Heart Disease Prediction Report

Introduction

This project aims to predict the likelihood of heart disease in individuals based on various health metrics using machine learning models. The primary goal is to leverage data-driven insights to identify patterns and risk factors associated with heart disease, enabling early detection and intervention. The dataset includes a range of health-related features, and three machine learning models—Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting—are employed to classify individuals as having heart disease or not.

Dataset Description

The dataset contains the following features:
- ID: Unique identifier for each individual
- Age: Age of the individual (in years)
- Gender: Gender of the individual (Male/Female)
- Height_cm: Height in centimeters
- Weight_kg: Weight in kilograms
- BMI: Body Mass Index
- Daily_Steps: Number of steps taken daily
- Calories_Intake: Daily calorie intake (in calories)
- Hours_of_Sleep: Hours of sleep per day
- Heart_Rate: Resting heart rate (in beats per minute)
- Blood_Pressure: Blood pressure reading (in mmHg)
- Exercise_Hours_per_Week: Hours of exercise per week
- Smoker: Smoking status (Yes/No)
- Alcohol_Consumption_per_Week: Alcohol consumption per week (in units)
- Diabetic: Diabetic status (Yes/No)
- Heart_Disease: Presence of heart disease (Yes/No, target variable)

The dataset provides a comprehensive view of each individual's health profile, with Heart_Disease as the target variable.

Data Preprocessing

To prepare the dataset for modeling, the following preprocessing steps were applied:
- Categorical Variable Encoding: Categorical features such as Gender (Male=0, Female=1), Smoker (No=0, Yes=1), Diabetic (No=0, Yes=1), and Heart_Disease (No=0, Yes=1) were mapped to numerical values.
- Blood Pressure Transformation: The Blood_Pressure feature was split into two numerical features: Max_BP (systolic) and Min_BP (diastolic).

These steps ensured the dataset was numerical and ready for modeling.

Exploratory Data Analysis

A line plot was created to show the proportion of individuals with heart disease across age groups, highlighting age as a potential risk factor.

Please insert the EDA plot (eda_plot.png) here from the notebook directory.

Model Training and Evaluation

Three models were trained:
- Random Forest: Ensemble of decision trees.
- KNN: Distance-based classifier.
- Gradient Boosting: Sequential tree ensemble.

Performance was evaluated using accuracy, precision, recall, and F1-score.

Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.915	0.837	0.915	0.874
KNN	0.915	0.887	0.915	0.890
Gradient Boosting	0.910	0.837	0.910	0.872

Model Performance Visualization

A bar plot compares the models' performance across metrics.
Please insert the performance plot (model_plot.png) here from the notebook directory.

Conclusion

The project demonstrates the use of machine learning to predict heart disease, with Gradient Boosting showing strong performance. Age was identified as a key risk factor. Limitations include potential class imbalance, which could be addressed in future work using techniques like SMOTE or AUC-ROC metrics. Additional features could further improve predictions. This work provides a foundation for early heart disease detection tools.