Esraa madian

221001883

DR: Mohamed ELsayeh

Cardiovascular Disease Prediction Project Report

1. Introduction

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide. Early prediction of CVDs can significantly improve patient outcomes through timely intervention and lifestyle modifications. This project aims to develop a machine learning model to predict the presence of cardiovascular disease based on various health metrics and lifestyle factors.

2. Dataset Overview

The dataset used in this project is a medical examination dataset containing information from over 60,000 individuals. It includes a wide range of health-related features such as age (in days), gender, height, weight, systolic and diastolic blood pressure, cholesterol and glucose levels, and lifestyle indicators such as smoking, alcohol consumption, and physical activity. The target variable is a binary indicator showing whether a person has been diagnosed with cardiovascular disease. This dataset provides a solid foundation for training machine learning models because it combines biometric data with behavioral risk factors, allowing for a more comprehensive risk assessment. Before model training, the dataset underwent thorough preprocessing including outlier removal, feature engineering, and scaling to ensure high-quality input data.

3. Exploratory Data Analysis (EDA) Insights

Exploratory Data Analysis was conducted to understand the dataset's characteristics and identify potential relationships between features and the target variable. Key findings include: - The dataset is relatively balanced, with an almost equal distribution of patients with and without cardiovascular disease. - Age is a significant factor, showing a strong correlation with the risk of CVD. - Both systolic and diastolic blood pressure levels are highly correlated with cardiovascular disease. - Patients with CVD generally exhibit a higher Body Mass Index (BMI). - Cholesterol and glucose levels demonstrate clear associations with the presence of CVD. - Lifestyle factors such as smoking, alcohol consumption, and physical activity also show some relationship with cardiovascular disease risk.

4. Methodology and Model Building

Several machine learning models were evaluated for their effectiveness in predicting cardiovascular disease: - **Logistic Regression:** A linear model used for binary classification. - **Random Forest:** An ensemble learning method that constructs multiple decision trees. - **Gradient Boosting:** Another ensemble method that builds models sequentially, with each new model correcting errors made by previous ones.

The models were trained and evaluated using standard machine learning practices, including data preprocessing (e.g., scaling numerical features) and hyperparameter tuning. The Gradient Boosting model consistently outperformed the others, achieving the highest accuracy.

5. Results and Evaluation

The Gradient Boosting model achieved an approximate accuracy of 74% in predicting cardiovascular disease. Further optimization through hyperparameter tuning contributed to this performance. The model's performance was assessed using various metrics, including accuracy, classification reports, confusion matrices, and ROC curves.

6. Web Application

A Streamlit web application ( ) was developed to provide an interactive interface for predicting cardiovascular disease risk. The application allows users to: - Input their health information. - View calculated health metrics like BMI and hypertension status. - Receive a prediction of their cardiovascular disease risk. - Access key risk factors and personalized recommendations.

7. Future Improvements

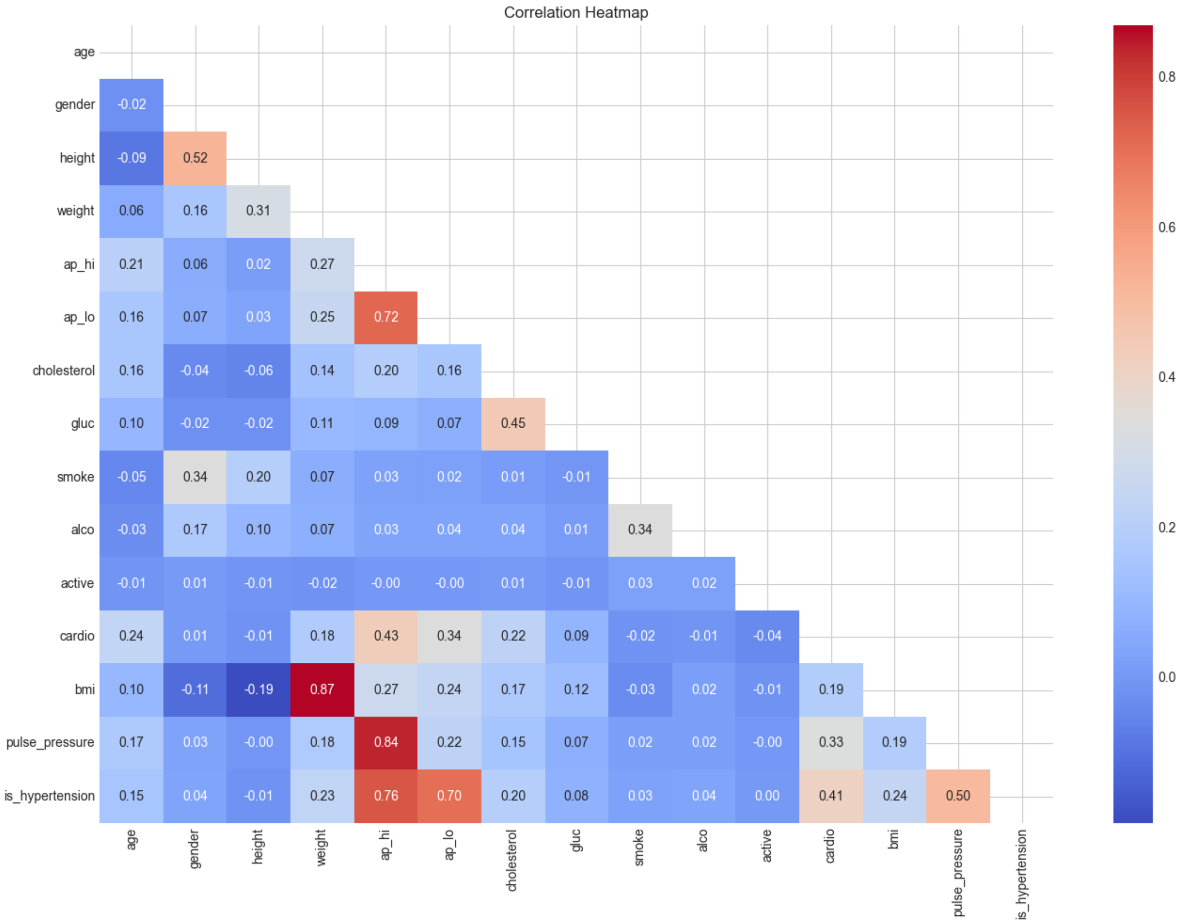
To further enhance the project, the following improvements are suggested: - **Data Collection:** Acquire additional data, especially on more detailed lifestyle factors and comprehensive medical history. - **Advanced Models:** Experiment with more sophisticated machine learning models, such as neural networks, to potentially capture more complex patterns. - **Personalized Strategies:** Develop more personalized risk reduction strategies based on an individual's identified risk factors. - **Medical Guidelines Integration:** Incorporate more detailed medical guidelines into the recommendation system for improved clinical relevance.

8. Visualizations

8.1Exploratory Data Analysis Visualizations

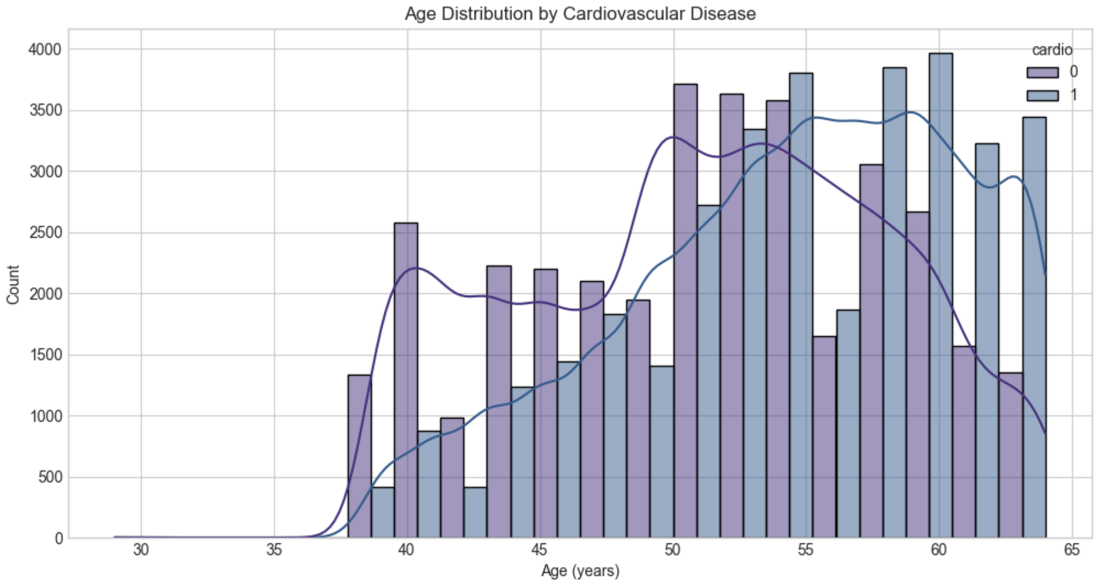
Correlation Heatmap

This heatmap illustrates the correlation between different features in the dataset, providing insights into their relationships.



Age Distribution

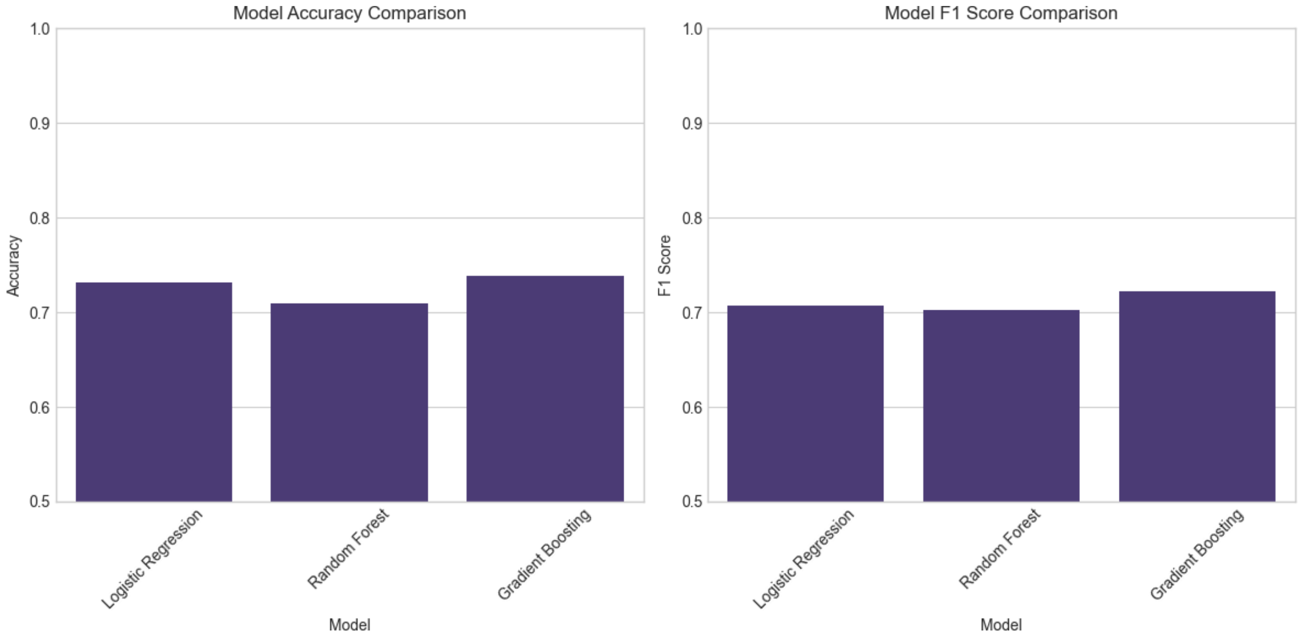
This plot shows the distribution of age within the dataset, highlighting the age groups most represented.



8.1 Model Performance Visualizations

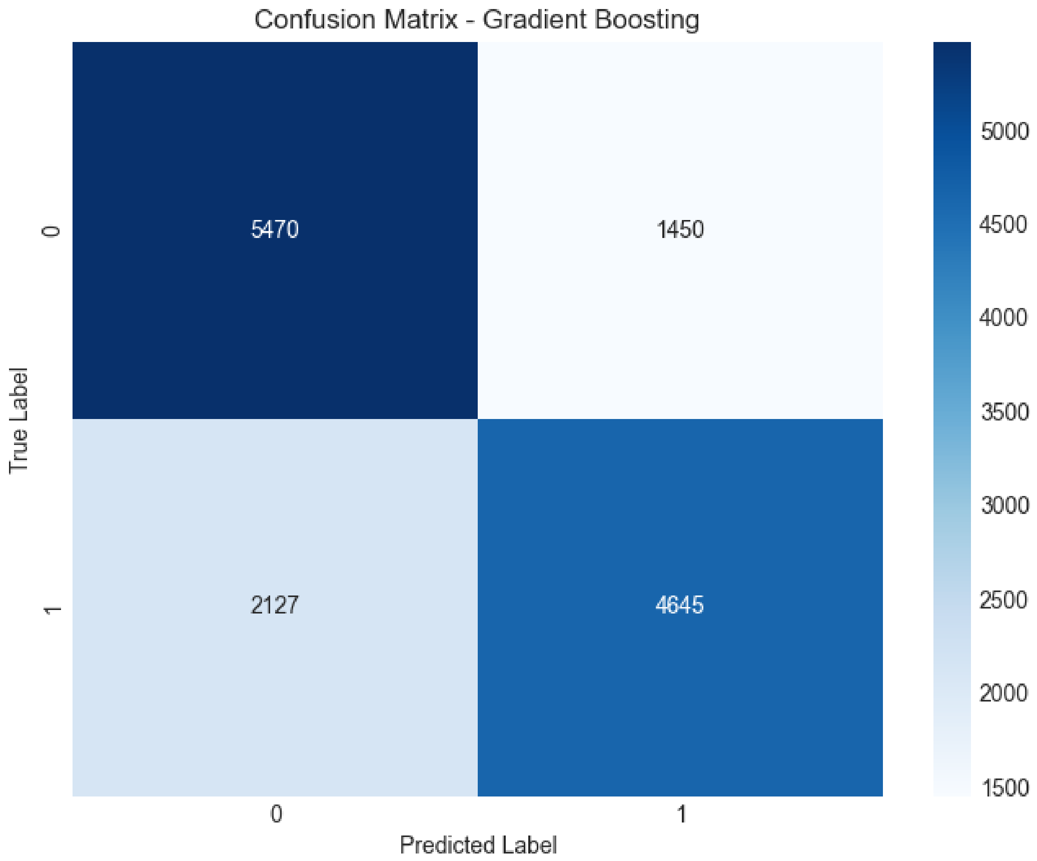
Model Comparison

This chart compares the performance of the different machine learning models evaluated in the project.



Gradient Boosting Confusion Matrix

The confusion matrix for the Gradient Boosting model provides a detailed breakdown of its classification performance, showing true positives, true negatives, false positives, and false negatives.



9. Conclusion

This project successfully developed a machine learning model for cardiovascular disease prediction, demonstrating the potential of data-driven approaches in healthcare. The Gradient Boosting model proved effective, and the accompanying web application provides a user-friendly tool for risk assessment and awareness. The identified areas for future improvement highlight opportunities for further research and development in this critical domain.