Predicting Heart Disease Using Machine Learning

TEAM MEMBERS:

Yaswitha Kalapala, Sujish Kumar Indrajith, Rebecca Dara, Lakshmi Wajja

PROJECT OVERVIEW:

Heart disease is a leading cause of mortality worldwide, and early detection can significantly improve patient outcomes. This project aims to develop a predictive model for heart disease using machine learning techniques. By leveraging the Cleveland Heart Disease dataset, we will preprocess data, apply multiple classification models, and determine the most effective one. The goal is to build a reliable model to aid healthcare professionals in early diagnosis.

DOMAIN AND ADDRESSED ISSUES:

This project falls within the **healthcare and data science domain**. It addresses key issues such as:

- The need for accurate and timely heart disease prediction.
- The enhancement of traditional diagnostic methods with machine learning.
- Feature selection and model interpretability to assist healthcare professionals in decision-making.

RELATED WORK:

Several studies have explored ML-based heart disease prediction. The Framingham Heart Study identifies cardiovascular risk factors, while deep learning approaches like Convolutional Neural Networks (CNNs) have been applied for ECG analysis. Traditional methods such as Logistic Regression, Support Vector Machines (SVMs), and ensemble models have been widely used. This project distinguishes itself by focusing on rigorous preprocessing, feature selection, and hyperparameter tuning to optimize model accuracy and interpretability.

DATA PLAN:

- **Dataset**: Cleveland Heart Disease dataset (UCI ML Repository)
- **Size:** 303 patient records, 13 features + 1 target variables
- **Target Variable:** Binary (1 = Presence of heart disease, 0 = Absence)
- **Reason for Selection**: Publicly available, well-documented, and widely used in medical research.

PREPROCESSING STEPS:

- Handle missing data (none detected).
- Detect & treat outliers using Z-score and IQR methods.
- Normalize numerical variables using Min-Max scaling.
- Select key features using Recursive Feature Elimination (RFE).
- Encode categorical variables via one-hot and binary encoding.

IMPLEMENTATION PLAN:

The data mining pipeline involves multiple stages to ensure accurate and efficient heart disease prediction. It begins with Data Collection & Cleaning, where raw data is gathered and processed to handle inconsistencies. Next, Exploratory Data Analysis (EDA) & Feature Engineering is conducted to uncover insights and optimize feature selection. The Model Selection & Training phase involves experimenting with various machine learning models, including Logistic Regression, KNN, SVM, Decision Tree, Random Forest, and Naive Bayes. To enhance performance, Hyperparameter

Tuning is applied using Grid and Random Search methods. The Model Evaluation & Validation step ensures the chosen model meets accuracy and reliability standards through metrics such as precision, recall, and ROC-AUC. Finally, the process concludes with Interpretation & Potential Deployment, where results are analyzed for real-world applicability and possible integration into healthcare systems.

TOOLS & LIBRARIES:

Python (pandas, numpy, scikit-learn, matplotlib, seaborn)

Jupyter Notebook, Google Colab/AWS

EVALUATION PLAN:

- **Primary Metric:** Accuracy
- **Secondary Metrics:** Precision, Recall, F1-score, ROC-AUC
- **Comparison:** Evaluate against baseline models (e.g., simple Logistic Regression).

GROUP COLLABORATION PLAN:

To collaborate efficiently and ensure seamless data pipeline management, we have divided tasks among team members based on their strengths, experience, and background.

- **Meetings:** Weekly (in-person & Zoom).
- **Version Control:** GitHub for code and data management.
- **Task Assignment:** Google Docs & Trello.

TIMELINE:

Week	Task
1-2	Literature review, dataset collection, preprocessing
3-4	EDA and feature selection
5-6	Model training, evaluation, and comparison
7	Hyperparameter tuning, performance optimization
8	Model validation, real-world testing
9-10	Documentation, final report, presentation

REFERENCES:

- UCI Machine Learning Repository - Cleveland Heart Disease Dataset: https://archive.ics.uci.edu/ml/datasets/heart+disease
- Framingham Heart Study: https://www.framinghamheartstudy.org/
- Research on ML-based Cardiology Diagnostics