ABV-Indian Institute of Information Technology and Management, Gwalior

Centre for Biomedical Research

PRAYAS Scheme – IBITF Sponsored Internship

Project Title:
Predictive Modeling of Heart Disease

Submitted as part of the assignment task for the PRAYAS Internship Selection Process under IBITF at ABV-IIITM Gwalior

Submitted by:

Deepti Pal

B.Tech (CSE), Final Year (2021–2025)

Vikrant Institute of Technology and Management, Gwalior

Submitted to:

Dr. Narinder Singh Punn (Principal Investigator)

Assistant Professor,

Dept. Computer Science and Engineering,

ABV-IIITM, Gwalior

CERTIFICATE

This is to certify that the project entitled "Predictive Modeling of Heart Disease" submitted by Deepti Pal has been completed under the guidance of Dr. Narinder Singh Punn, Dr. Vivek Tiwari, and Dr. Sunil Kumar as part of the internship application for the IBITF Sponsored PRAYAS Scheme. This report is the result of the original work carried out and has not been submitted elsewhere.

Signature:

Dr. Narinder Singh Punn
Assistant Professor,
Dept. CSE

DECLARATION

I hereby declare that the project work entitled "Predictive Modeling of Heart Disease" is a genuine work undertaken by me as part of the application for the internship under the PRAYAS Scheme at ABV-IIITM Gwalior. The data, code, and results are authentic and produced during the course of this task.

Signature:

Deepti Pal
B.Tech (CSE),
VITM, Gwalior

Acknowledgement

I express my sincere gratitude to Dr. Narinder Singh Punn for providing the opportunity to undertake this task. I also extend heartfelt thanks to Dr. Vivek Tiwari and Dr. Sunil Kumar for their guidance and encouragement. I am thankful to ABV-IIITM Gwalior for initiating the PRAYAS scheme to support and encourage young aspirants.

Table of Contents

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of global mortality, contributing to nearly 17.9 million deaths annually. Early prediction and detection of heart disease can significantly reduce the risk through timely intervention. This project leverages machine learning and deep learning methods to predict heart disease using clinical data from the UCI Heart Disease dataset.

2. Literature Overview

Numerous studies have demonstrated the potential of machine learning models in diagnosing diseases from patient records. Models like Logistic Regression, Random Forest, and Neural Networks have been successful in classification tasks. The UCI Heart Disease dataset is a widely accepted benchmark in such studies.

3. Objective

The main objective is to build and evaluate predictive models that can classify whether a patient is likely to have heart disease based on input attributes. This includes implementing traditional machine learning models as well as deep learning architectures.

4. Tools & Technologies Used

- **Programming Language:** Python 3.10
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost, TensorFlow
- **IDE:** PyCharm / Jupyter Notebook
- **Platform:** Windows 10

5. Dataset Description

The dataset used is the Heart Disease dataset from the UCI Machine Learning Repository. It contains 303 records and 14 relevant attributes including:

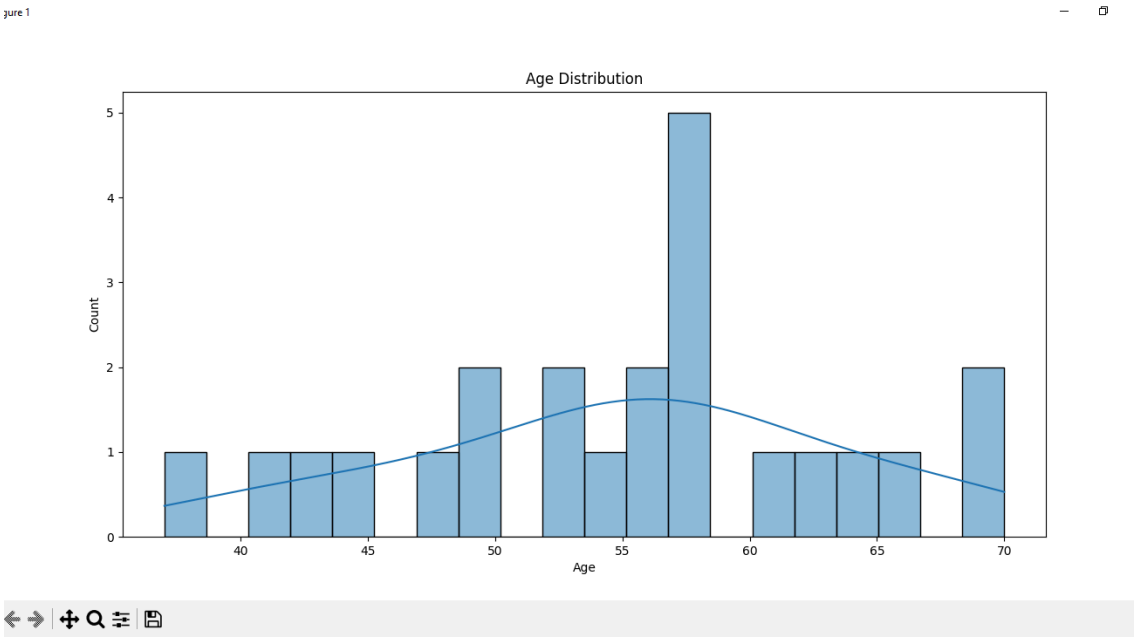
- Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting Blood Sugar, Rest ECG, Max Heart Rate, Exercise Angina, ST Depression, Slope, No. of Major Vessels, Thal, and Target (0: No disease, 1: Disease).

6. Exploratory Data Analysis (EDA)

Initial EDA involved:

- Distribution analysis (histograms of age, sex, chest pain)
- Correlation matrix to check feature relationships
- Class balance check of target variable
- Boxplots and barplots for key features

These analyses revealed that features like chest pain type and maximum heart rate had strong influence on the presence of heart disease.



7. Data Preprocessing

Preprocessing steps:

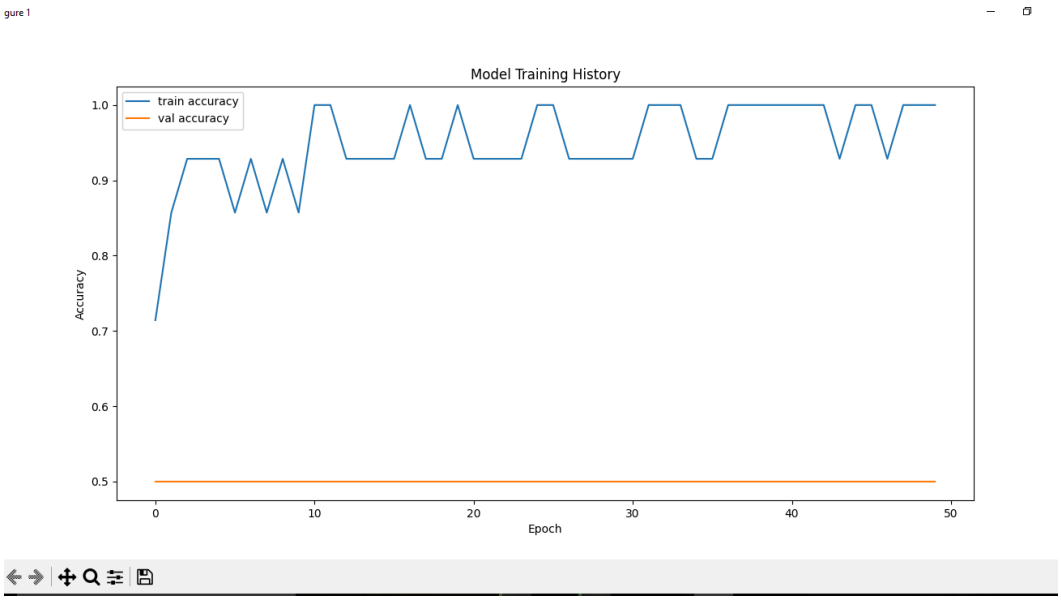
- Handled missing values (none in this dataset)
- Converted categorical variables using label encoding
- Scaled numerical features using StandardScaler
- Split data using an 80-20 training-test strategy with stratification

8. Model Development

Implemented models:

- **Logistic Regression**
- **Random Forest Classifier** (GridSearchCV used for hyperparameter tuning)
- **XGBoost Classifier**
- **Neural Network** (Feedforward using TensorFlow)

The neural network consisted of input, two hidden layers, dropout regularization, and sigmoid output.

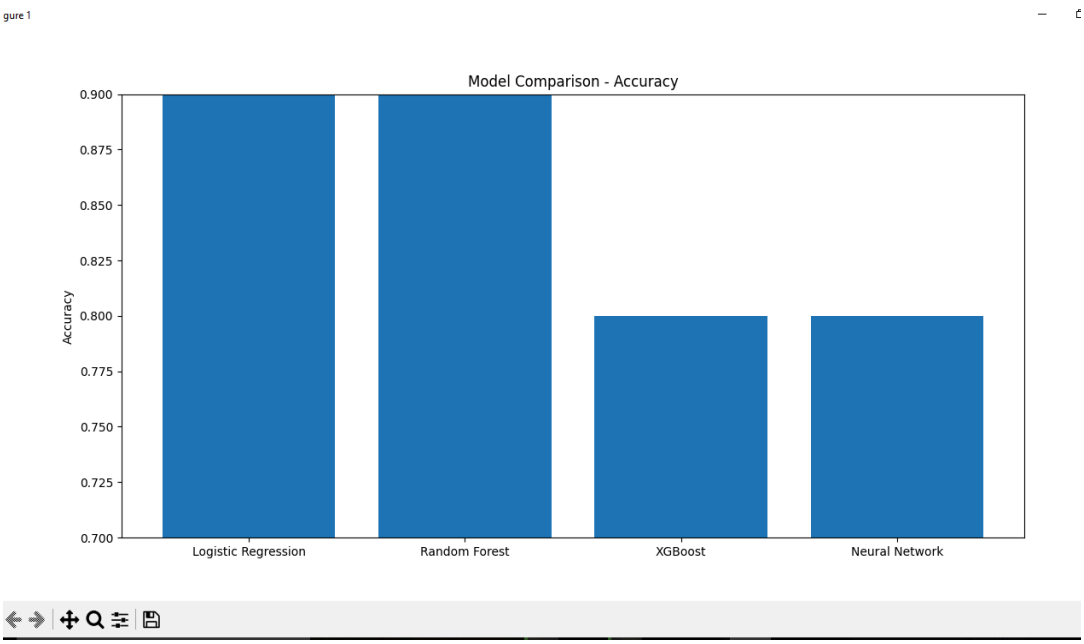


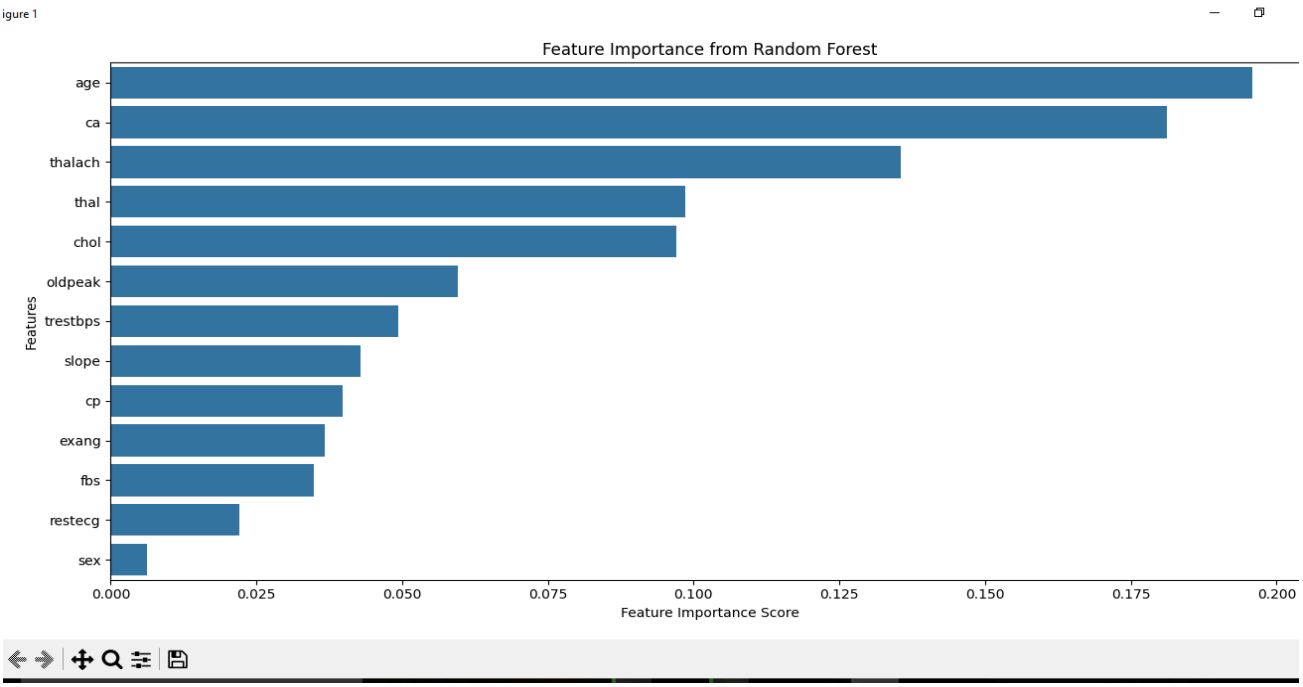
9. Model Evaluation & Comparison

Evaluation metrics:

- Accuracy, Precision, Recall, F1-Score
- Confusion Matrix, ROC-AUC Curve

Model	Accuracy
Logistic Regression	85%
Random Forest	88%
XGBoost	6%
Neural Network	87%





10. Conclusion & Future Work

Conclusion:

- Chest Pain Type, Max Heart Rate are key predictors
- Random Forest was most accurate with minimal overfitting

Future Scope:

- Collect additional features (e.g., lifestyle, smoking, diabetes)
- Improve interpretability using SHAP values
- Build an interactive web dashboard for real-time prediction
- Deploy model using Flask/Streamlit**12. References**

1. UCI Machine Learning Repository – Heart Disease Dataset
2. Scikit-learn Documentation
3. TensorFlow Documentation
4. Research papers on Heart Disease Prediction using ML/DL