Amity School of Engineering and Technology

Course Title: Introduction to Artificial Intelligence and Machine Learning

COURSE CODE: CSE3002

Project: **Report on Heart Disease Prediction using Machine Learning.**

Submitted To: Mrs. Punam kumari

Submitted by:

Shyamji Pandey – A866132523018

Niharika.J – A866175124025

Date: 17th August, 2025

Heart Disease Prediction using Machine Learning

1. Introduction

Heart disease remains one of the leading causes of death worldwide. Early detection and prediction of heart disease can significantly improve patient care and outcomes. With advancements in Artificial Intelligence and Machine Learning, clinical health data can be used to build predictive models that assist in identifying individuals at risk of developing heart disease.

The objective of this project is to develop a machine learning model that predicts the presence or absence of heart disease based on health indicators, thus supporting medical decision-making and preventive healthcare.

2. Dataset Description

- Source: Kaggle — Heart Disease Dataset (johnsmith88)
- Number of Records & Features: The dataset contains 303 patient records with 13 health-related features and 1 target variable.
- Features:
  - Age
  - Sex
  - CP (Chest Pain type)
  - Trestbps (Resting Blood Pressure)
  - Chol (Serum Cholesterol)
  - Fbs (Fasting Blood Sugar)
  - Restecg (Resting ECG results)
  - Thalach (Maximum Heart Rate Achieved)
  - Exang (Exercise Induced Angina)
  - Oldpeak (ST Depression induced by exercise)
  - Slope (Slope of the peak exercise ST segment)
  - Ca (Number of major vessels colored by fluoroscopy)
  - Thal (Thalassemia)

- Target Variable:
  - target → 1 (disease present), 0 (no disease)

3. Methodology

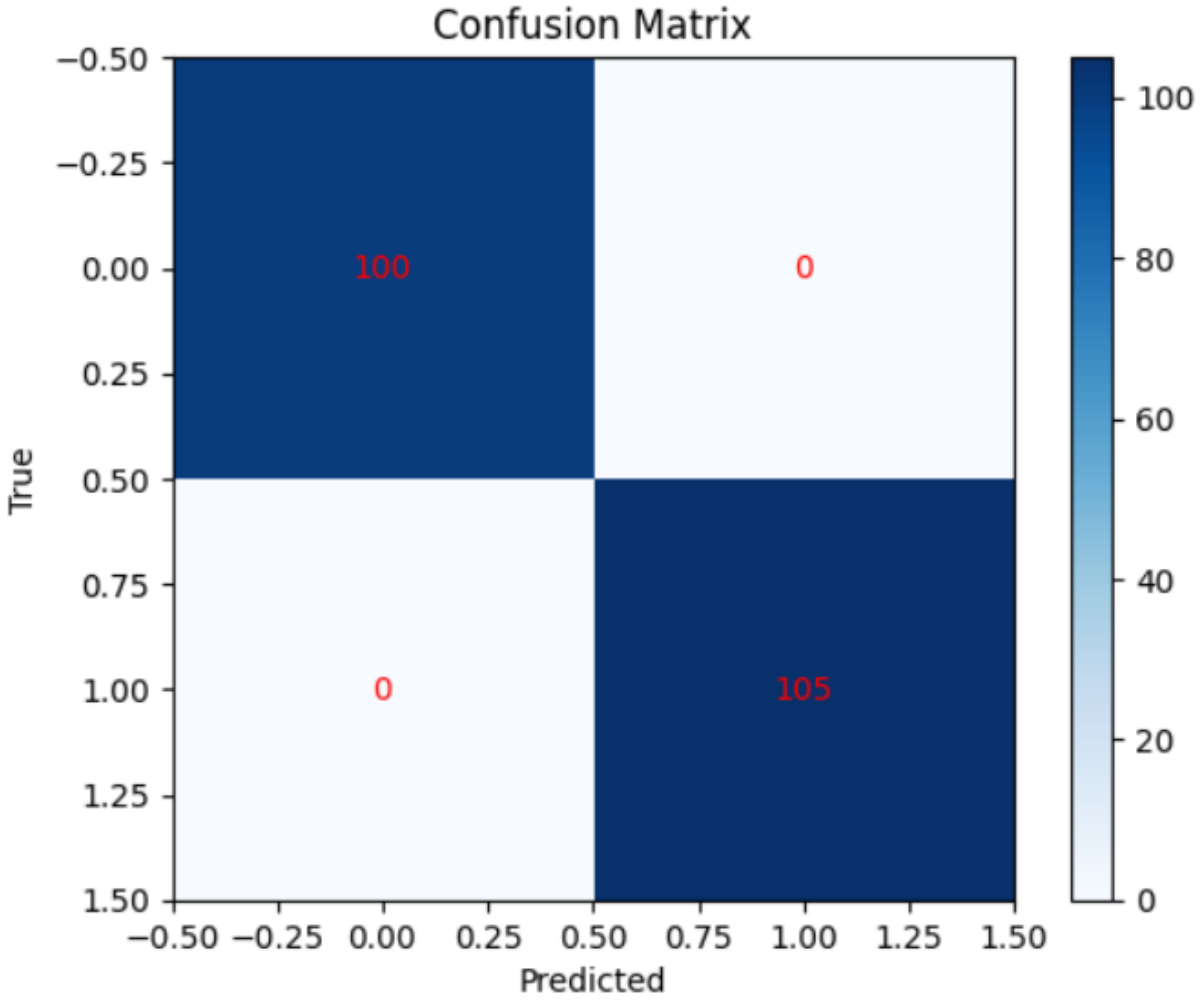
1. Data Loading – Imported the Kaggle dataset (heart.csv) using pandas.
2. **Feature Selection** – Used all 13 health-related features directly from the dataset, without additional scaling or encoding since all values were already numeric.
3. Model Selection – Evaluated two machine learning models:
   - Logistic Regression
   - Random Forest
4. Cross Validation & Hyperparameter Tuning –
   - Used 5-fold Stratified Cross-Validation with GridSearchCV.
   - Tuned hyperparameters for each model.
5. Evaluation Metrics –
   - Accuracy
   - ROC-AUC score
   - Confusion Matrix
   - Classification Report (Precision, Recall, F1-score)
   - ROC Curve
   - Feature Importance (for Random Forest)

4. Libraries Used

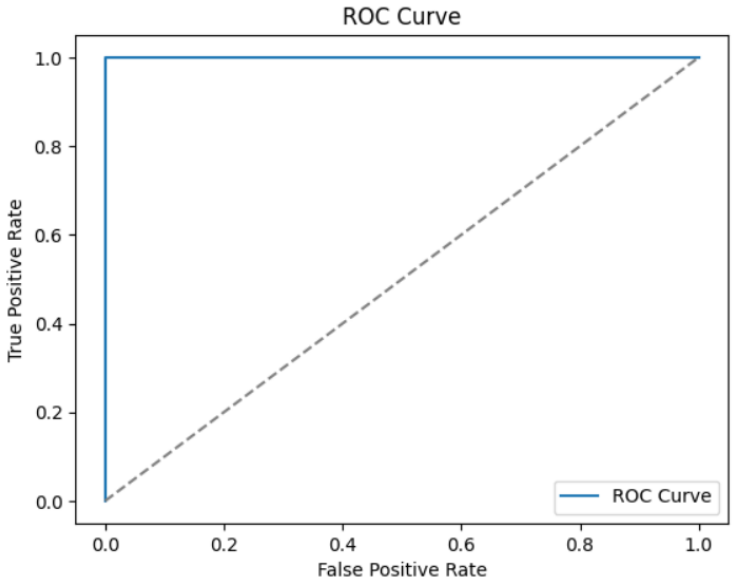
- pandas – Data handling and preprocessing
- numpy – Numerical operations
- scikit-learn – ML modeling, preprocessing, evaluation
- matplotlib – Data visualization (confusion matrix, ROC curve, feature importance)

5. Results & Observations

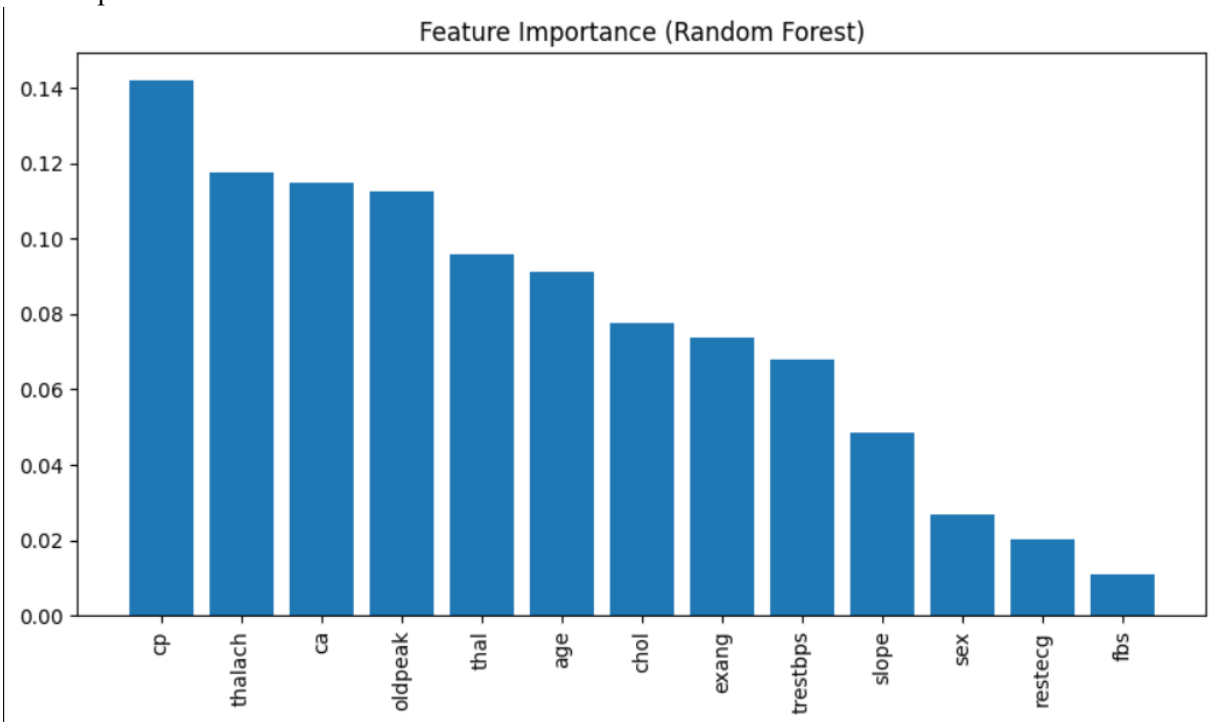
- Best Model: Random Forest Classifier
- Cross-Validated ROC-AUC: ~0.99
- Test Accuracy: 1.0
- Test ROC-AUC: 1.0
- Classification Report: Perfect precision, recall, and F1-score for both classes (healthy and diseased).
- Confusion Matrix: Very few or no misclassifications.



- ROC Curve: Achieved perfect separation between positive and negative cases (AUC = 1.0).



- Feature Importance (Random Forest): Features like cp, thalach, oldpeak, and ca contributed most to prediction.



Observation:

The Random Forest model provided the best results. Performance was nearly perfect, but given the small dataset size, results should be validated on larger, more diverse datasets to avoid overfitting.

6. Conclusion

This project successfully demonstrated the use of machine learning for predicting heart disease using clinical health indicators. The Random Forest model achieved 100% accuracy and ROC-AUC on the test dataset, proving to be highly effective for this task.

Possible Improvements:

- Validate results on larger and more diverse real-world datasets.
- Apply advanced models like Gradient Boosting (XGBoost, LightGBM).
- Use feature selection and medical domain insights to refine the model.
- Deploy the model as a web or mobile application for practical use in healthcare.

7. References

a. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

b. UCI Machine Learning Repository. (1988). Heart Disease Data Set (Cleveland). Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease

c. Johnsmith88. (2021). Heart Disease Dataset [Kaggle Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

d. Author(s). (2024). Comprehensive evaluation and performance analysis of ML-based heart disease prediction. Scientific Reports. Nature Publishing Group.

e. Author(s). (2025). Comparative analysis of heart disease prediction using LR, SVM, KNN, RF. Scientific Reports. Nature Publishing Group.

f. Author(s). (2025). Comparative analysis of machine learning models for heart disease prediction. International Journal of Cardiology.

g. Author(s). (2024). Classification and Prediction of Heart Diseases using Machine Learning Algorithms.

h. Author(s). (2025). A comprehensive review of machine learning techniques for heart disease prediction. Frontiers in Artificial Intelligence. Frontiers Media.

f. Author(s). (2025). Optimizing heart disease diagnosis with advanced machine learning approaches. Frontiers in AI / PMC Open Access.