G5. Steve Ramasamy

Project Proposal: Heart Attack Risk Assessment using Machine Learning

1. Introduction: The project aims to develop a Heart Attack Risk Assessment system using machine learning techniques. This system will predict an individual's risk of experiencing a heart attack based on various health parameters and lifestyle factors.

2. Problem Statement: Heart attacks remain a significant health concern globally, and a quick, accessible assessment tool is needed. This project addresses the need for a rapid and straightforward heart attack risk assessment system.

3. Objectives:

Develop a machine learning model to predict the risk of a heart attack based on the provided dataset containing health parameters and lifestyle factors.

Create a building a GUI for users to input their data and receive their heart attack risk prediction.

Demonstrate the feasibility of using machine learning for rapid heart attack risk assessment based on diverse factors.

4. Methodology:

Data Collection: Utilize the provided dataset containing various health parameters and lifestyle factors.

Data Preprocessing: Cleanse the dataset, handle missing values, encode categorical variables, and prepare it for modelling.

Model Development: Use Scikit-learn to build a predictive model to assess heart attack risk based on the dataset's features.

Model Evaluation: Assess the model's performance using appropriate metrics such as accuracy, precision, and recall.

Interface Development: Create a basic command-line interface for users to input their data and receive their heart attack risk prediction.

5. Technologies and Tools:

Python: Utilize Python for data manipulation, model development, and interface creation.

Scikit-learn: Use Scikit-learn for building and evaluating the machine learning model.

Pandas and NumPy: Handle data manipulation and numerical operations.

8. Conclusion: The project will showcase the potential of machine learning in assessing heart attack risk based on diverse health parameters and lifestyle factors. The developed system will offer users a simple interface to estimate their risk of experiencing a heart attack.

Reference

https://link.springer.com/article/10.1007/s11042-023-14817-z#Fun

Introduction

Cardiovascular Disease (CVD) remains a significant threat, impairing the functionality of our heart and blood vessels. Detecting CVD at an early stage is crucial to saving lives. Numerous efforts have been made to achieve this, yet opportunities persist to enhance performance and reliability in detection methods. This study introduces a machine learning-driven approach aimed at predicting heart attacks by analysing diverse risk factors. It employs several machines learning techniques, including Support Vector Machines, Logistic Regression, XGBoost, and a combination of methods, to enhance accuracy and precision in CVD prediction.

**The Data Processing Steps:**

**Data Gathering:** Initially, data was collected from publicly available sources. This involved assessing physical conditions and transforming numerical samples into a format usable by computers.

**Pre-processing:** The second stage focused on addressing data issues such as missing values, outlier identification, and the removal of redundant information to ensure dataset cleanliness.

**Integration:** Through the use of Python, disparate libraries and subsets were consolidated by importing independent modules and merging them for essential experimental procedures.

**Variable Comparison:** Following integration, a comparison of variables was conducted to comprehend correlations, allowing for a comprehensive analysis of similar variables.

**Application of ML Algorithms**: The final phase cantered on selecting optimal ML models for predictive purposes and developing a heart attack risk prediction app using Streamlet.

**The Data Frame**

The dataset consists of 26 columns, with 25 columns representing features and the remaining column, "Heart Attack Risk," serving as the target variable for prediction.

Brief important column explanations:

Age: Age of the individual.

Sex: Gender of the individual (1 for male, 0 for female).

Cholesterol: Levels of cholesterol in the blood.

Heart Rate: The individual's heart rate.

Diabetes: Presence or absence of diabetes (1 for present, 0 for absent).

Family History: History of heart-related issues in the family.

Smoking: Smoking status (1 for smoker, 0 for non-smoker).

Obesity: Indicates obesity status (1 for obese, 0 for non-obese).

Alcohol Consumption: Level of alcohol consumption.

Exercise Hours Per Week: Weekly exercise hours.

Diet: Dietary habits or preferences.

Previous Heart Problems: History of previous heart problems.

Medication Use: Usage of medications.

Stress Level: Level of stress experienced.

Sedentary Hours Per Day: Hours spent in a sedentary state per day.

Income: Income level.

BMI: Body Mass Index of the individual.

Triglycerides: Levels of triglycerides in the blood.

Physical Activity Days Per Week: Days of physical activity per week.

Sleep Hours Per Day: Hours of sleep per day.

Continent: Continent of residence.

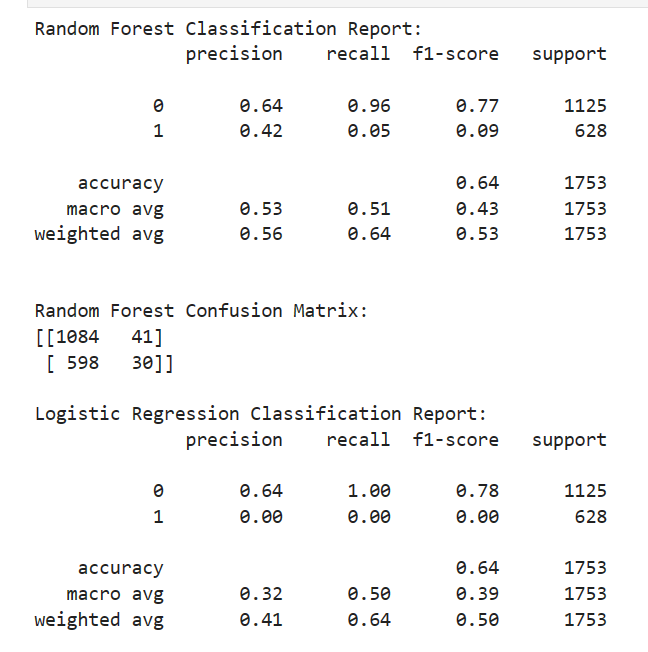
Hemisphere: Hemisphere of residence.

Heart Attack Risk: Target variable indicating the risk of a heart attack (0 for low risk, 1 for high risk).

Systolic\_BP: Systolic blood pressure.

Diastolic\_BP: Diastolic blood pressure.

The "Heart Attack Risk" column contains binary values, framing the problem as a classification task. The dataset comprises a mix of categorical and numerical features, necessitating preprocessing to prepare for analysis. Initial modelling employed Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN). These models achieved accuracy scores of 65%, 52%, 64%, and 66%, respectively. Efforts were made to improve model performance by exploring data reengineering techniques, aiming to augment the dataset and enhance predictive capabilities and overall accuracy.

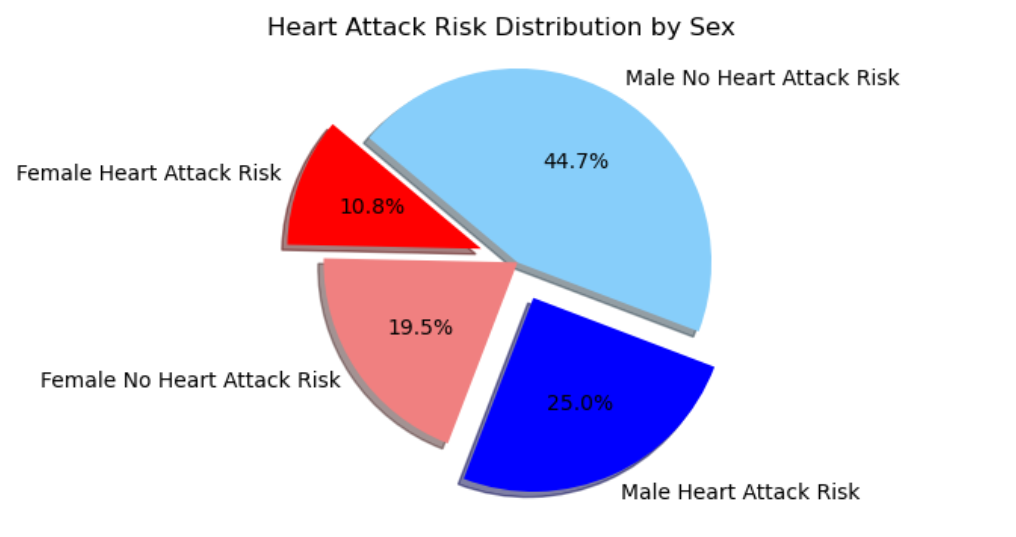
A green and blue rectangular shapes

Description automatically generated

A blue and orange pie chart

Description automatically generated

A graph of a heart attack

Description automatically generated  A graph of blue and green bars

Description automatically generated

A graph with blue and black text

Description automatically generated with medium confidence

