Disclaimer of Liability

The material and information contained on this website is for general information, reference, and self-learning purposes only. You should not rely upon the material or information on the website as a basis for making any academic, business, legal or any other decisions. You should not copy any material or information on the website into any of your academic, business, legal or any other non-private usages. ZHANG Wengyu will not be responsible for any consequences due to your violations.

Whilst ZHANG Wengyu endeavours to keep the information up to date and correct, ZHANG Wengyu makes no representations or warranties of any kind, express or implied about the completeness, accuracy, reliability, suitability or availability with respect to the website or the information, products, services or related graphics contained on the website for any purpose. Any reliance you place on such material is therefore strictly at your own risk.

ZHANG Wengyu will not be liable for any false, inaccurate, inappropriate or incomplete information presented on the website.

Although every effort is made to keep the website up and running smoothly, due to the nature of the Internet and the technology involved, ZHANG Wengyu takes no responsibility for and will not be liable for the website being temporarily unavailable due to technical issues (or otherwise) beyond its control or for any loss or damage suffered as a result of the use of or access to, or inability to use or access this website whatsoever.

Certain links in this website will lead to websites which are not under the control of ZHANG Wengyu. When you activate these you will leave ZHANG Wengyu's website. ZHANG Wengyu has no control over and accepts no liability in respect of materials, products or services available on any website which is not under the control of ZHANG Wengyu.

To the extent not prohibited by law, in no circumstances shall ZHANG Wengyu be liable to you or any other third parties for any loss or damage (including, without limitation, damage for loss of business or loss of profits) arising directly or indirectly from your use of or inability to use, this site or any of the material contained in it.

THE HONG KONG POLYTECHNIC UNIVERSITY

Department of Computing

Individual Project Report

COMP4433 Data Mining and Data Warehousing

ZHANG Wengyu

Dec. 2023

Table of Content

1. Exploratory Data Analysis

1.1 Dataset Overview

The Heart Attack Analysis and Prediction Dataset contains **303 records** and **13 attributes** with one target variable (`output`). The detailed description of each attribute is as follows:

- `age` : Age of the patient
- `sex` : Sex of the patient
  - `1` , `0`
- `cp` : Chest pain type
  - `0` = Typical Angina, `1` = Atypical Angina, `2` = Non-anginal Pain, `3` = Asymptomatic
- `trtbps` : Resting blood pressure (in mm Hg)
- `chol` : Cholestoral in mg/dl fetched via BMI sensor
- `fbs` : (fasting blood sugar > 120 mg/dl)
  - `1` = True, `0` = False
- `restecg` : Resting electrocardiographic results
  - `0` = Normal, `1` = ST-T wave normality, `2` = Left ventricular hypertrophy
- `thalachh` : Maximum heart rate achieved
- `oldpeak` : Previous peak
- `slp` : Slope
  - `0` , `1` , `2`
- `caa` : Number of major vessels
  - `0` , `1` , `2` , `3` , `4`
- `thall` : Thalium Stress Test result
  - `0` , `1` , `2` , `3`
- `exng` : Exercise induced angina
  - `1` = Yes, `0` = No
- `output` : Target variable
  - `1` = Heart disease, `0` = No heart disease

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

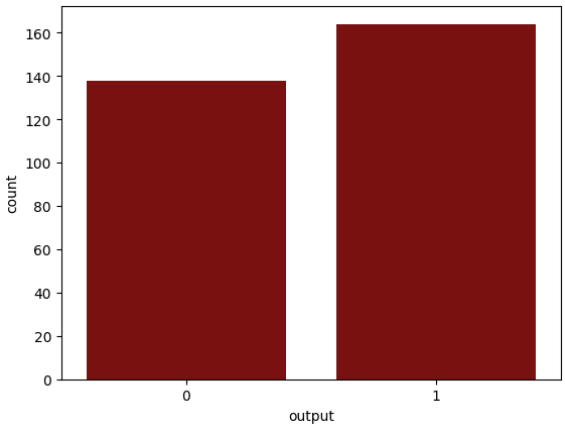
- **Categorical Features**: These 8 of them, including 'sex', 'cp' (chest pain type), 'fbs' (fasting blood sugar), 'restecg' (resting electrocardiographic results), 'slp' (slope), 'caa' (number of major vessels), 'thall' (Thalium Stress Test), and 'exng' (exercise induced angina). Each of these features represents a key aspect of a patient's medical profile relevant to heart health.
- **Numerical Features**: These are 5 of them, including 'age', 'trtbps' (resting blood pressure), 'chol' (cholesterol), 'thalachh' (maximum heart rate achieved), and 'oldpeak'. They provide quantifiable measures of a patient's health status.

Missing Data: The dataset has no missing data, which indicates that no missing data handling is required.

Duplicate Data: Examination for duplicate records indicates one redundant record with index `164`.

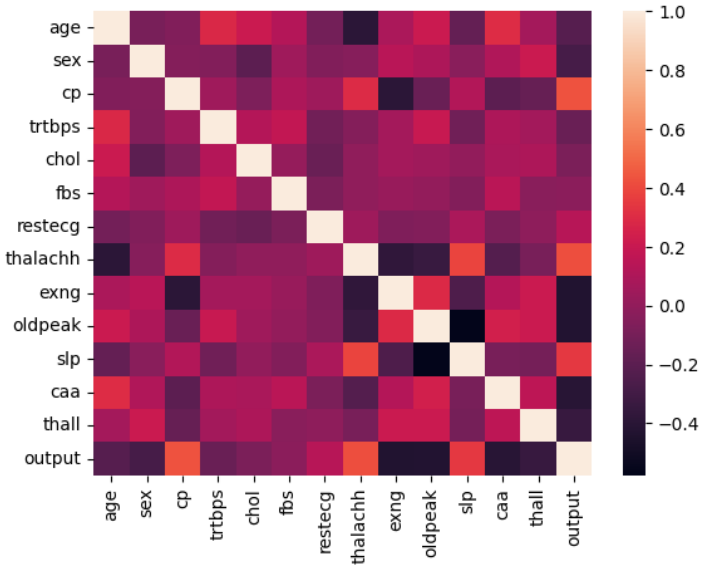
Distribution of Target Variable (`output`)

The distribution of the target variable showed a relatively balanced dataset concerning heart attack risk as shown in the figure below.



Distribution of Target Variable

1.2 Correlated Feature Analysis on `output`

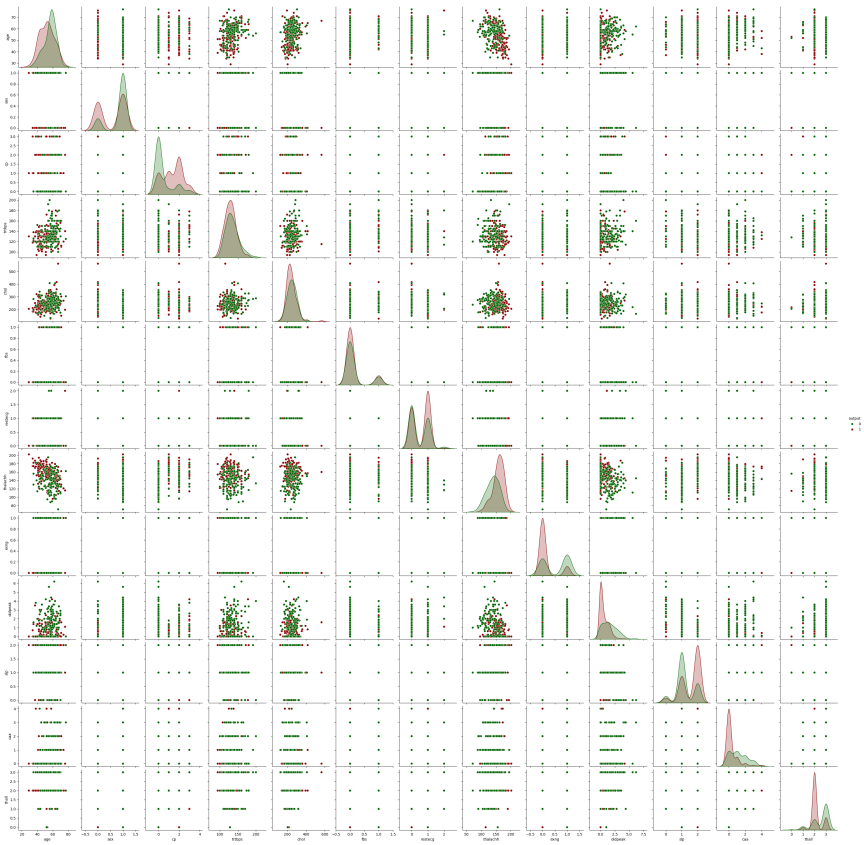


Heatmap of Correlations between Features and Output

The correlation matrix provides valuable insights into the relationships between various features and the heart attack risk (`output`). Key observations include:

- **Positive Correlations with Output:**
  - **Chest Pain Type (`cp`):** A strong positive correlation (0.434) suggests that certain types of chest pain are more associated with a higher risk of heart attack.
  - **Maximum Heart Rate Achieved(`thalachh`):** A significant positive correlation (0.422) indicates that higher maximum heart rates are linked to increased heart attack risk.
- **Negative Correlations with Output:**
  - **Exercise Induced Angina (exng):** Exhibits a notable negative correlation (-0.437), implying that the presence of exercise-induced angina decreases the likelihood of a heart attack.

    - **Oldpeak (Previous Peak):** A negative correlation (-0.431) suggests that higher Previous Peak is associated with a lower risk of heart attack.
- **Important Correlations:**
    - **Age:** Shows a negative correlation (-0.225), indicating that older age might be linked to a lower risk, which could be counterintuitive to the general perception of heart attack risk.
    - **Sex:** Also negatively correlated (-0.281) with `output`, indicating a possible lower risk in one of the sexes.
- **Weak or No Significant Correlations:**
    - Features such as **Fasting Blood Sugar (fbs)** and **Cholesterol (chol)** showed weaker correlations with the heart attack risk.



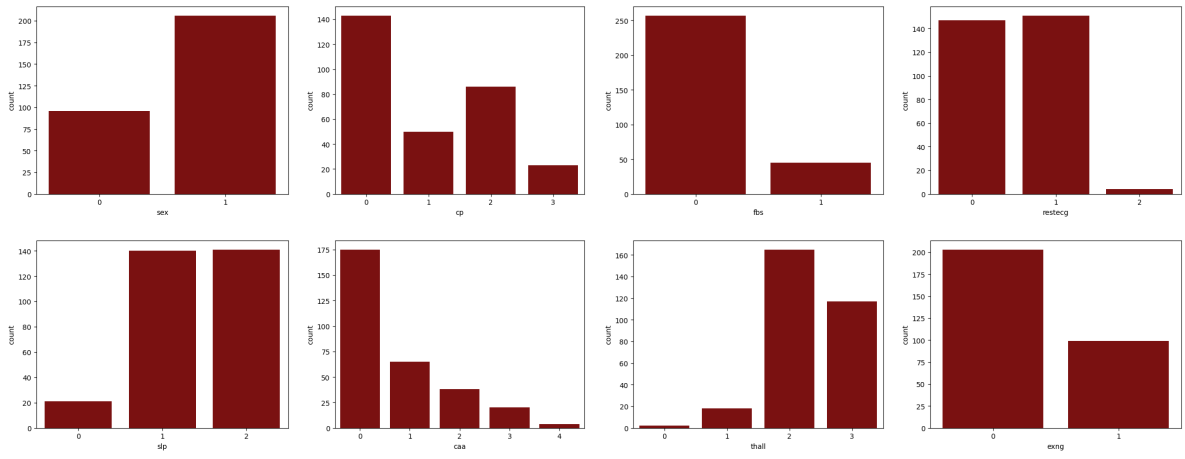
Pair Plot of Correlations between Features and Output

1.3 Category Features

The distribution of categorical features in the dataset provides insights into the characteristics of the subjects and their potential associations with heart attack risk. Following are some observations:

- **Sex (Gender):**
    - The dataset includes 207 instances of sex as '1' and 96 as '0'.
    - This suggests a higher representation of one gender in the dataset, which could be a important factor in the analysis and prediction of heart attack risk.
- **Chest Pain Type (cp):**
    - The majority of subjects (143) experiences type '0' chest pain.
    - Types '2' (87), '1' (50), and '3' (23) follow in frequency.

    - This indicates that different types of chest pain vary significantly among subjects and might be a crucial factor in predicting heart attack risk.
- **Fasting Blood Sugar (fbs):**
    - A majority of subjects (258) have fasting blood sugar below the threshold, while 45 were above.
    - This feature may provide insights into the metabolic health of the subjects.
- **Resting Electrocardiographic (restecg):**
    - Subjects are almost evenly split between types '1' (152) and '0' (147), with a small number (4) having type '2'.
    - This feature could be indicative of underlying heart conditions.
- **Slope(slp):**
    - Types '2' (142) and '1' (140) are almost equally represented, while type '0' was less common (21).
    - The slope might correlate with heart health.
- **Major Vessels (caa):**
    - Most subjects (175) have '0' major vessels colored by fluoroscopy, with decreasing counts for '1' (65), '2' (38), '3' (20), and '4' (5).
    - This could relate to the severity of coronary artery disease.
- **Thalium Stress Test Result (thall):**
    - The most common type is '2' (166), followed by '3' (117), '1' (18), and '0' (2).
    - The presence and type of thall might be a significant factor in heart attack risk.
- **Exercise Induced Angina (exng):**
    - A significant portion of subjects (204) do not experience angina induced by exercise, while 99 does.
    - This feature might be important in understanding exercise-related cardiac events.



Category Feature Count Plots

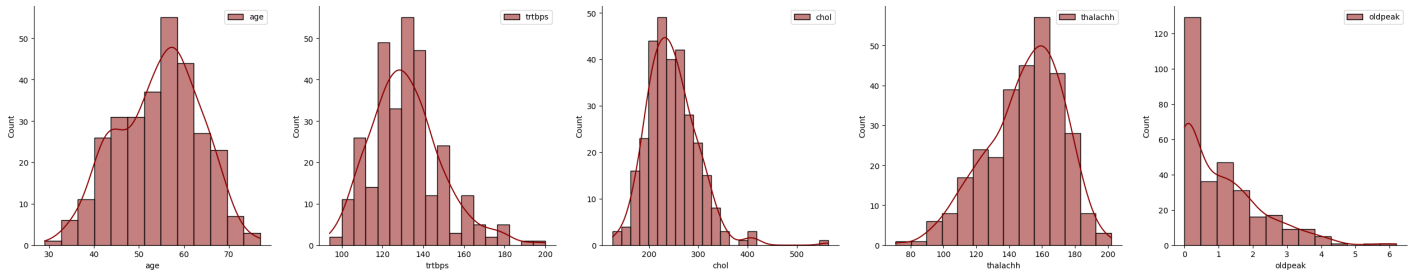
1.4 Numerical Features

The distribution of numerical features provides insights into the range and distribution patterns of key physiological and medical test data in the dataset. Following are some significant findings:

- **Age:**

- The age distribution appears somewhat normally distributed with a slight right skew, indicating a higher representation of middle-aged subjects.
- Most subjects fall in the age range of late 50s to 60s.

- **Resting Blood Pressure (trtbps):**
  - The distribution of resting blood pressure shows a normal pattern with a concentration around the 120-140 mmHg range.
  - This indicates a prevalence of normal to slightly elevated blood pressure among subjects.
- **Cholestoral (chol):**
  - The cholesterol levels are right-skewed, indicating that a significant portion of subjects have cholesterol levels in the higher range.
  - This skewness might be relevant in assessing the risk factors for heart attack.
- **Maximum Heart Rate Achieved (thalachh):**
  - The distribution of maximum heart rate shows a left skew, with most subjects achieving higher maximum heart rates.
  - This feature might correlate with physical fitness and cardiac function.
- **Previous Peak (oldpeak):**
  - ST depression shows a highly right-skewed distribution with a significant number of subjects exhibiting lower values.
  - The variability in this feature could be indicative of differing cardiac responses to exercise.



Numerical Feature Count Plots

2. Feature Engineering

2.1 Data Cleaning and Preparation

The initial step in the feature engineering involved removing duplicate entries from the dataset. This step is crucial to ensure the model is trained on unique data points, avoiding any bias that might occur due to repeated information. After the removal of index-164 duplicate, the dataset includes 302 records, each with 13 attributes, and one target variable (`output`).

2.2 One-Hot Encoding of Categorical Features

To effectively handle categorical variables in the dataset, one-hot encoding is applied. This process transformed categorical features into a format that could be better interpreted by the models, converting them into binary columns. This transformation increases the number of attributes in the dataset from the original set to 22 attributes. One-hot encoding is an essential step in preprocessing as it allows the model to recognize categorical data without assuming a natural ordering in categories.

2.3 Normalization

Normalization of the data is performed using `StandardScaler`. This step is critical in bringing all features to the same scale, thereby ensuring that no single feature dominates the model due to its scale. This standardization helps in improving the model's performance by reducing the chances of biased or skewed outcomes due to variable scales in the data.

3. Model Training

The dataset was subsequently divided into training and testing sets with a test size of 20%, following the common practice to evaluate the model's performance on unseen data. The training set consisted of 241 samples, and the test set had 61 samples.

3.1 Classification Based Model

A variety of classification models are trained and evaluated to predict the likelihood of heart attacks. The models including:

- **Logistic Regression**: A statistical model that predicts the probability of a binary outcome (like heart attack risk) based on one or more predictor variables. It's widely used for binary classification tasks.
- **Support Vector Machine (SVM)**: SVM is a powerful classifier that works by finding the best hyperplane to separate different classes in the feature space. It's effective in high-dimensional spaces.
- **K-Nearest Neighbors (kNN)**: A non-parametric method that classifies a data point based on how its neighbors are classified. It's simple and effective, particularly for datasets where similar cases are close to each other.
- **Gaussian Naive Bayes**: Based on Bayes' theorem, this model assumes independence among predictors and is particularly suited for classification tasks with features following a normal distribution.
- **Decision Tree**: A tree-like model that splits the data based on certain conditions. It's intuitive and easy to interpret, making it popular for classification tasks.
- **Random Forest**: An ensemble of decision trees, typically trained with the "bagging" method. It's known for its robustness and ability to avoid overfitting.
- **X Gradient Boosting (XGBoost)**: An advanced implementation of gradient boosting algorithms, known for its speed and performance. It builds trees in a sequential manner, where each tree corrects the errors of the previous ones.

Model Optimization

To fine-tune kNN and Random Forest, hyperparameter optimization was conducted. The optimal number of neighbors for KNN is found to be 6, resulting in an accuracy of 90.16% and for Random Forest, the optimal number of estimators was 13 with an accuracy of 100%.

Cross-Validation Scores

To assess model stability and generalization, 10-fold cross-validation was applied.

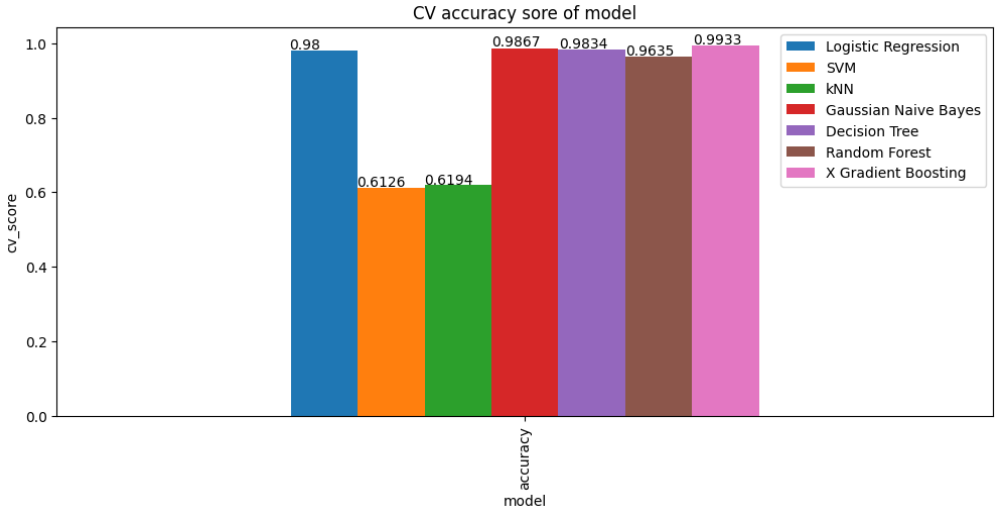
Performance Evaluation

The high accuracy scores, as show in the following table, particularly for Logistic Regression, SVM, Gaussian Naive Bayes, Random Forest, and XGBoost, indicate strong predictive capabilities in these models. However, the notable disparity between test accuracy and cross-validation scores for SVM and KNN suggests potential overfitting or model-specific sensitivity to the training data.

The cross-validation scores, as show in the following plot, provide a more realistic assessment of model performance, highlighting the robustness of Logistic Regression (98.0%), Gaussian Naive Bayes (98.67%), Decision Tree (98.34%), Random Forest (96.35%), and X Gradient Boosting (99.33%) in this task.

Model	Validation Accuracy
Logistic Regression	100.0 %
SVM	100.0 %
KNN	90.16 %
Gaussian Naive Bayes	100.0 %
Decision Tree	98.36%
Random Forest	100.0 %
X Gradient Boosting	100.0 %

Model Validation Accuracy



Model Cross-Validation Scores

3.2 Clustering Based Model

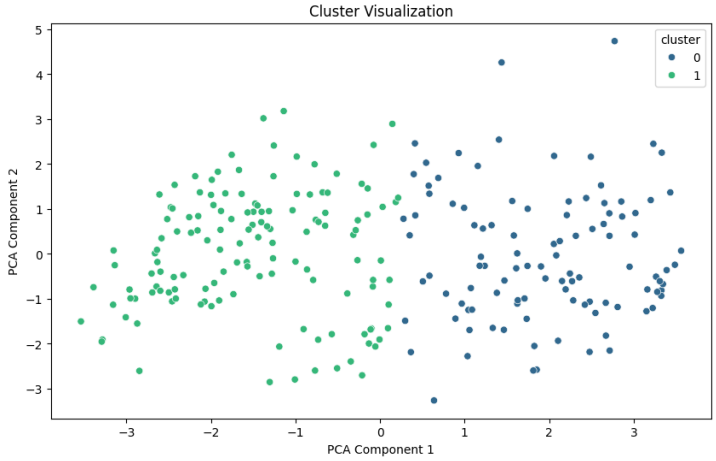
In addition to traditional classification models, a clustering-based approach was also employed to predict heart attack risk. This method involved the use of KMeans clustering, which is generally used for unsupervised learning but can be adapted for predictive tasks.

Clustering

KMeans clustering is performed on the transformed training data with an optimal number of 2 clusters.

This step aimed to group the data into two distinct clusters, potentially correlating with the heart attack risk levels.

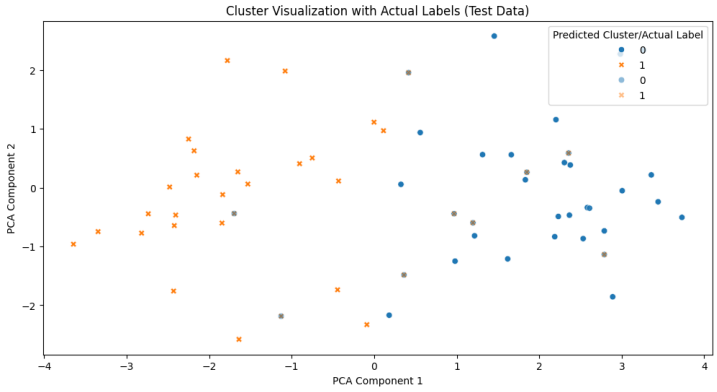
The clusters were visualized in the two-dimensional PCA space, both for the training and test data, providing an intuitive understanding of the data distribution and cluster separation, as shown in the following figures.



Dataset Clustering Visualization in 2 Clusters

Predictive Accuracy

The clusters were then used to predict heart attack risk in the test set, achieving an accuracy of approximately 85.24%. This accuracy level indicates that the clusters formed were relatively indicative of the heart attack risk, the result is shown in the following figure.



Dataset Clustering Visualization in 2 Clusters

The clustering-based method's performance, with over 85% accuracy, suggests that the dataset contains distinct groups that are relevant to predicting heart attack risk. This method's success demonstrates the potential of using unsupervised learning techniques in a semi-supervised manner for predictive tasks. The visualization of clusters also provided useful insights into the nature of the data and its inherent groupings.

4. Association Rule Mining

ARM Approach in Heart Attack Analysis

Association Rule Mining, a key technique in data mining, is applied to the Heart Attack Analysis and Prediction dataset to uncover relationships and patterns that might be indicative of heart attack risk.

The Apriori algorithm is used in this task.

Data Transformation for ARM

A transformation process is performed on the dataset to suit the needs of ARM:

- Continuous variables like 'age', 'trtbps', 'chol', 'thalachh', and 'oldpeak' are discretized into three range bins each, facilitating the identification of association rules.
- Categorical variables and the target variable ('output') are labeled in a binary format, e.g., 'sex-0', 'sex-1', to clearly distinguish different categories.

Rule Mining and Key Findings

With min_supp=0.25 and min_conf=0.7, the ARM process focuses on finding rules that have a strong association with the occurrence of heart attacks ('output-1'). The results indicates several interesting associations:

R-1. **Previous Peak, Exercise Induced Angina and Heart Attack Risk:**

- Rule: oldpeak(-0.0062, 2.067], exng-0 $\Rightarrow$ output-1
- Support=0.45, Confidence=0.74
- A high likelihood of heart attack risk with certain range of Previous Peak and absence of exercise-induced angina.

R-2. **Number of Major Vessels and Heart Attack Risk:**

- Rule: caa-0 $\Rightarrow$ output-1,
- Support=0.43, Confidence=0.74
- The absence of major vessels is strongly associated with heart attack risk.

R-3. **Thalassemia Type and Heart Attack Risk:**

- Rule: thall-2 $\Rightarrow$ output-1,
- Support=0.43, Confidence=0.78
- The type 2 of Thalium Stress Test result ('thall-2') shows a significant association with heart attack risk.

R-4. **Previous Peak, Thalassemia Type and Heart Attack Risk:**

- Rule: oldpeak(-0.0062, 2.067], thall-2 $\Rightarrow$ output-1
- Support=0.42, Confidence=0.82
- A high likelihood of heart attack risk with certain range of Previous Peak and type 2 of Thalium Stress Test result.

R-5. **Thalassemia Type, Exercise Induced Angina and Heart Attack Risk:**

- Rule: thall-2, exng-0 $\Rightarrow$ output-1,
- Support = 0.38, Confidence = 0.84

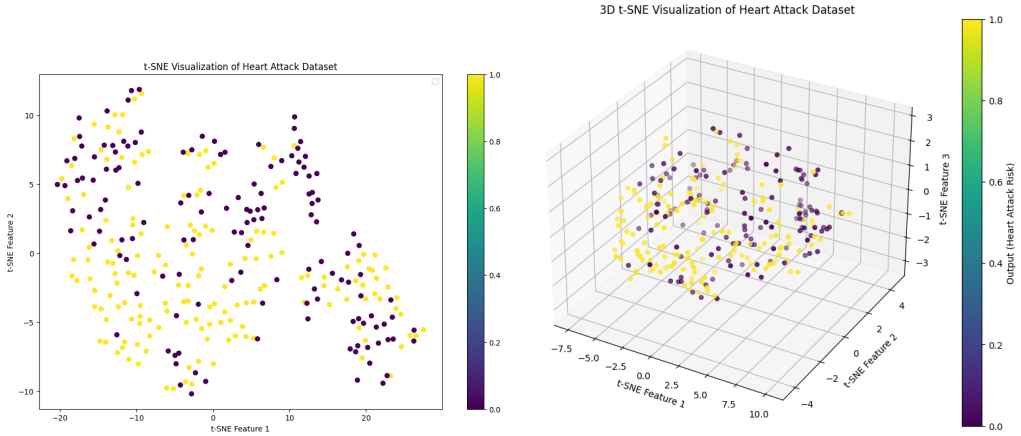
- The type 2 of Thalium Stress Test result ('thall-2') and absence of exercise-induced angina shows a significant association with heart attack risk.

Findings of ARM in Heart Attack Prediction

The rules mined offer valuable insights into the factors that might be associated with an increased risk of heart attacks. These associations provide a deeper understanding of the interplay between various clinical and physiological features and heart attack risk. They also highlight potential areas for further investigation and might aid in developing targeted strategies for heart attack prevention and treatment.

These findings are going to be used in the further formulation in the next section.

5. Further Formulation



Left: 2D t-SNE Visualization of Dataset; Right: 3D t-SNE Visualization of Dataset

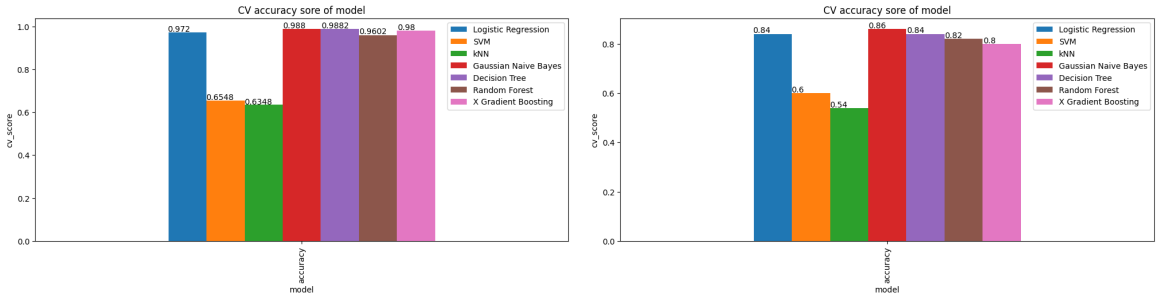
By observing the t-SNE visualizations of the dataset, I believe that some groups or clusters of the dataset may have more significant association with heart attack risk than the others.

Therefore, based on the findings of the EDA and ARM analysis, the following formulation is proposed to further analyze the dataset and predict heart attack risk:

- Some attributes including `oldpeak`, `caa`, `sex`, `age`, `chol` and `exng` are notably correlated with the heart attack risk. These attributes are selected and are used to split the dataset into different groups based on their values. Then perform the classification-based model on each group to see if the model is more accurate on the one group or the other group.

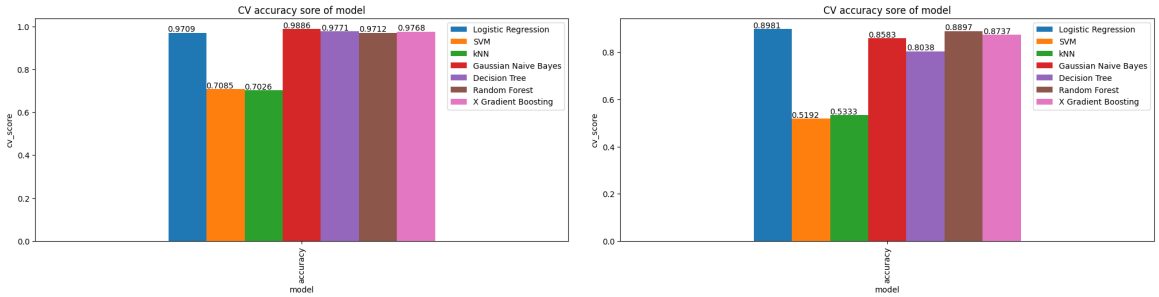
The major findings are as follows:

1. Group with **oldpeak** in range [-0.0062, 2.067] has a **higher accuracy** than the group with **oldpeak** not in range [-0.0062, 2.067] on predicting heart attack risk, with a accuracy improvement in **9%~18%**.
    - This finding can be supported by the ARM rule R-1 and R-4.



Left: CV Score on oldpeak in [-0.0062, 2.067] group; Right: CV Score on oldpeak NOT in [-0.0062, 2.067] group

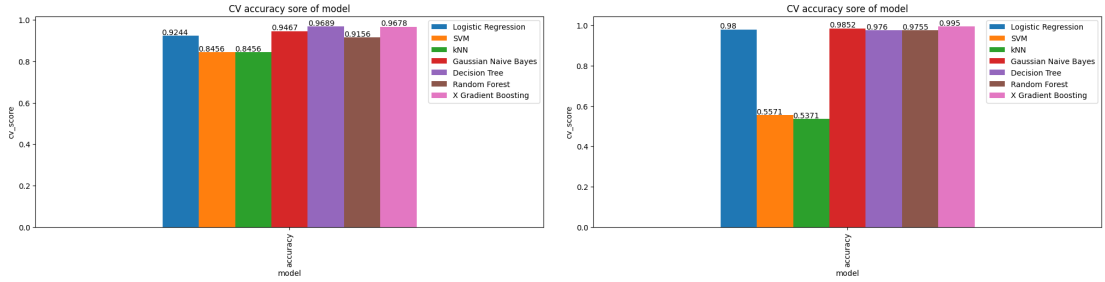
2. Group with **caa** = 0 has a **higher accuracy** than the group with **caa** = 1, 2, 3 on predicting heart attack risk, with a accuracy improvement in **8%~17%**.
    - This finding can be supported by the ARM rule R-2.



Left: CV Score on caa = 0 group; Right: CV Score on caa != 0 group

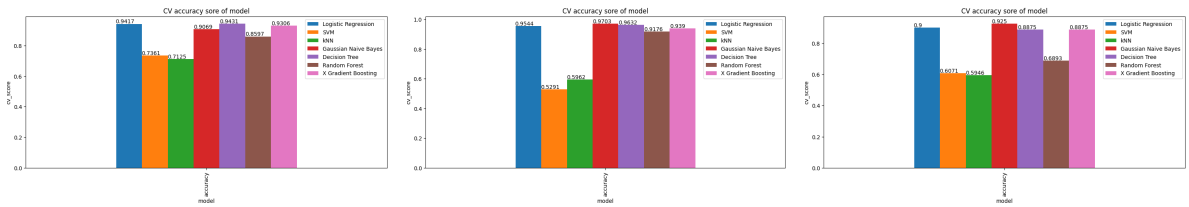
The following are some other findings that are not as significant as the above two findings:

Group with sex=1 has a **higher accuracy** than the group with sex=0 on predicting heart attack risk.



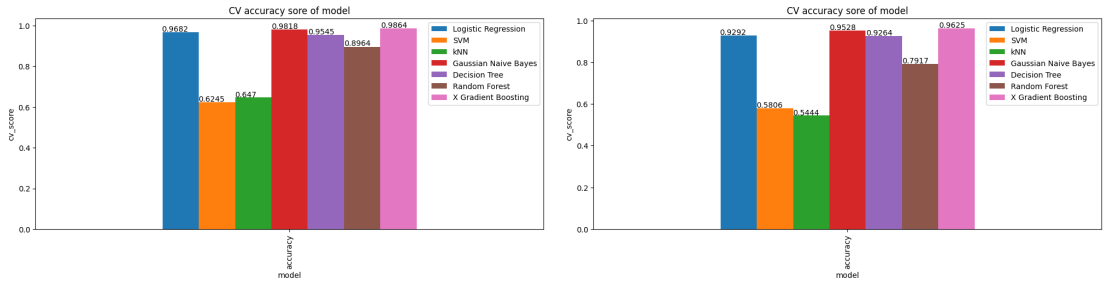
Left: CV Score on sex=0 group; Right: CV Score on sex=1 group

Group with age in [50, 60] has a **higher accuracy** than the group with age out of [50, 60] on predicting heart attack risk.



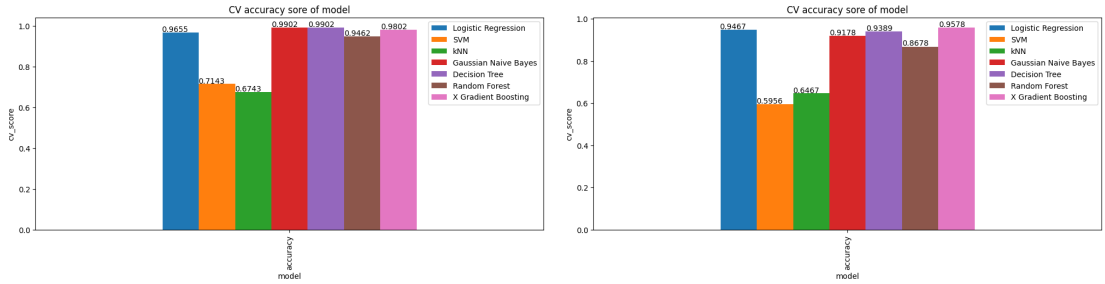
Left: CV Score on age < 50 group; Middle: CV Score on age 50-60 group; Right: CV Score on age > 60 group

Group with chol in [125.562, 272.0] has a **higher accuracy** than the group with chol out of [125.562, 272.0] on predicting heart attack risk.



Left: CV Score on chol in [125, 272] group; Right: CV Score on chol not in [125, 272] group

Group with exng = 0 has a **higher accuracy** than the group with exng != 0 on predicting heart attack risk.



Left: CV Score on exng = 0 group; Right: CV Score on exng = 1 group

6. Conclusion

In this project, on the Heart Attack Analysis and Prediction Dataset, I have analyzed the dataset, performed feature engineering, trained and evaluated classification-based and clustering-based models, which achieved a highest cross-validation accuracy of **99.33%** with X Gradient Boosting model.

I also performed Association Rule Mining on the dataset, which uncovered several interesting associations between features and heart attack risk. For example, a high chance of heart attack risk with `oldpeak` Previous Peak in range of [-0.0062, 2.067] and absence of exercise-induced angina. These findings are valuable in further formulating the dataset and predicting heart attack risk.

Finally, I proposed a further formulation based on the findings of the EDA and ARM analysis, which is to split the dataset into different groups based on the values of some important attributes, and then perform the classification-based model on each group to observe if the model is more accurate on the one group than the other group. The results showed that the model is more accurate on the group with `oldpeak` (Previous Peak) in range [-0.0062, 2.067] and the group with `caa` (Number of Major Vessels) = 0, which are consistent with the findings of ARM analysis.

However, limited by the number of samples in the dataset, the findings of this project may not be conclusive and need to be further verified on a larger dataset.

References

[1] "Association Rules with Python," *kaggle.com*. https://www.kaggle.com/code/mervetorkan/association-rules-with-python

[2] "heart-attack-analysis python," *kaggle.com*. https://www.kaggle.com/code/licgsg/heart-attack-analysis-python.

[3] "HeartAttack prediction with 91.8 % Accuracy," *kaggle.com*. https://www.kaggle.com/code/fahadmehfoooz/heartattack-prediction-with-91-8-accuracy.