Cardiovascular Disease Prediction Project Report

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for millions of lives lost each year. Early detection and prediction of cardiovascular diseases can save lives by enabling timely intervention. This project leverages **machine learning** to predict the likelihood of cardiovascular disease based on various health metrics. The goal is to build a predictive model that can assist healthcare professionals in identifying at-risk individuals.

This report provides a comprehensive overview of the project, covering all the steps from data collection to model deployment, in a way that is easy to understand for beginners.

Project Overview

The project is divided into the following key steps:

1. **Data Collection and Preprocessing**
2. **Exploratory Data Analysis (EDA)**
3. **Feature Engineering**
4. **Model Building**
5. **Model Evaluation**
6. **Predictive System Development**

Each step is explained in detail below, with visualizations and code snippets to make the process clear and engaging.

Step 1: Data Collection and Preprocessing

Data Source

The dataset used in this project is stored in a CSV file named cardio_data_processed.csv. It contains **68,205 records** and **17 features**, including age, gender, height, weight, blood pressure, cholesterol levels, and more.

Data Loading

The dataset is loaded using the pandas library:

Data Inspection

- **Shape**: The dataset has **68,205 rows** and **17 columns**.
- **Columns**: Features include id, age, gender, height, weight, ap_hi (systolic blood pressure), ap_lo (diastolic blood pressure), cholesterol, gluc (glucose levels), smoke, alco (alcohol consumption), active (physical activity), cardio (target variable), age_years, bmi, bp_category, and bp_category_encoded.
- **Missing Values**: There are **no missing values** in the dataset, making it clean and ready for analysis.

Data Preprocessing

- **Mapping Categorical Variables**: The bp_category and bp_category_encoded columns are mapped to numerical values for easier processing:
- **Data Type Conversion**: Columns like gender, cholesterol, gluc, smoke, alco, and active are converted to integers.
- **Gender Mapping**: The gender column is mapped to binary values (0 for female, 1 for male)

Step 2: Exploratory Data Analysis (EDA)

Data Summary

- **Mean Values**: The average age is **52.82 years**, the average BMI is **27.51**, and the average blood pressure category is **1.19**.
- **Distribution**: The dataset is balanced, with **34,533 individuals without cardiovascular disease** and **33,672 with cardiovascular disease**.

Grouped Analysis

- **Cardiovascular Disease vs. No Disease**: Individuals with cardiovascular disease tend to have:
    - Higher average age (**54.46 years** vs. **51.23 years**).
    - Higher average weight (**76.67 kg** vs. **71.59 kg**).
    - Higher average blood pressure (**133.45/84.44 mmHg** vs. **119.60/78.17 mmHg**).
    - Higher cholesterol levels (**1.51** vs. **1.22**).
    - Higher BMI (**28.50** vs. **26.54**).

Step 3: Feature Engineering

Feature Selection

The target variable is cardio, which indicates the presence (1) or absence (0) of cardiovascular disease. All other columns are used as features.

Data Standardization

The features are standardized using StandardScaler to ensure that all features contribute equally to the model:

Step 4: Model Building

Data Splitting

The dataset is split into **training (80%)** and **testing (20%)** sets using train_test_split:

Model Selection

A **Logistic Regression** model is chosen for this classification task due to its simplicity and effectiveness for binary classification problem

Step 5: Model Evaluation

Accuracy on Training Data

The model achieves an accuracy of **72.67%** on the training data:

Accuracy on Test Data

The model achieves an accuracy of **72.90%** on the test data, indicating good generalization:

Step 6: Predictive System Development

Input Data Preparation

An input data sample is prepared for prediction.

Standardization of Input Data

The input data is standardized using the same scaler used for training.

Prediction

The model predicts whether the individual has cardiovascular disease.

Conclusion

This project successfully builds a predictive model for cardiovascular disease using **Logistic Regression**. The model achieves an accuracy of approximately **72.9%** on the test data, demonstrating its potential for early detection of cardiovascular disease.

Key Takeaways

- **Data Preprocessing**: Proper handling of categorical variables and standardization are crucial for model performance.
- **Model Selection**: Logistic Regression is a good starting point for binary classification problems.
- **Evaluation**: Accuracy is a useful metric, but other metrics like precision, recall, and F1-score should also be considered for imbalanced datasets.
- **Predictive System**: The developed system can be integrated into healthcare applications for real-time predictions.

Final Thoughts

This project provides a solid foundation for further exploration and improvement in the field of cardiovascular disease prediction using machine learning. By leveraging data-driven insights, we can make significant strides in early detection and prevention, ultimately saving lives.

Thank you for reading! 🚀