FINAL PROJECT

OBJECTIVES OF THE PROJECT

The objectives of this project are as follows;

1. Create a robust predictive model for assessing a patient's risk of experiencing a heart attack.
2. Identify the primary explanatory factors influencing the occurrence of heart attacks.

SIGNIFICANCE OF THE PROJECT

This project holds significant value as it addresses two crucial aspects in the field of healthcare and medical research. Firstly, by developing an accurate predictive model for heart attack risk, it can potentially save lives by enabling early intervention and personalized care for patients. Secondly, uncovering the key explanatory variables influencing heart attacks contributes to a deeper understanding of the disease, aiding in the development of preventive measures and targeted treatments, thus advancing the field of cardiology and overall public health.

DATA

- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest pain type
  - 1: typical angina
  - 2: atypical angina
  - 3: non-anginal pain
  - 4: asymptomatic
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholesterol in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results
  - 0: normal

    - 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise-induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- num: suffering from heart attack (1 = yes; 0 = no)

DATA EXPLORATION

Missing Values

- I deleted the columns with many missing values, which are slope, ca and thal
- The rows with missing values were removed.

Outliers

- Z-score method was used to check for outliers
- There were no outliers in the data

Others

The target variable was renamed from 'num' to 'heart attack'

CLASSIFICATION MODELS

This project used six (6) models. As seen in the table below, K-Nearest Neighbour is the best performing model in terms of accuracy, precision, F-score and Recall. Therefore I selected KNN for this project.

MODELS	ACCURACY	PRECISION	F_SCORE	RECALL
LOGISTIC REGRESSION	0.811321	0.75	0.75	0.75
K-NEAREST NEIGBOUR	0.867925	0.842105	0.820513	0.8
SUPPORT VECTOR MACHINES	0.811321	0.75	0.75	0.75
RANDOM FOREST	0.811321	0.75	0.75	0.75
GRADIENT BOOSTING	0.792453	0.736842	0.717949	0.7
STACKING	0.792453	0.764706	0.702703	0.65

KEY FINDINGS

Objective 1: Create a robust predictive model for assessing a patient's risk of experiencing a heart attack.

The provided results are from a K-Nearest Neighbors (KNN) classification model, and they include various performance metrics, including accuracy, precision, F-score, and recall. These metrics are used to evaluate the model's performance in classifying whether a patient has a heart attack (1) or does not have a heart attack (0).

1. Accuracy: Accuracy is a measure of the overall correct classification rate of the model. In this case, the KNN model has an accuracy of approximately 0.8679, which means it correctly classifies around 86.79% of the samples. This is a good overall accuracy.

2. Precision: Precision measures the ability of the model to make correct positive predictions (1: patient has a heart attack) among all positive predictions made. The precision of approximately 0.8421 indicates that when the model predicts that a patient has a heart attack (1), it is correct about 84.21% of the time.

3. F-Score: The F-score is the harmonic mean of precision and recall. It provides a balance between these two metrics. An F-score of approximately 0.8205 is a good score and indicates that the model achieves a good trade-off between precision and recall.

4. Recall: Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify all actual positive cases (1: patient has a heart attack). The recall value of 0.8 means that the model correctly identifies 80% of the patients who actually have a heart attack.

In summary, the KNN model shows promising results for predicting whether a patient has a heart attack or does not have a heart attack. It has a high accuracy, indicating that it correctly predicts the majority of cases. The precision and recall values are also relatively high, suggesting that the model is good at correctly identifying patients with heart attacks and making positive predictions when necessary.

Objective 2: Identify the primary explanatory factors influencing the occurrence of heart attacks.

Permutation Feature Importance method was used to identify which variables are most informative for the model's predictions. The findings indicate that the four (4) most important variables that determines whether or not a person has or will have heart attack are as follows;

1. ST depression induced by exercise relative to rest
2. chest pains
3. resting blood pressure
4. maximum heart rate achieved

LIMITATION OF THE MODEL AND FUTURE REVISION

The dataset I used was imbalanced. Imbalanced classes can pose challenges in machine learning, as models trained on such data tend to be biased towards the majority class, potentially leading to poor performance on the minority class.

To address class imbalance, I may consider employing techniques such as resampling (oversampling or undersampling), using different algorithms, adjusting classification thresholds, or applying cost-sensitive learning, depending on my specific problem and dataset. These strategies can help ensure that my model performs well for both the majority and minority classes.