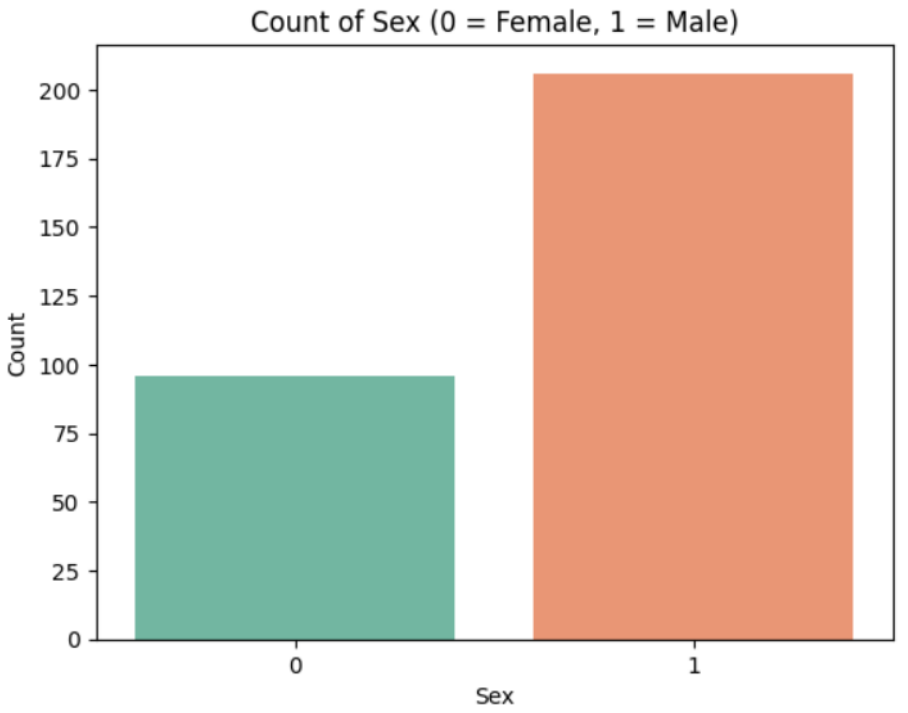


Exploratory Data Analysis (EDA)

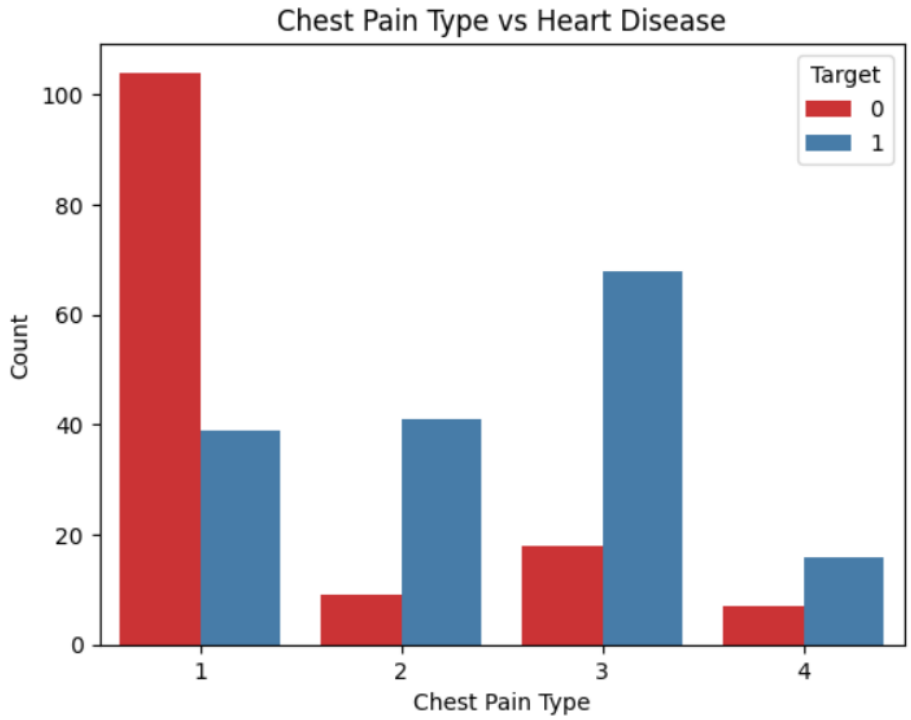
Key statistics:

- Mean Age: 54.4 years
- Mean Cholesterol: 246.5 mg/dL
- Mean Max Heart Rate: 149 bpm

Univariate Analysis revealed a fairly normal age distribution and a male-dominant dataset (68.2% male).

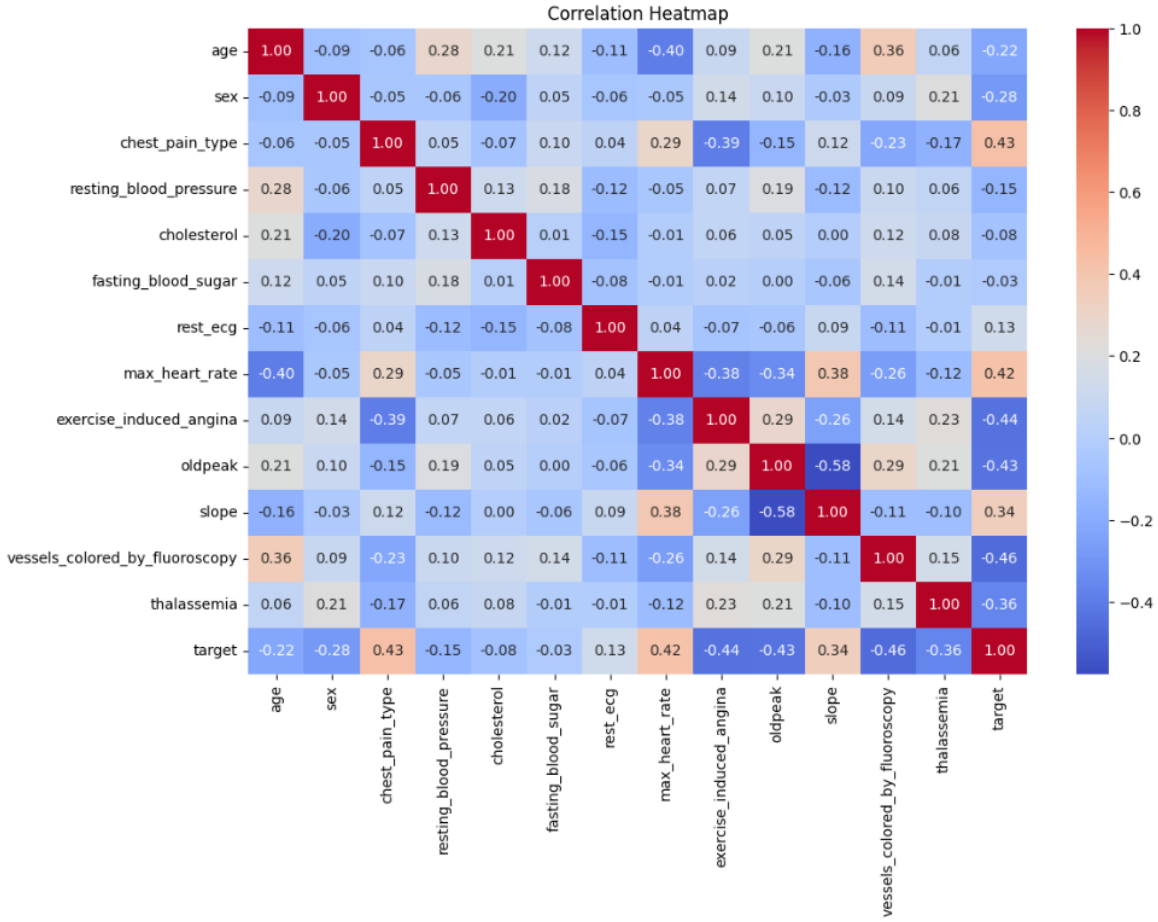


Bivariate Analysis showed chest pain type strongly differentiate between patients with and without disease.



Multivariate Correlations with target:

- Chest Pain Type: 0.43
- Max Heart Rate: 0.42
- ST Segment Slope: 0.34
- Age and Sex were negatively correlated (-0.22 and -0.28 respectively).



Feature Engineering

Key features used include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression (oldpeak), slope of ST segment, number of vessels colored by fluoroscopy, and thalassemia status.

Model Development

Four classification models were implemented:

1. **Logistic Regression (LR):** A linear model estimating the probability of heart disease presence.

2. **Support Vector Machine (SVM):** Finds the optimal hyperplane to separate classes, capable of handling non-linear relationships.
3. **K-Nearest Neighbors (KNN):** Classifies based on the majority class among nearest neighbors in feature space.
4. **Random Forest (RF):** Ensemble of decision trees voting for the most probable class.

Workflow Diagram

```
Raw Data → Data Cleaning → Feature Engineering → Train-Test Split → Model Training → Model Evaluation → Performance Comparison
```

Evaluation Metrics

Models were assessed using:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC AUC

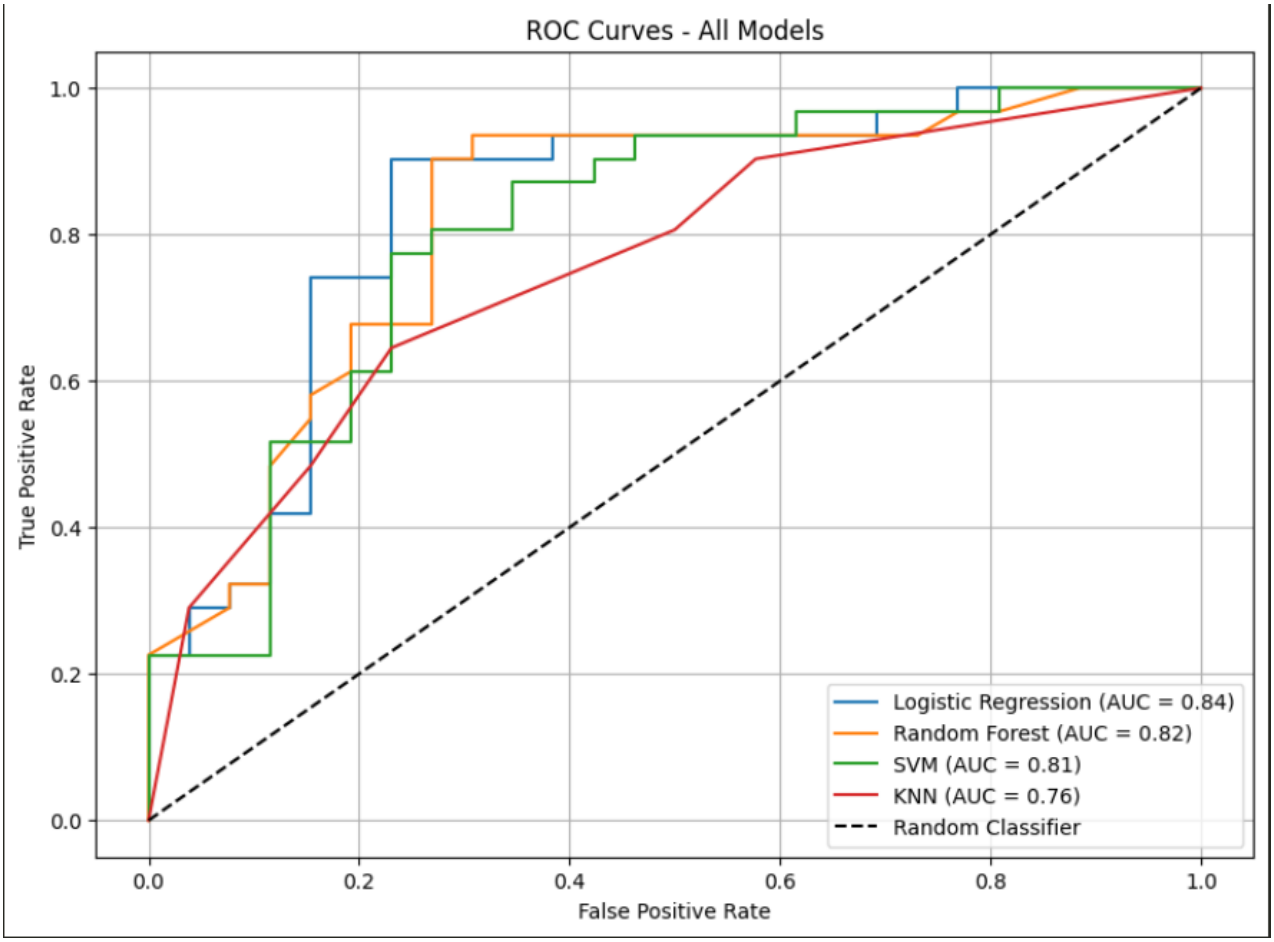
Results

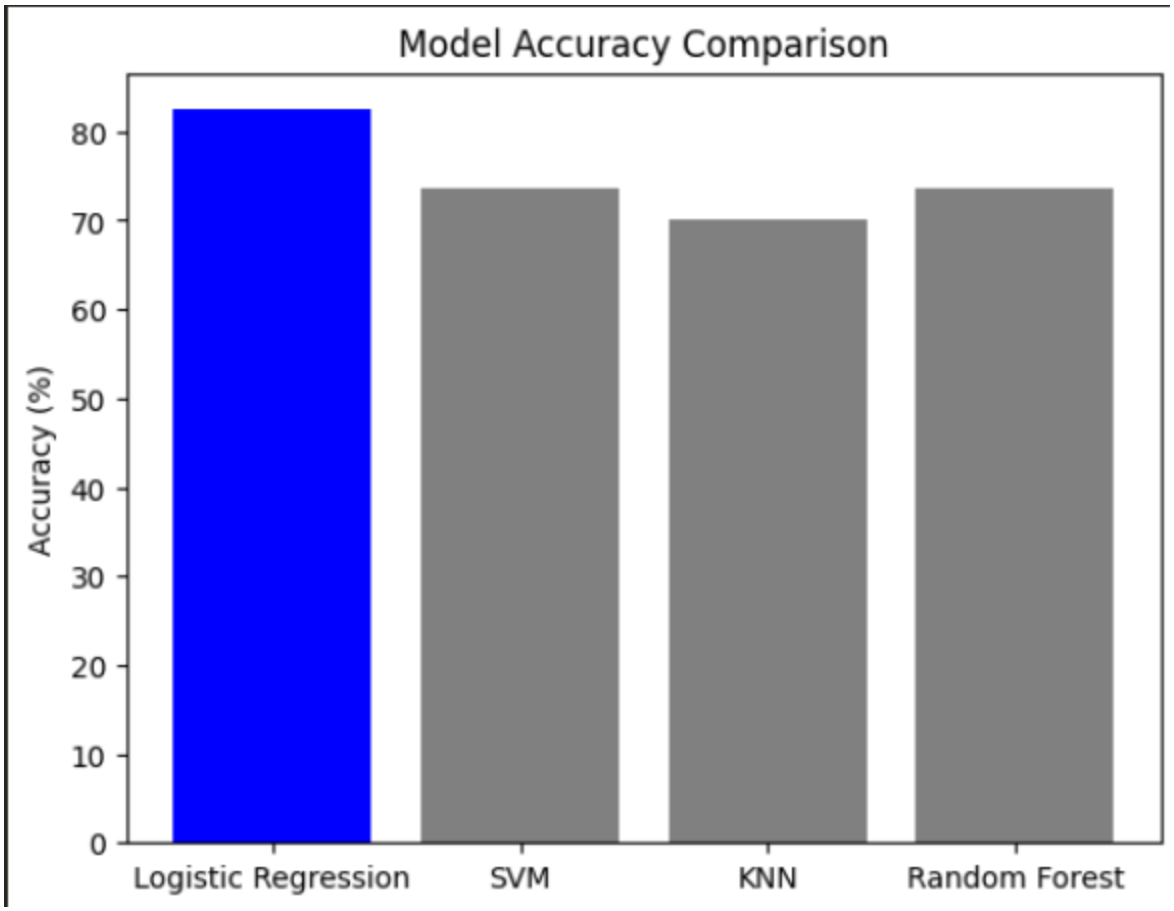
Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	82.46%	81.82%	87.10%	84.38%	83.75%
Random Forest	73.68%	76.67%	74.19%	75.41%	82.26%
Support Vector Machine	73.68%	78.57%	70.97%	74.58%	80.52%
K-Nearest Neighbors	70.18%	76.92%	64.52%	70.18%	75.74%

Analysis

- **Logistic Regression** achieved the highest overall performance, indicating that the relationship between features and heart disease is effectively captured by a linear model.

- **Random Forest** and **SVM** showed moderate accuracy but differed in precision and recall trade-offs.
- **KNN** had the lowest performance, suggesting proximity-based classification is less effective for this dataset.





Conclusion

This comparative study demonstrates that Logistic Regression outperforms more complex models like Random Forest and SVM in predicting heart disease from clinical data, achieving an accuracy of 82.46%. The findings emphasize that simpler, interpretable models can be highly effective for medical diagnosis tasks. Future work should explore feature importance analysis, larger datasets, and integration of temporal data for dynamic prediction. These results support the development of machine learning-based clinical decision support systems for early detection and prevention of cardiovascular diseases.

References

1. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access.

2. Malavika G. et al. (2023). Prediction of Heart Disease Based on Machine Learning Using Cleveland Dataset. PMC.
3. American Heart Association. (2022). Heart Disease and Stroke Statistics.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.
5. Kumar-laxmi. (2021). Heart Disease Prediction System using Machine Learning. GitHub Repository.