International Journal of Recent Advances in Multidisciplinary Topics
Volume 6, Issue 4, April 2025
https://www.ijramt.com | ISSN (Online): 2582-7839

A Comparative Study of Machine Learning Models for Predicting Chronic Diseases

Ishit Bajpai*
AI/ML Researcher, Chandigarh, India

***Abstract*: In this study, we explore the predictive capabilities of various machine learning models in identifying two major chronic conditions: diabetes and heart disease. Using publicly available datasets, we trained and evaluated Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models. Performance was assessed using Accuracy and ROC-AUC metrics. For heart disease, the best-performing model was Random Forest with an accuracy of 0.88. For diabetes, Logistic Regression performed best with an accuracy of 0.75. Our findings reinforce the value of ML in preventive healthcare and suggest promising directions for future improvements.**

***Keywords*: Machine Learning, Health Care, Diabetes, Heart diseases.**

1. Introduction

Chronic diseases such as diabetes and heart disease are leading causes of death worldwide. Early prediction and intervention can significantly reduce their impact. Traditional diagnostic approaches often rely on clinical expertise and extensive testing, which may be time-consuming or resource-intensive.

Machine learning (ML) offers a data-driven alternative that can support early detection through automated analysis of patient data. This study aims to evaluate and compare multiple ML models for predicting the presence of diabetes and heart disease using historical health records. The ultimate goal is to identify reliable models that can assist in preventive healthcare settings.

2. Dataset Description

- *Diabetes Dataset*
  - *Source*: UCI Machine Learning Repository
  - *Samples*: 768
  - *Features*: 8 numerical features + 1 target variable
  - *Target*: 0 (No Diabetes), 1 (Diabetic)
  - *Preprocessing*: Feature scaling applied using Standard Scaler.
- *Heart Disease Dataset*
  - *Source*: UCI Heart Disease Dataset
  - *Samples*: 303
  - *Features*: 13 numerical/categorical features + 1 target variable
  - *Target*: 0 (No Disease), 1 (Disease)
  - *Preprocessing*: Categorical features were encoded using one-hot encoding, scaling for numerical features.
- *Class Distribution*

Both datasets exhibit moderate class imbalance, which was considered during model evaluation through metrics like ROC-AUC.

3. Methodology

- *Train/Test Split*: 80/20 ratio
- Models Used
  - Logistic Regression (LR)
  - Random Forest (RF)
  - K-Nearest Neighbors (KNN)
  - Support Vector Machine (SVM)
- Evaluation Metrics
  - Accuracy
  - ROC-AUC Score
  - Confusion Matrix
  - Classification Report
- Preprocessing
  - Feature Scaling (Standard Scaler)
  - One-hot encoding (for categorical heart features)

4. Results & Discussion

Table 1
Diabetes dataset performance

Model	Accuracy	ROC-AUC
Logistic Regression	0.75	0.81
Random Forest	0.72	0.81
KNN	0.68	0.76
SVM	0.72	0.80

- *Best Model*: Logistic Regression
- *Insights:*
  - Linear models worked better here possibly due to simpler feature-target relationships.
  - Confusion matrices showed LR had the lowest false positives.

*Corresponding author: ib7147@gmail.com

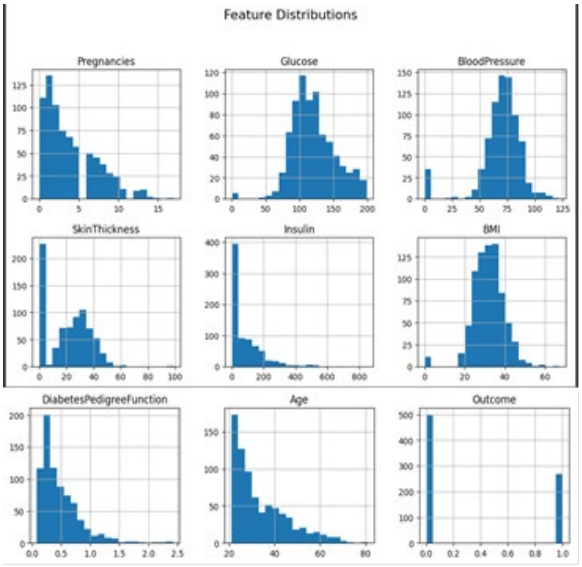

Fig. 1.

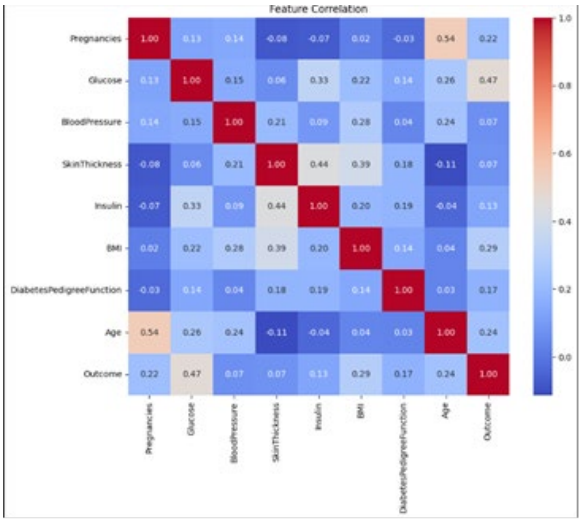

Fig. 2.

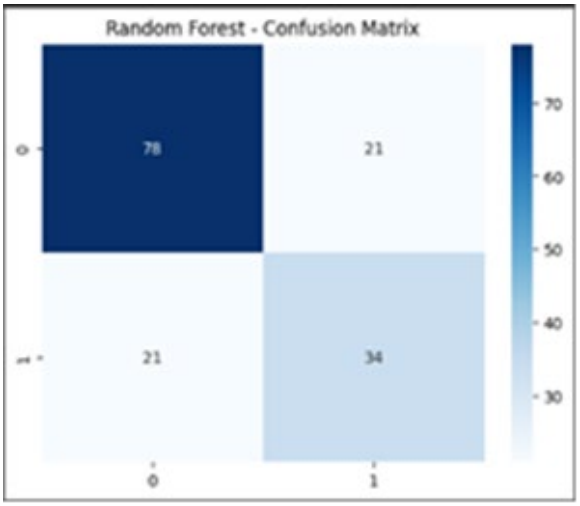

Fig. 3.

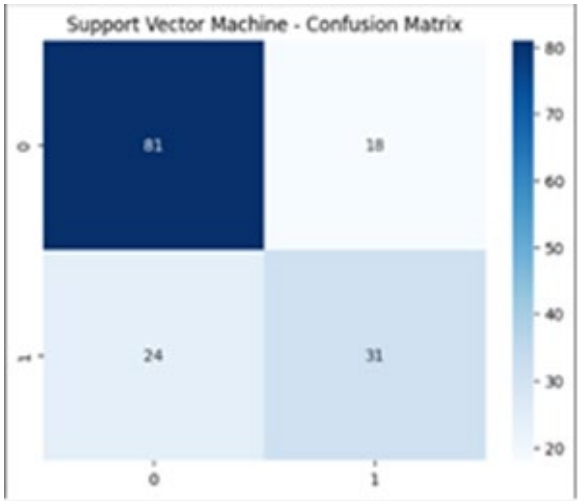

Fig. 4.

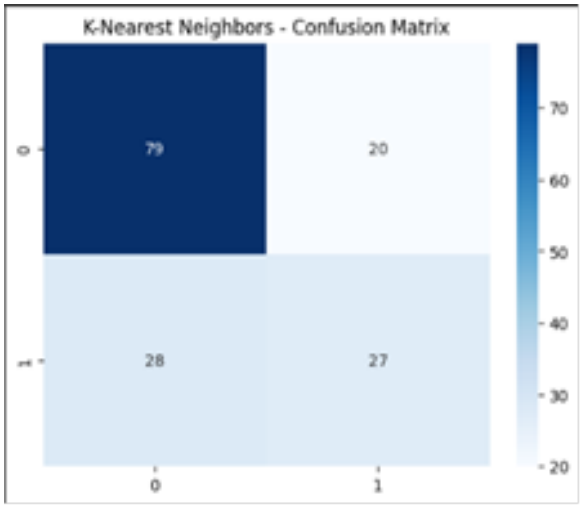

Fig. 5.

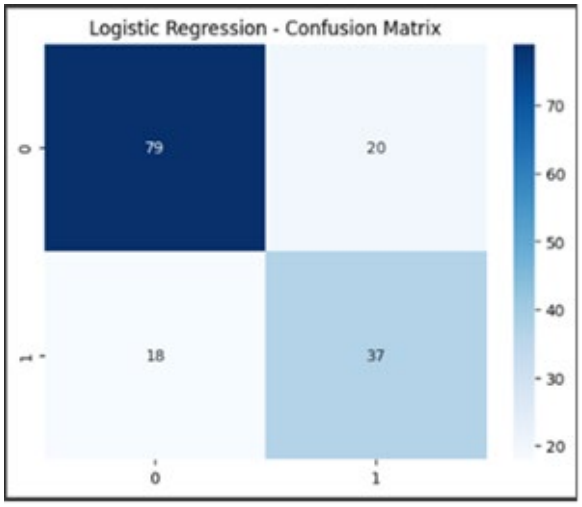

Fig. 6.

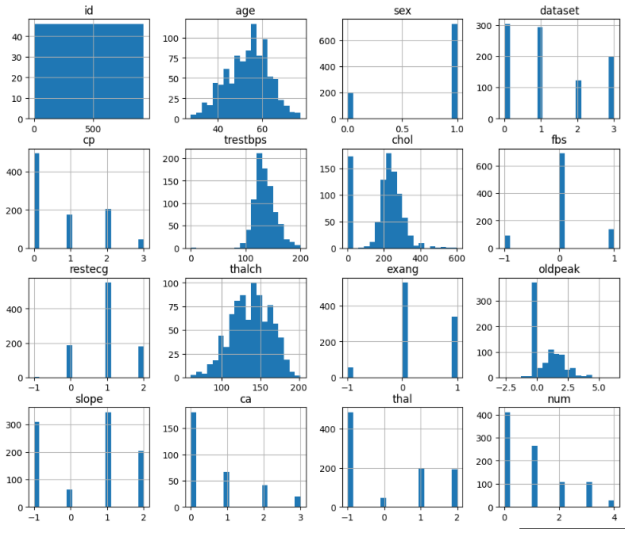

Fig. 7.

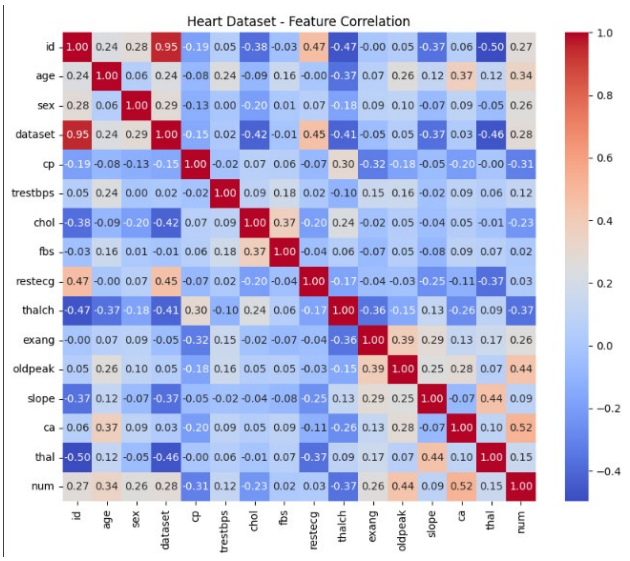

Fig. 8.

Table 2
Heart disease dataset performance

Model	Accuracy	ROC-AUC
Logistic Regression	0.78	0.89
Random Forest	0.88	0.94
KNN	0.82	0.89
SVM	0.86	0.91

- *Best Model*: Random Forest
- *Insights:*
  - RF handles feature interactions well and is robust to noise.
  - KNN suffered slightly due to high dimensionality and non-linearity.

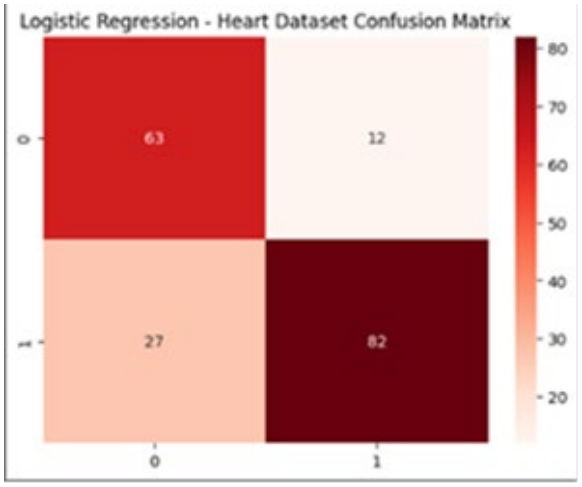

Fig. 9.

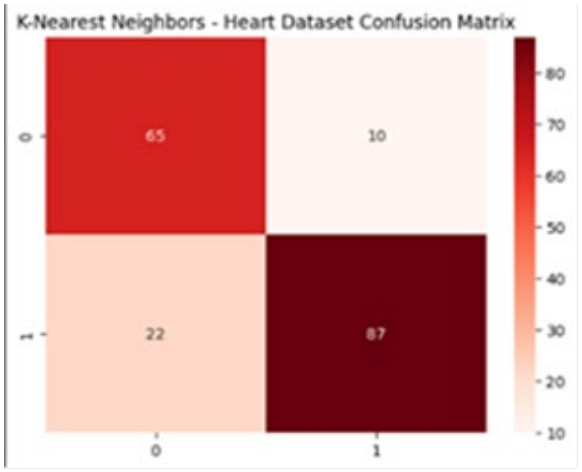

Fig. 10.

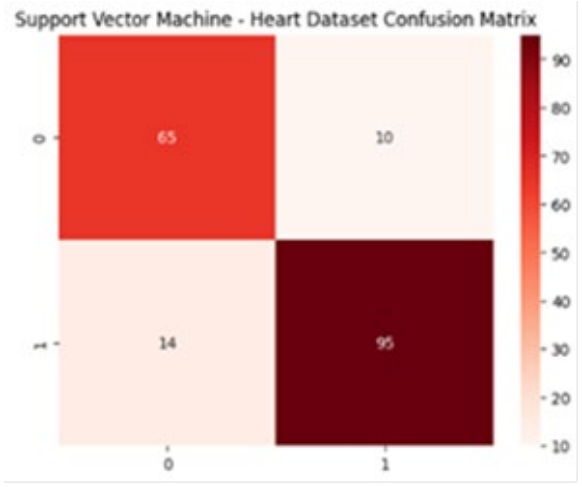

Fig. 11.

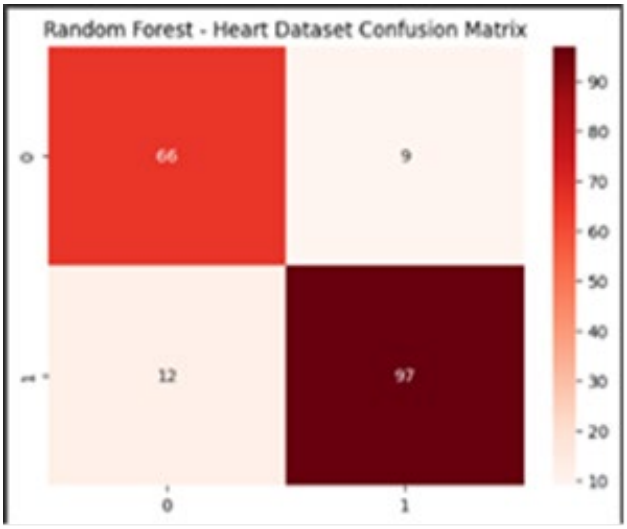

Fig. 12.

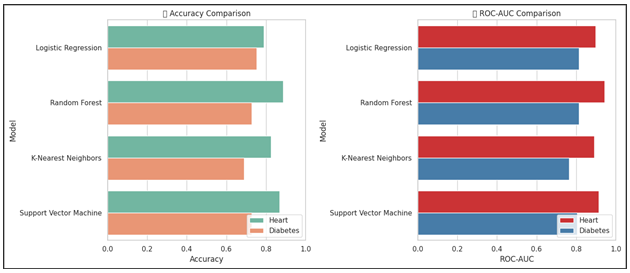

Fig. 13.

Comparative Insights:

- Heart dataset was easier to model, possibly due to stronger feature correlations and clearer patterns.
- Diabetes dataset was more challenging, possibly due to overlapping feature values and higher noise.

5. Conclusion

This study demonstrates that ML models can effectively predict chronic diseases from patient data. Logistic Regression was the top performer for the diabetes dataset, while Random Forest led in heart disease predictions.

These results highlight the potential for deploying ML in real-world healthcare applications, where accurate early prediction can lead to timely interventions and better outcomes.

Future Work:

- Hyperparameter optimization (e.g., GridSearchCV)
- Deep learning methods (e.g., MLP, CNN)
- Explainability (e.g., SHAP, LIME)
- Deploying as a clinical decision support tool

References

[1] Pima Indians Diabetes Database, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[2] UCI Heart Disease Data, https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data?select=heart_disease_uci.csv

[3] Scikit-learn: Machine Learning in Python

[4] Pandas, Seaborn, Matplotlib – Data analysis and visualization libraries.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[6] McKinney, W. (2010). Data structures for statistical computing in Python. *In 9th Python in Science Conference* (pp. 51-56).

[7] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.