COMP4433 Data Mining & Data Warehousing

Heart Attack Analysis & Prediction Dataset

Individual Project Report

Siyu Zhou

(21094655D)

Dec 14th , 2024

Table of Contents

1. Introduction

This report is basically based on the dataset from Kaggle, Heart Attack Analysis & Prediction Dataset [1]. The objective is to analyze datasets, implement data mining solutions with different models, and then evaluate their performance. The analysis and methods mentioned in this report are implemented in notebook link in Appendix 1.

2. Exploratory Data Analysis

2.1 Dataset Overview

The Heart Analysis Prediction Dataset contains 303 $records$ and 13 $attributes$ with 1 target variable ($output$).

The features are categorized into categorical features and numerical types for focused analysis.

- **Categorical**: sex, cp (chest pain type), fbs (fasting blood sugar), $restecg$ (resting electrocardiographic results), slp (slope), caa (number of major vessels), $thall$ (Thallium Stress Test), $exng$ (exercise induced angina). These features represent different aspects of a patient's medical profile relevant to heart health.
- **Numerical:** age, $trtbps$ (resting blood pressure), $chol$ (cholesterol), $thalachh$ (maximum heart rate achieved), $oldpeak$. These features represent measures of health status.

Missing Data: the dataset has no missing data.

Duplicate Data: for records in dataset, there is one redundant record with index 164.

Distribution of output: The distribution showed a relatively balanced data set for output in Figure 1.

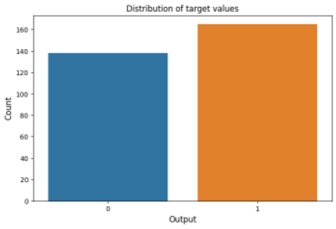

Figure 1 Distribution of Output

2.2 Features Correlation Analysis

A correlation analysis was performed between different features, the result shown in Figure 2.

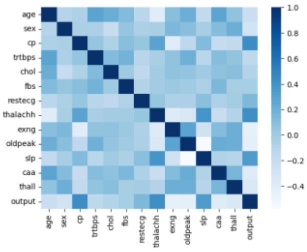

Figure 2 Correlation between features

Positive Correlations with output:

- **Chest Pain Type ($\boldsymbol{cp}$)**: A strong positive correlation (0.434) suggests chest pain types (with higher number) are more associated with higher risk of heart attack.
- **Maximum Heart Rate Achieved ($\boldsymbol{thalachh}$):** A strong positive correlation here suggests that higher maximum heart rates are linked to higher heart attack possibility.

Negative Correlations with output

- **Previous Peak ($\boldsymbol{oldpeak}$):** A significant negative correlation (-0.431) indicates that the higher previous peak is associated with lower risk of heart attack.
- **Exercise Induced Angina ($\boldsymbol{exng}$):** A significant negative correlation (-0.437) indicates that presence of exercise-induced angina is higher associated with lower probability of heart attack.

Other Important correlation

- **Age:** negative correlation (-0.225) with output, older age might lead to lower risk.
- **Sex:** negative correlation (-0.281) with output, possibly lower risk in heart attack.

Other features

- Features such as fbs (Fasting Blood Sugar) and $chol$ (Cholesterol) show a weaker correlation with the probability of heart attack.

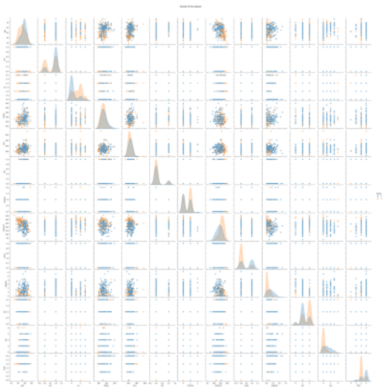

Figure 3 Pair Plot Correlation between Features and Output

From Figure 4, we can induce the following findings from pair plot of each feature with the output.

- Object with $cp = 2$ have more heart attack, with $cp = 0$ have less heart attack.
- Object with $rest\_ecg = 1$ have higher chance of heart attack (having ST-T wave abnormality).
- Object with higher $thalachh$ have higher heart attack risk (higher Maximum heart rate achieve).
- Object with $exng = 0$ have higher chance of heart attack.
- Object with lower $oldpeak$ have higher chances of heart attack.
- Object with $slp = 2$ have higher heart attack risk; with $slp = 1$ have lower heart attack risk.
- Object with $caa = 0$ have higher chance of heart attack.
- Object with $thall = 2$ have much higher chance of heart attack.

2.3 Categorical Feature

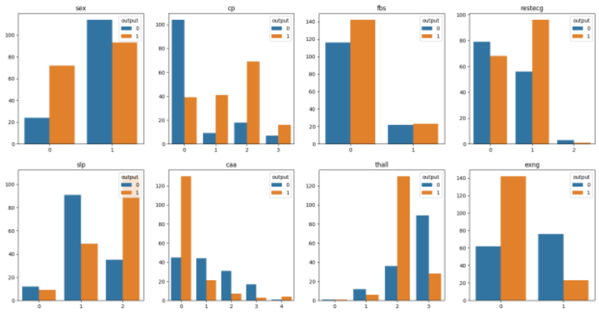

Figure 4 Categorical Features' Counter Plot

From Figure 5 above, it shows the distribution between features and the target variable and their underlining associations with heart attack risk. The following are observations.

- **Sex ($\boldsymbol{gender}$):** 207 instances in the dataset of sex are 1 and others 96 are 0. Higher representation of one gender in dataset could be an important factor in prediction.
- **Chest Pain Type ($\boldsymbol{cp}$):** 143 subject experiences type '0' chest pain, type '2' subject are 87, type '1' are 50, object with type '3' have 3. Different types of chest pain vary significantly among subjects and might be a crucial factor in predicting heart attack risk.
- **Fasting blood sugar ($\boldsymbol{fbs}$):** Objects having $fbs - 0$ are more than four times the objects having $fbs - 1$. This feature provides insights into patients' health.
- **Resting Electrocardiographic ($\boldsymbol{restecg}$):** $restecg - 0$ (147) and $restecg - 1$ (152) have similar counts, $restecg - 2$ have the lowest count, $restecg - 2$ is almost negligible. This feature indicates some underlying heart conditions.
- **Slope ($\boldsymbol{slp}$):** $slp - 1$ (140) & $slp - 2$ (142) have similar counts, $slp - 0$ have smaller counts (21). The slope may correlate with the patient's heart health.
- **Major vessels ($\boldsymbol{caa}$):** $caa - 0$ have the largest count (175), with the decreasing counts for $caa - 1$ (65), $caa - 2$ (38), $caa - 3$ (20), and $caa - 4$ have the lowest count. This might relate to the severity of coronary diseases.
- **Thallium Stress Test Result** ($thall$): $thall - 2$ have the largest count, following with $thall - 3$ (117), $thall - 1$ (18), and $thall - 0$ have the lowest count (2). The presence and type of thallium stress might be a significant factor in heart attack risk.
- **Exercise Induced Angina** ($exng$): $exng - 0$ (204) have more counts than $exng - 1$ (99). This feature may be important in understanding exercise-related heart attack risk.

2.4 Numeric Features

The distribution of numerical features provides insights into the range and distribution patterns of key numbers and medical test data in the dataset. The following are observations for Figure 6.

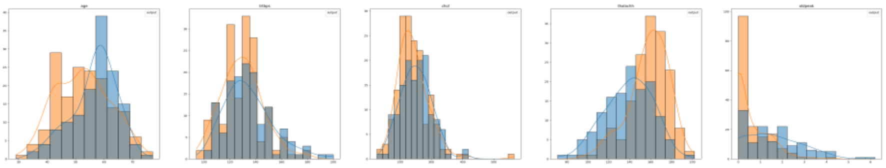

Figure 5 Numerical Feature Counter Plots

- $\boldsymbol{age}$: the age distribution appears normally distributed with slight right skewed, indicating higher probability of heart attacks in age group of $40 - 60$.
- **Resting Blood Pressure ($\boldsymbol{trtbps}$):** the distribution of resting blood pressure shows normal distribution, $trtbps$ group range in $120 - 140$ mmHg has more heart attacks.
- **Cholesterol ($\boldsymbol{chol}$):** The cholesterol distribution is right skewed. The $chol$ group range $200 - 300$ have higher probability of heart attacks.
- **Maximum Heart Rate Achieved ($\boldsymbol{thalachh}$):** The distribution of $thalachh$ is left-skewed. The $thalachh$ group 150-200 have a higher chance of heart attacks.
- **Previous Peak ($\boldsymbol{oldpeak}$):** the $oldpeak$ distribution shows a highly right-skewed distribution with a higher number exhibiting lower values. $oldpeak$ Group 0-2 have more heart attacks.

3. Data Preprocessing

3.1 Data Cleaning

The initial step of data preprocessing is to remove duplicated data from the dataset. This step is crucial to ensure the model is being trained to ensure data quality, integrity and avoid any bias that might occur due to repeated information. After removing the duplicate record ($index$: 164), the dataset includes 302 $records$ with 13 $attributes$ and 1 $target\ variable$ ($output$). And for missing value, the dataset does not have any missing value. The dataset does not need to be handled through imputation or removal.

3.2 One-hot Encoding for categorical features

The next step is to handle the categorical variable like cp and $thall$. One-hot encoding is applied to convert categorical feature into numerical format through $get\_dummies()$ function which could be better interpreted by the models. Each category is converted into separate binary columns, ensuring that no information is lost. This transformation increases the number of attributes in the dataset from the original 13 attributes to 22 attributes. One-hot encoding is crucial in preprocessing for recognizing categorical data instead of assuming natural order in categories.

3.3 Normalize numerical features

The third step of preprocessing data is performed using $StandardScaler$. This step is critical in making all features to the same scale based on mean and standard deviation. This normalization helps to improve model's performance by reducing skewness due to variable scales in data.

4. Modelling Methods Comparison

The dataset ware divided into training and testing set, with the test size of 20%. The evaluation metrics based on $accuracy\_score$. The training set consists of 241 samples, test set has 61 sample.

4.1 Classification based Approaches

Different kinds of classification models are trained and evaluated to predict heart attacks.

4.1.1 Classification Models

- **Logistic Regression**: A statistical model predicts probability of heart attack based on variables.
- **Decision Tree**: A tree-based regression model that splits data based on feature thresholds to minimize variance in target variable within each node. This model is useful for its interpretability but requires regularization to prevent overfitting.
- **Random Forest**: This is an ensemble methods of decision trees, with bagging methods. This method is robust and better ability in avoid overfitting.
- **Gradient Boosting**: boosting combines prediction of multiple weak learner to create single more accurate stronger learning algorithm.
- **Support Vector Regressor (SVM)**: A regression model that aims to fit a function within a margin of tolerance around actual data.
- **K-Nearest Neighbors Regressor (KNN)**: A non-parametric regression model that predicts value by averaging values of k nearest neighbors in the feature space. KNN is simple and intuitive but scales poorly with dataset size.
- **Gaussian Naïve Bayes**: This method is based Bayes theorem, the model is suitable for classification with features following normal distribution.
- **X Gradient Boosting (XGBoost)**: An advanced gradient boosting framework optimized for speed and performance, including regularization parameters to prevent overfitting and handle non-linear relationships effectively.

4.1.2 k-fold cross validation

Cross-validation minimizes overfitting by simulating performance on unseen data. A **k-fold cross-validation** strategy was employed with $k = 10$. The dataset was split into 10 equal parts. In each fold, 9 parts were used for training, and 1 part was used for validation. The process was repeated 10 times, with each fold as the validation set, ensuring that model is evaluated robustly by testing its performance on multiple splits, minimizing the overfitting [2].

Classification Models	Accuracy on Testing Data
Logistic Regression	100.00 %
Decision Tree	100.00 %
Random Forest	100.00 %
Gradient Boosting	100.00 %
XGBoost	100.00 %
SVM	100.00 %
KNN	95.08%
Gaussian Naive Bayes	100.00 %
Neural Network	98.36 %

Table 1 Classification Model's Accuracy

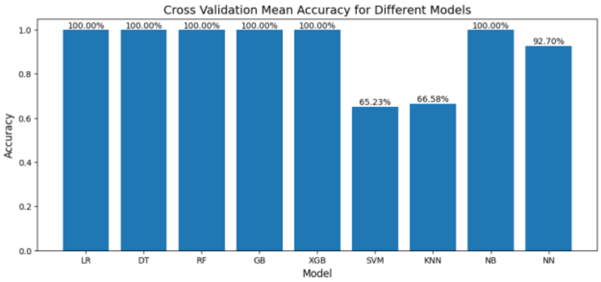

Figure 6 Cross validation Scores

Figure 7 above illustrates cross-validation score for various models. Ensemble methods like Random Forest, Gradient Boosting, and XGBoost achieved 100% accuracy, due to their robustness. Simple models like Logistic Regression and Naive Bayes also achieve 100% accuracy. Neural Networks with 92.7%, high accuracy show adaptability. In contrast, traditional models like SVM and KNN performed lower at 65.23% and 66.58%, likely due to their limitations with high-dimensional data.

4.2 Clustering-based Approaches

Clustering techniques, such as K-Means and Agglomerative Clustering, and DBSCAN models, were applied.

4.2.1 PCA Visualization of Clustering Prediction Methods

To enhance interpretability of clustering results, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset to two principal components [3]. This enabled visualization of the clusters as following figures shows:

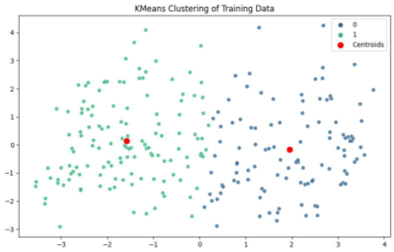

Figure 7 PCA – KMeans Training

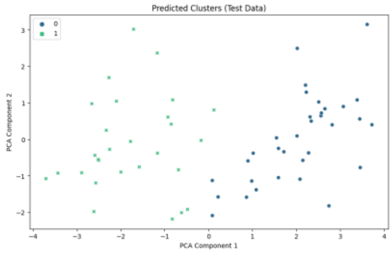

Figure 8 PCA - KMeans Predicted Cluster

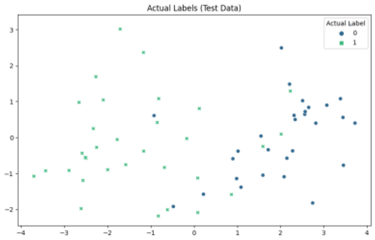

Figure 9 PCA - Actual Labels

Figure 8 – 10 shows the PCA Visualization for KMeans Clustering method, the PCA scatter plot shows two well-separated clusters. For figure 8, the separation of clusters in training data validates the effectiveness of K-Means on the training data, although some overlap is visible due to noise and feature limitations. For the testing data, the cluster would have some overlapping regions, which indicates misclassifications, and needs enhancement like feature engineering. Other Clustering Methods are also shown in PCA Feature Visualization in the notebook.

4.2.2 Clustering Models

Three clustering models were implemented to group similar patient profiles:

- **K-Means Clustering**: Partitioned the data into k clusters based on the mean values of features. This method highlighted distinct patient groups but was sensitive to initialization and outliers.
- **Agglomerative Clustering**: A hierarchical approach that iteratively merged clusters. It provided a dendrogram to visualize relationships between data points, effective for small datasets.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Focused on density regions, effectively handling noise and outliers. However, performance was sensitive to the choice of epsilon and minimum samples.

4.2.3 Evaluation Accuracy

For accuracy on the testing data shown in Table 2, **KMeans Clustering** achieved an accuracy of 86.89%, reflecting its ability to form reasonable distinct clusters that align with true labels, but the sensitivity to noise may impact the precision, and may lead to some overlaps. **Agglomerative Clustering's** accuracy is slightly lower at 85.25%, indicating its hierarchical approach understand most of underlying structure, but struggle with near datapoint. **DBSCAN** has the lowest accuracy at 67.21%, due to its density-based characteristics, which may misclassify sparse regions. This approach is less suitable for Heart Attack prediction.

Clustering Models	Accuracy on Testing Data
KMeans	86.89%
Agglomerative	85.25%
DBSCAN	67.21%

Table 2 Clustering Models' Accuracy

4.3 Models comparison

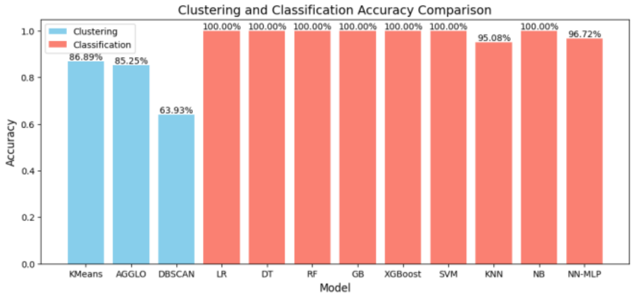

Figure 10 Clustering & Classification Accuracy Comparison

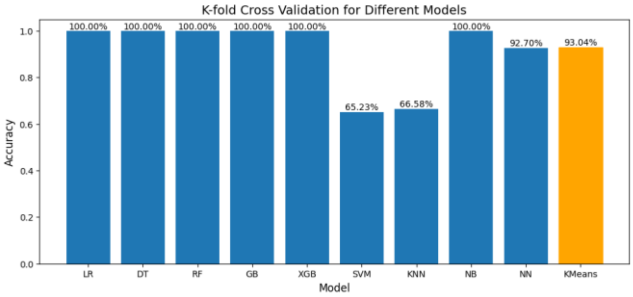

Figure 11 Cross-validation Score comparison

The accuracy comparison of clustering and classification models is illustrated in Figure 11. Clustering models, led by K-Means (86.89%) and Agglomerative Clustering (85.25%), demonstrated effective grouping but fell short of classification models, which all achieved near-perfect accuracies. Classification methods like Random Forest, Gradient Boosting, and XGBoost achieved 100% accuracy, highlighting their robustness for predictive tasks.

Figure 12 presents the results of k-fold cross-validation. Classification models consistently achieved high accuracies, with Neural Networks (92.70%) and K-Means clustering (93.04%) also performing strongly. This comparison emphasizes the dominance of classification models for predictive accuracy, while clustering methods, though slightly less accurate, remain valuable for exploratory analysis and segmentation tasks. These visualizations emphasize the superiority of classification models for predictive accuracy in this dataset, with ensemble methods excelling in both training and cross-validation phases.

5. Associate Rule Mining – Apriori Analysis

Association Rule Mining could be applied to the data set to reveal uncovered patterns that might indicate the factors that are related to heart attack [2]. Apriori algorithm was applied here.

Data Transformation

- Continuous variables are split into three bins for each feature, which facilitates the identification of association rules [4].
- Categorical variables and output are labeled in binary format, for example, convert sex to $sex - 0$ and $sex - 1$, which is clearer to distinguish categories [4].

Rule Mining

Association rules mining uncovered key patterns within the dataset. With the $min\_sup = 0.25$ and $min\_conf = 0.75$, the ARM focus on rules that have strong associations with the probability of heart attack. The top rules identified are summarized below:

- **Rule 1:** $\boldsymbol{exng - 0, oldpeak(-0.0062, 2.067] \rightarrow output - 1}$
  - Support: 0.4488 | Confidence: 0.7432
  - Patients with $exng = 0$ (no exercise-induced angina) and $oldpeak$ in the range of $(-0.0062, 2.067]$ are likely to belong to the high-risk group ($output = 1$).
- **Rule 2:** $\boldsymbol{caa - 0 \rightarrow output - 1}$
  - Support: 0.4290 | Confidence: 0.7428
  - Patients with $caa = 0$ (no major vessels colored by fluoroscopy) are likely to belong to $output = 1$.
- **Rule 3:** $\boldsymbol{thall - 2 \rightarrow output - 1}$
  - Support: 0.4290 | Confidence: 0.7831
  - Patients with $thall = 2$ (fixed defect in the thalassemia test) are at high risk of having heart attack ($output = 1$).
- **Rule 4:** $\boldsymbol{thall - 2, oldpeak(-0.0062, 2.067] \rightarrow output - 1}$
  - Support: 0.4158 | Confidence: 0.8235
  - Patients with $oldpeak$ in range (-0.0062, 2.067] and $thall = 2$ show a stronger association with high risk ($output = 1$).
- **Rule 5:** $\boldsymbol{caa - 0, oldpeak(-0.0062, 2.067] \rightarrow output - 1}$
  - Support: 0.3861 | Confidence: 0.7405
  - Patients with $caa = 0$, $oldpeak$ in range $(-0.0062, 2.067]$ are likely to belong to $output = 1$.
- **Rule 6:** $\boldsymbol{exng - 0, thall - 2 \rightarrow output - 1}$
  - Support: 0.3762 | Confidence: 0.8444
  - Patients with $exng = 0$ (no exercise-induced angina) and $thall = 2$ have a strong association with $output = 1$.

These rules highlight critical risk factors such as $thall$, $oldpeak$, and caa in determining high-risk groups. The high confidence levels associated with these rules demonstrate their potential usefulness in predictive analytics and guiding medical decisions.

6. Improvement

To finetune KNN and Random Forest classification model, parameter optimized is conducted to optimize accuracy on testing data [5]. As Table 1 shows, the optimal number of neighbors for KNN is found to be 23, which results in accuracy of 95.08%. The optimal number of estimators for Random Forest is found to be 3 with 100% accuracy.

Reference

[1] “Heart Attack Analysis & Prediction Dataset,” *www.kaggle.com*, 2020. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data (accessed Dec. 12, 2024).

[2] “heart-attack-analysis python,” *kaggle.com*, 2022. https://www.kaggle.com/code/licgsg/heart-attack-analysis-python (accessed Dec. 12, 2024).

[3] “DLInBMI_Assignment1_zz2721,” *Kaggle.com*, Feb. 18, 2022. https://www.kaggle.com/code/zixiangzhao001/dlinbmi-assignment1-zz2721 (accessed Dec. 12, 2024).

[4] “Association Rules with Python,” *kaggle.com*, 2020. https://www.kaggle.com/code/mervetorkan/association-rules-with-python

[5] rocklen, “Heart Attack Analysis & Prediction,” *Kaggle.com*, Feb. 26, 2022. https://www.kaggle.com/code/rocklen/heart-attack-analysis-prediction (accessed Dec. 12, 2024).

Appendix 1 Notebook Link

Notebook Link: https://www.kaggle.com/code/siyuuzoeyzhou/heart-attack-prediction