Heart Attack Risk Prediction *using Random Forest and Decision Tree Classifier.*

Google Colab Link: https://colab.research.google.com/drive/1FIAbg_UBkr3PV-NY5Yr-4H1a4H8VIsij#scrollTo=9DUPNIYBdcj1

Dataset Link: https://www.kaggle.com/datasets/m1relly/heart-attack-prediction/code

Introduction

- The objective of this study is to develop machine learning models capable of predicting the risk of heart attack based on various demographic, medical information, lifestyle, and geographical factors. The dataset used for this project contains information about patients, which includes; age, sex, cholesterol levels, blood pressure, heart rate, diabetes status, family history, lifestyle habits, and more. These features are major contributing factors or element of heart risks.

Data Exploration and Preprocessing

Dataset Overview

- The dataset comprises 26 columns, including: Patient ID, Age, Sex, Cholesterol, Blood Pressure, Heart Rate ,Diabetes, Family History, Smoking, Obesity ,Alcohol Consumption, Exercise Hours Per Week, Diet, Previous Heart Problems, Medication Use, Stress Level
- Sedentary Hours Per Day, Income, BMI, Triglycerides, Physical Activity Days Per Week
- Sleep Hours Per Day, Country, Continent, Hemisphere, Heart Attack Risk (Target Variable)

Data Preprocessing,

- Checked for null values and data types.
- Encoded categorical variables (Sex and Diet) to numerical values.
- Split the "Blood Pressure" column into two separate columns (Pressure1 and Pressure2).
- Dropped unnecessary columns (Country, Continent, Hemisphere).

Feature selection

The following features were selected to get the heart risks rates

- Blood Pressure (Pressure1 and Pressure2): The blood pressure has two values, and it was splitted into 2 columns as each value is represents the risk of high blood pressure or not .
- Age, Sex: The sex male and female was converted into binary 1 to represent male and 0 to represent female due to the the fact that the models only takes in numerical values.
- Cholesterol, Heart Rate, Diabetes, Family History
- Smoking, Obesity, Alcohol Consumption
- Exercise Hours Per Week, Diet
- Previous Heart Problems, Medication Use, Stress Level
- Sedentary Hours Per Day, income
- BMI ,Triglycerides
- Physical Activity Days Per Week, Sleep Hours Per Day

Model Development (Random forest classifier)

The dataset was split into training and testing sets using an 80-20 ratio:
Training set: 7010 samples
Testing set: 1753 samples

- **Random Forest Classifier**

A Random Forest Classifier was chosen as the machine learning model due to its ability to handle complex datasets and capture nonlinear relationships between features.

- **Hyperparameter tuning**

GridSearchCV was employed to find the optimal hyperparameters for the Random Forest Classifier. The best hyperparameters identified were; number of estimators: 100, max depth: 10, min_samples_split: 10

- **Model Evaluation**

The evaluation metrics obtained from the classification report are as follows:
Precision: Class 0 (No Heart Attack Risk): 0.63
Class 1 (Heart Attack Risk): 0.00

The accuracy of the random forest classifier model is 63%

Model Development (Decision Tree classifier)

The Decision Tree Classifier was used as an alternative to the Random Forest Classifier to compare their performance in predicting the risk of heart attack based on various factors. Decision trees are straightforward to understand, interpret, and visualize, making them a popular choice for classification tasks. Additionally, Decision Trees can capture complex relationships and interactions between features, potentially providing valuable insights into the underlying patterns in the data.

- **Model evaluation**;

The evaluation metrics obtained from the classification report are as follows:

Precision:	Class 0 (No Heart Attack Risk): 0.66
	Class 1 (Heart Attack Risk): 0.40

Accuracy of the decision tree classifier model is 55 %

Model comparison (Random forest vs Decision tree)

	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	63%	0.00	0.00	0.00
Decision Tree Classifier	55%	0.40	0.43	0.41

Model Comparison (Random Forest vs Decision Tree)

- Random Forest Classifier performs better in terms of accuracy compared to the Decision Tree Classifier (63% vs. 55%).
- Although the Precision, Recall, and F1-score for Class 1 are still not ideal for both models, the Decision Tree Classifier shows slightly better performance with a precision of 0.40, recall of 0.43, and F1-score of 0.41 compared to the Random Forest Classifier's values of 0.00 for precision, recall, and F1-score.
- The Confusion Matrix for the Decision Tree Classifier indicates a more balanced distribution of True Positives and False Positives compared to the Random Forest Classifier, which had a significant imbalance with 1110 True Negatives and 641 False Negatives.

Recommendations

- Both models require further improvement, particularly in predicting positive cases (Heart Attack Risk = 1).
- Addressing the class imbalance, feature engineering, and trying different machine learning algorithms or ensemble methods could help improve the performance of both models.
- Fine-tuning hyperparameters and conducting a more in-depth analysis of the dataset may also contribute to enhancing the predictive power of the models.

Recommendations

- Future work may involve conducting a more in-depth analysis of the dataset to identify additional factors or interactions between variables that could influence the risk of heart attack. Additionally, exploring advanced machine learning techniques and ensemble methods could further enhance the predictive power of the model.