RWANDA POLYTECHNIC

HUYE COLLEGE

Module code & title: ITLML801- Machine Learning

Department: ICT

Program: IT

Level: Level 8, Year 4, Btech

RegNº: 25RP20037

Name: Nelson TUYISHIMIRE

Heart Disease Risk Prediction System

1. Introduction

Heart disease is a leading cause of mortality worldwide. Timely detection and risk stratification are essential for effective clinical management. Traditional assessment methods require manual evaluation by clinicians, which can be time-consuming and prone to human error. This project automates risk prediction, enabling faster and more consistent decision-making.

Objectives:

- Develop a machine learning model to classify heart disease risk.
- Deploy the model via a Flask API.
- Provide an intuitive and responsive frontend for hospital staff.

2. Dataset Description

- **Total Records:** 5,000 patients.
- **Features (13):**
  - ✓ Age
  - ✓ Sex

- ✓ Chest pain type (cp)
- ✓ Resting blood pressure (trestbps)
- ✓ Serum cholesterol (chol)
- ✓ Fasting blood sugar (fbs)
- ✓ Resting ECG results (restecg)
- ✓ Max heart rate (thalach)
- ✓ Exercise-induced angina (exang)
- ✓ ST depression (oldpeak)
- ✓ Slope of ST segment (slope)
- ✓ Number of major vessels (ca)
- ✓ Thalassemia status (thal)

- **Target:** 5 classes (No Disease, Very Mild, Mild, Severe, Immediate Danger)

3. Exploratory Data Analysis (EDA)

- **Statistics:** Mean, median, min/max for numerical features.
- **Missing Values:** Imputed using median (numerical) and mode (categorical).
- **Outliers:** Visualized using boxplots; extreme values handled via preprocessing.
- **Class Distribution:** Approximately balanced among 5 classes.
- **Visualization:** Histograms, scatter plots, and correlation heatmaps for feature relationships.

4. Data Preprocessing

- **Categorical Features:** Imputation + OneHotEncoding.
- **Pipeline:** Combined preprocessing using ColumnTransformer.
- **Train/Test Split:** Stratified sampling to preserve class distribution.

5. Model Development

5.1 Algorithms Tested

- Random Forest

- Gradient Boosting
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Multi-Layer Perceptron (MLP/ANN)

5.2 Hyperparameter Tuning

- Performed using **GridSearchCV**.
- Metrics optimized: accuracy and F1-score.

5.3 Best Model

- Random Forest Classifier (with preprocessing pipeline).
- Accuracy on test set: 99.9%.
- Classification Report:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Immediate Danger	1.00	0.99	0.99	199
Mild	1.00	1.00	1.00	205
No Disease	1.00	1.00	1.00	206
Severe	0.99	1.00	0.99	195
Very Mild	1.00	1.00	1.00	195
Accuracy	0.99	0.99	0.99	0.99

6. Feature Importance

- Extracted from Random Forest classifier.
- Top features influencing prediction:

1. Max Heart Rate (thalach)

2. Chest Pain Type (cp)
3. Resting Blood Pressure (trestbps)
4. Age
5. ST Depression (oldpeak)

- Visualization: Bar chart of top 10 important features.

7. Flask API & Frontend

- **API Endpoint:** /predict accepts patient data (JSON) → returns prediction + probabilities.
- **Frontend:** Responsive HTML form for inputting 13 features.
- **Results Display:** Color-coded prediction boxes; per-class probabilities.
- **Testing:** Verified on desktop, tablet, and mobile.

8. Deployment

- **Saved Model:** best_heart_disease_model.pkl
- **Feature Columns:** feature_columns.pkl
- **Class Names:** class_names.pkl

9. Conclusion

- Developed a fully functional heart disease risk prediction system.
- Achieved high accuracy with Random Forest classifier.
- Deployed with Flask API and responsive frontend.
- Ready for hospital use at CHUB, with robust testing and documentation.