COMP 4433 Data Mining and Data Warehousing

Individual project report

Lai Ka Chung 22080062d

Code: Kaggle link

Contents

1. Data Exploration

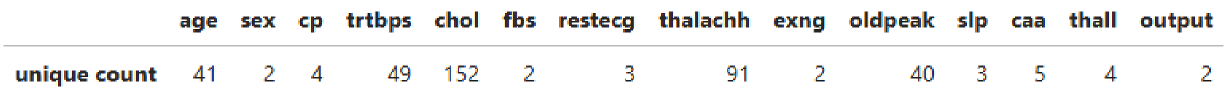
1.1 Dataset Overview [1]

The Heart Attack Analysis and Prediction Dataset contains 303 records, 13 features, and one target variable (output). A detailed description of each attribute:

- age - Age of the patient
- sex - Sex of the patient
    - (1 = male, 0 is female)
- 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
- caa - Number of major vessels
- thall - Thalium Stress Test result ~ (0,1,2, 3)
- exng - Exercise induced angina
    - (1 = Yes, 0 = No)
- output - Target variable
    - (1 = Yes, 0 = No)

1.2 Data separation

By checking the unique count of each column:



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
unique count	41	2	4	49	152	2	3	91	2	40	3	5	4	2

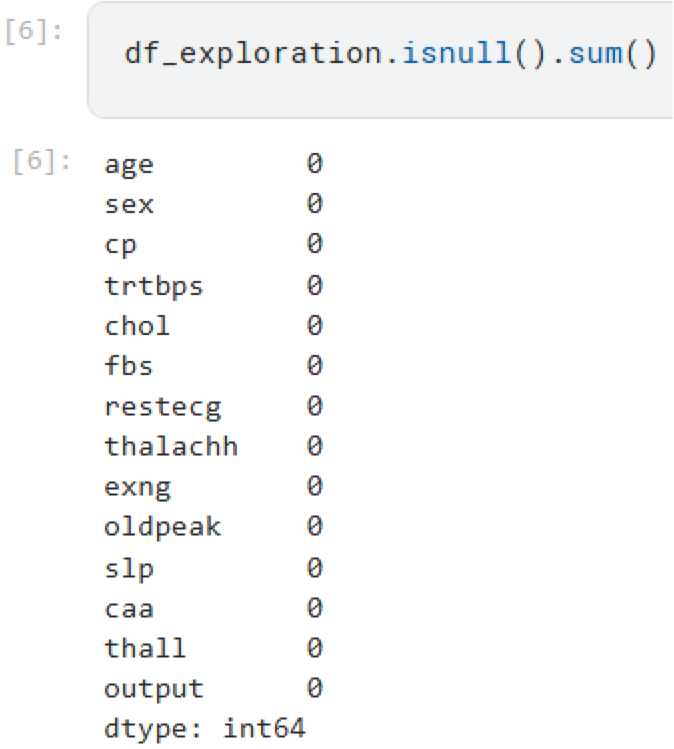
We can find two types of columns: categorial columns and continuous columns.

Categorial column: sex, exng, caa, cp, fbs, restecg, slp, thall

Continuous column: age, trtbps, chol, thalachh, oldpeak

1.3 Missing Data:

There is no null value in this dataset, which means there are no missing data

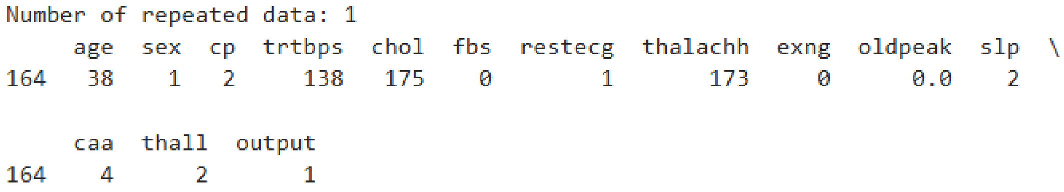


```
[6]:  df_exploration.isnull().sum()
```

```
[6]:  age         0
      sex         0
      cp          0
      trtbps      0
      chol        0
      fbs         0
      restecg     0
      thalachh    0
      exng        0
      oldpeak     0
      slp         0
      caa         0
      thall       0
      output      0
      dtype: int64
```

1.4 Duplicate Data:

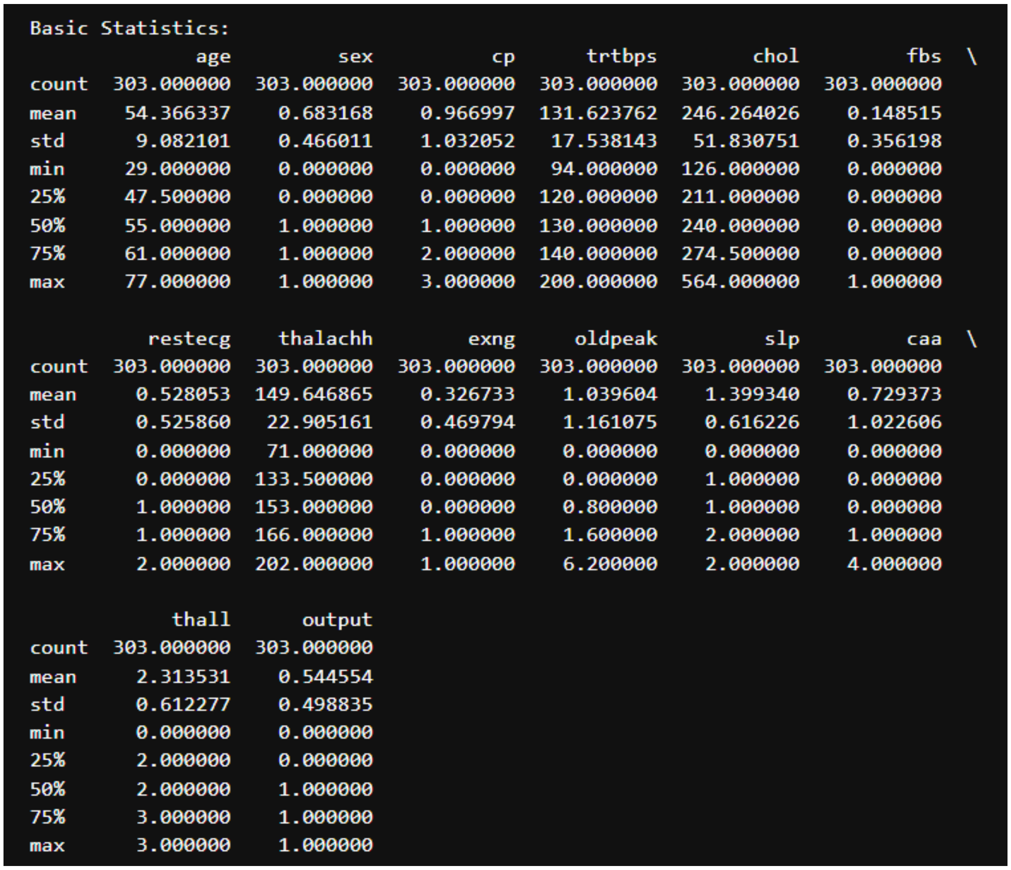
There is one duplicate record, which is index 164. The duplicate data will be deleted to provide a better understanding of the data.



```
Number of repeated data: 1
     age  sex  cp  trtbps  chol  fbs  restecg  thalachh  exng  oldpeak  slp  \
164   38    1   2     138   175    0        1       173     0      0.0    2   

     caa  thall  output  
164    4      2       1  
```

1.5 Basic Statistics of all the features

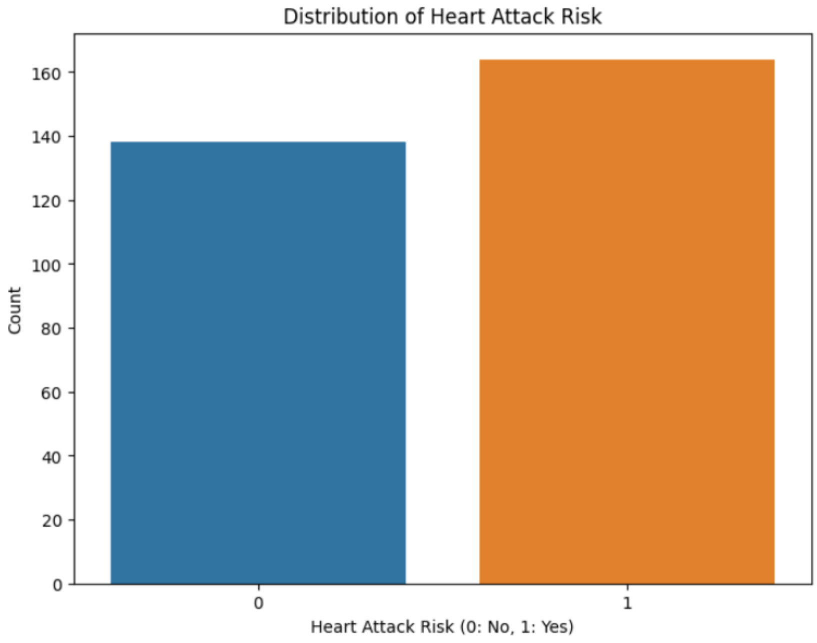
The following figure shows the basic statistics of all the features: count, mean, standard deviation, minimum, Q1, Q2, Q3 and the maximum.



```
Basic Statistics:
              age         sex          cp      trtbps        chol         fbs  \
count  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000
mean    54.366337    0.683168    0.966997  131.623762  246.264026    0.148515
std      9.082101    0.466011    1.032052   17.538143   51.830751    0.356198
min     29.000000    0.000000    0.000000   94.000000  126.000000    0.000000
25%     47.500000    0.000000    0.000000  120.000000  211.000000    0.000000
50%     55.000000    1.000000    1.000000  130.000000  240.000000    0.000000
75%     61.000000    1.000000    2.000000  140.000000  274.500000    0.000000
max     77.000000    1.000000    3.000000  200.000000  564.000000    1.000000

          restecg    thalachh        exng     oldpeak         slp         caa  \
count  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000
mean     0.528053  149.646865    0.326733    1.039604    1.399340    0.729373
std      0.525860   22.905161    0.469794    1.161075    0.616226    1.022606
min      0.000000   71.000000    0.000000    0.000000    0.000000    0.000000
25%      0.000000  133.500000    0.000000    0.000000    1.000000    0.000000
50%      1.000000  153.000000    0.000000    0.800000    1.000000    0.000000
75%      1.000000  166.000000    1.000000    1.600000    2.000000    1.000000
max      2.000000  202.000000    1.000000    6.200000    2.000000    4.000000

            thall      output
count  303.000000  303.000000
mean     2.313531    0.544554
std      0.612277    0.498835
min      0.000000    0.000000
25%      2.000000    0.000000
50%      2.000000    1.000000
75%      3.000000    1.000000
max      3.000000    1.000000
```

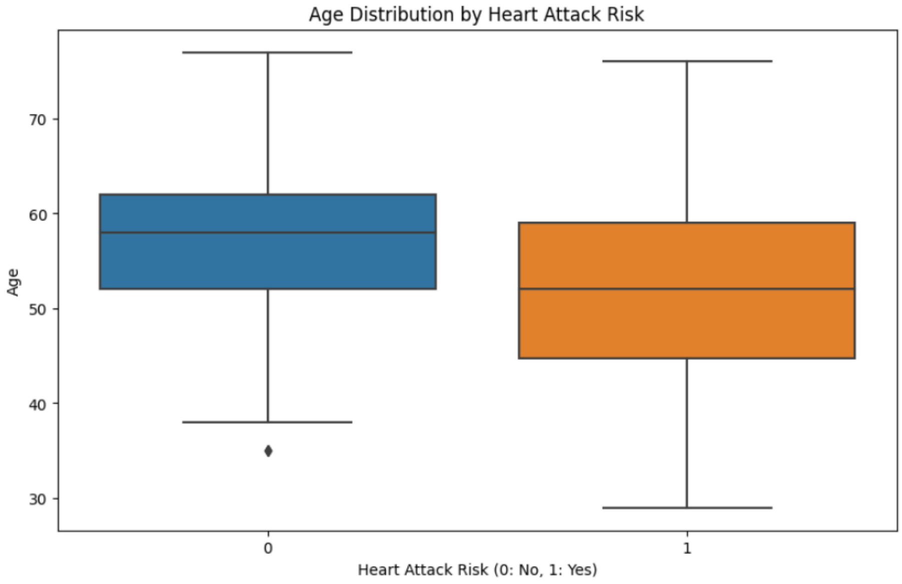
1.6 Distribution of Heart Attack Risk

The distribution of the target variable is quite balanced, as shown in the figure.



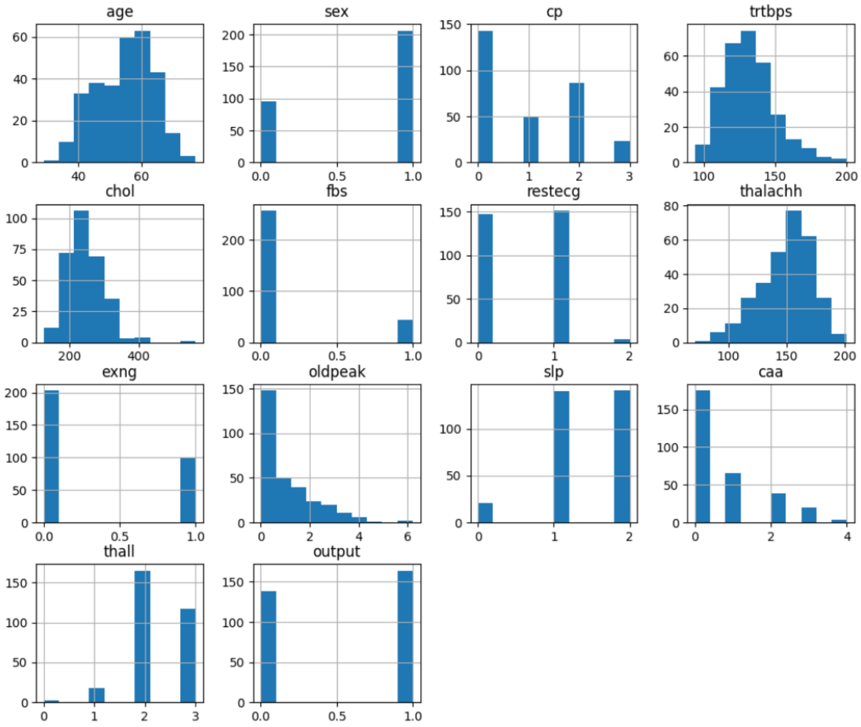
1.7 Age Distribution by Heart Attack Risk

The following figure shows that the age variance of having heart attack risk is more significant than people who don't have heart attack risk.

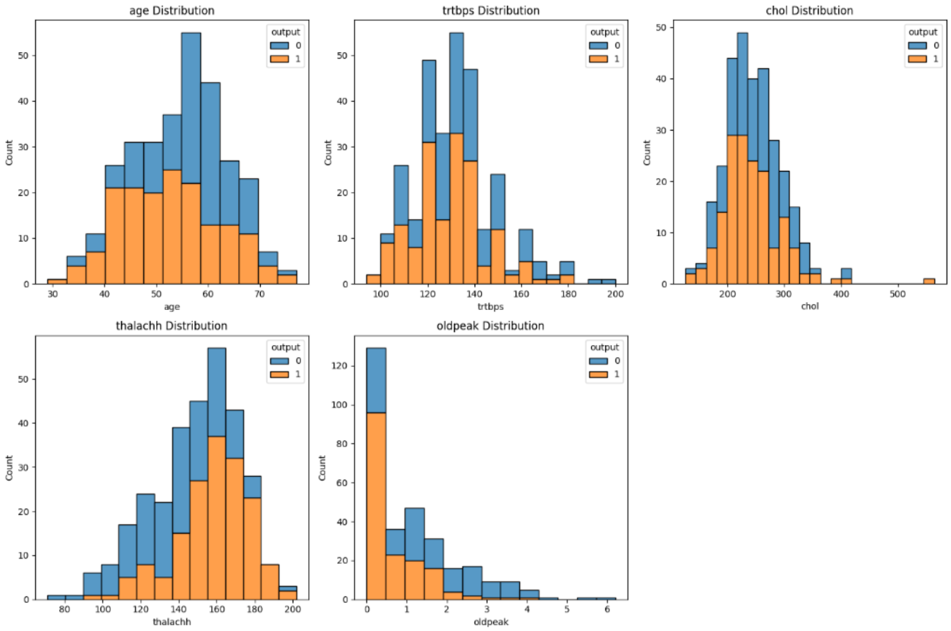


1.8 Feature Distributions

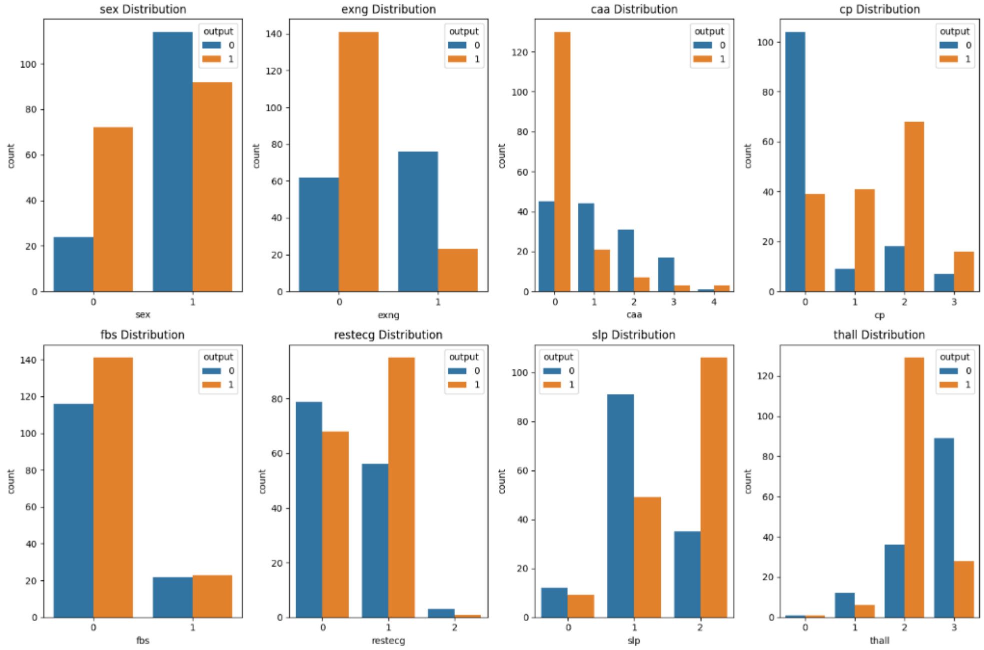
The following figure shows the overall distribution of the feature and distribution of the output.



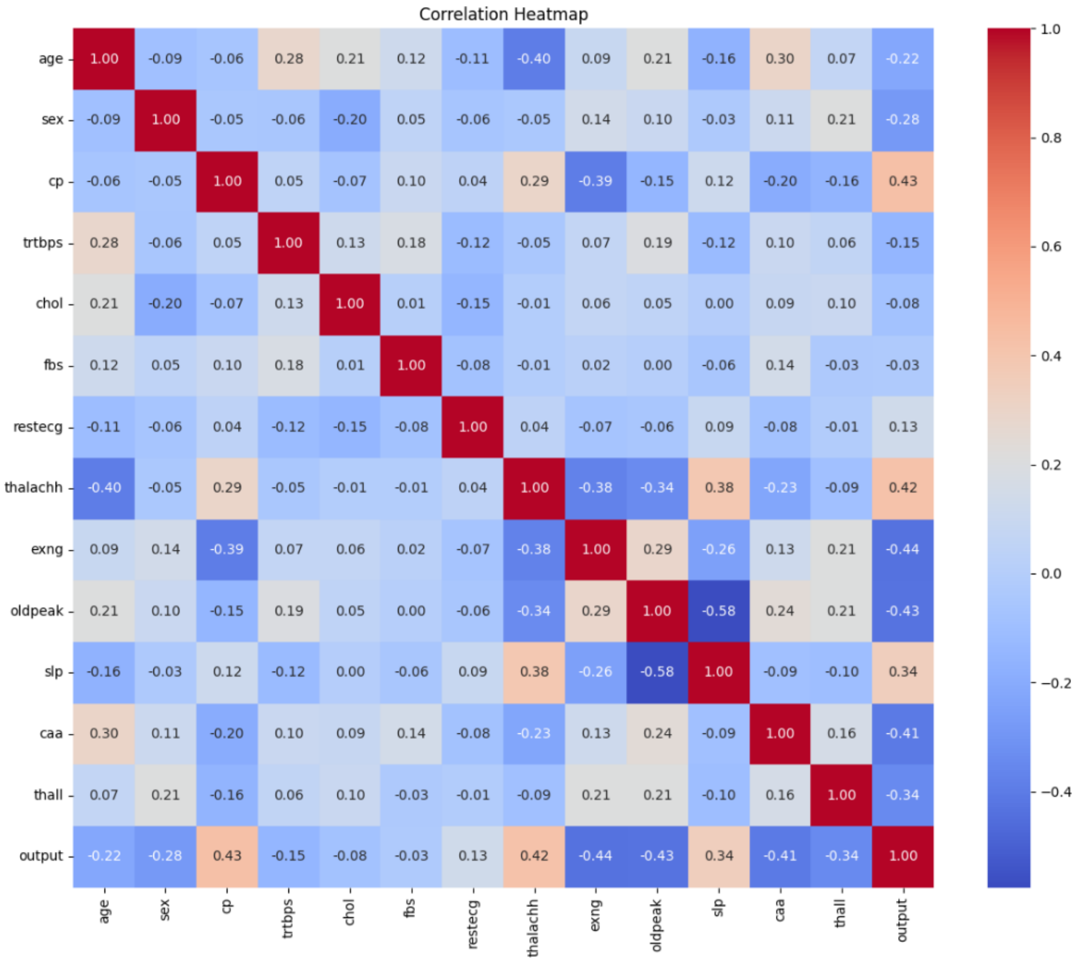
The following figure shows the distribution of continuous columns, which shows most of the features are slightly right-skewed or left-skewed, but oldpeak's distribution is Heavily right-skewed.



The following figure shows the output difference between different medical measurements and conditions, which shows that having 0 exng,0 caa, 2 slp, and 2 thall has a higher rate of heart attack risk.



1.9 Data Correlation map



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.00	-0.09	-0.06	0.28	0.21	0.12	-0.11	-0.40	0.09	0.21	-0.16	0.30	0.07	-0.22
sex	-0.09	1.00	-0.05	-0.06	-0.20	0.05	-0.06	-0.05	0.14	0.10	-0.03	0.11	0.21	-0.28
cp	-0.06	-0.05	1.00	0.05	-0.07	0.10	0.04	0.29	-0.39	-0.15	0.12	-0.20	-0.16	0.43
trtbps	0.28	-0.06	0.05	1.00	0.13	0.18	-0.12	-0.05	0.07	0.19	-0.12	0.10	0.06	-0.15
chol	0.21	-0.20	-0.07	0.13	1.00	0.01	-0.15	-0.01	0.06	0.05	0.00	0.09	0.10	-0.08
fbs	0.12	0.05	0.10	0.18	0.01	1.00	-0.08	-0.01	0.02	0.00	-0.06	0.14	-0.03	-0.03
restecg	-0.11	-0.06	0.04	-0.12	-0.15	-0.08	1.00	0.04	-0.07	-0.06	0.09	-0.08	-0.01	0.13
thalachh	-0.40	-0.05	0.29	-0.05	-0.01	-0.01	0.04	1.00	-0.38	-0.34	0.38	-0.23	-0.09	0.42
exng	0.09	0.14	-0.39	0.07	0.06	0.02	-0.07	-0.38	1.00	0.29	-0.26	0.13	0.21	-0.44
oldpeak	0.21	0.10	-0.15	0.19	0.05	0.00	-0.06	-0.34	0.29	1.00	-0.58	0.24	0.21	-0.43
slp	-0.16	-0.03	0.12	-0.12	0.00	-0.06	0.09	0.38	-0.26	-0.58	1.00	-0.09	-0.10	0.34
caa	0.30	0.11	-0.20	0.10	0.09	0.14	-0.08	-0.23	0.13	0.24	-0.09	1.00	0.16	-0.41
thall	0.07	0.21	-0.16	0.06	0.10	-0.03	-0.01	-0.09	0.21	0.21	-0.10	0.16	1.00	-0.34
output	-0.22	-0.28	0.43	-0.15	-0.08	-0.03	0.13	0.42	-0.44	-0.43	0.34	-0.41	-0.34	1.00

- Positive Correlations with Output:
  - Chest Pain Type (Cp): Strong positive correlation with 0.43 means a high relationship between having the risk of a heart attack.
  - Maximum Heart Rate Achieved (thalachh): A strong positive correlation with 0.42 means a high relationship between having the risk of heart attack.
- Negative Correlations with Output:
  - Exercise induced angina (exng): Strong negative correlation with -0.44, which means a high relationship between not having the risk of a heart attack.
  - Previous peak (oldpeak): Strong negative correlation with -0.43, which means a high relationship between not having the risk of a heart attack.

- Number of major vessels (caa): Significant negative correlation with -0.41, which means a high relationship between not having the risk of a heart attack.

2. Data Preprocessing

2.1 Drop duplicated data

Since in the dataset, there is one duplicated record. First, training bias and duplicated records will cause the model to give more weight to repeated instances. Second, in an Overfitting problem, the model might memorise the repeated patterns instead of learning the actual relationship. Third, Skewed Validation, if duplicates appear in both training and test sets, model evaluation becomes unreliable. That's why the duplicates will be dropped.

2.2 Handle outliers by using the IQR method

We have noticed slightly skewed data, and extreme values are errors that may significantly impact the analysis. So, the values out of Q1 (25th percentile) and Q3 (75th percentile) will be considered outliers.

2.3 Feature Scaling

Using the function StandardScaler() to scale all the features to the same scale prevents larger-scale features from dominating, which helps the model learn from all features equally.

3. Feature Engineering

To enhance the model's ability to learn the pattern by adding context.

3.1 Adding a new feature

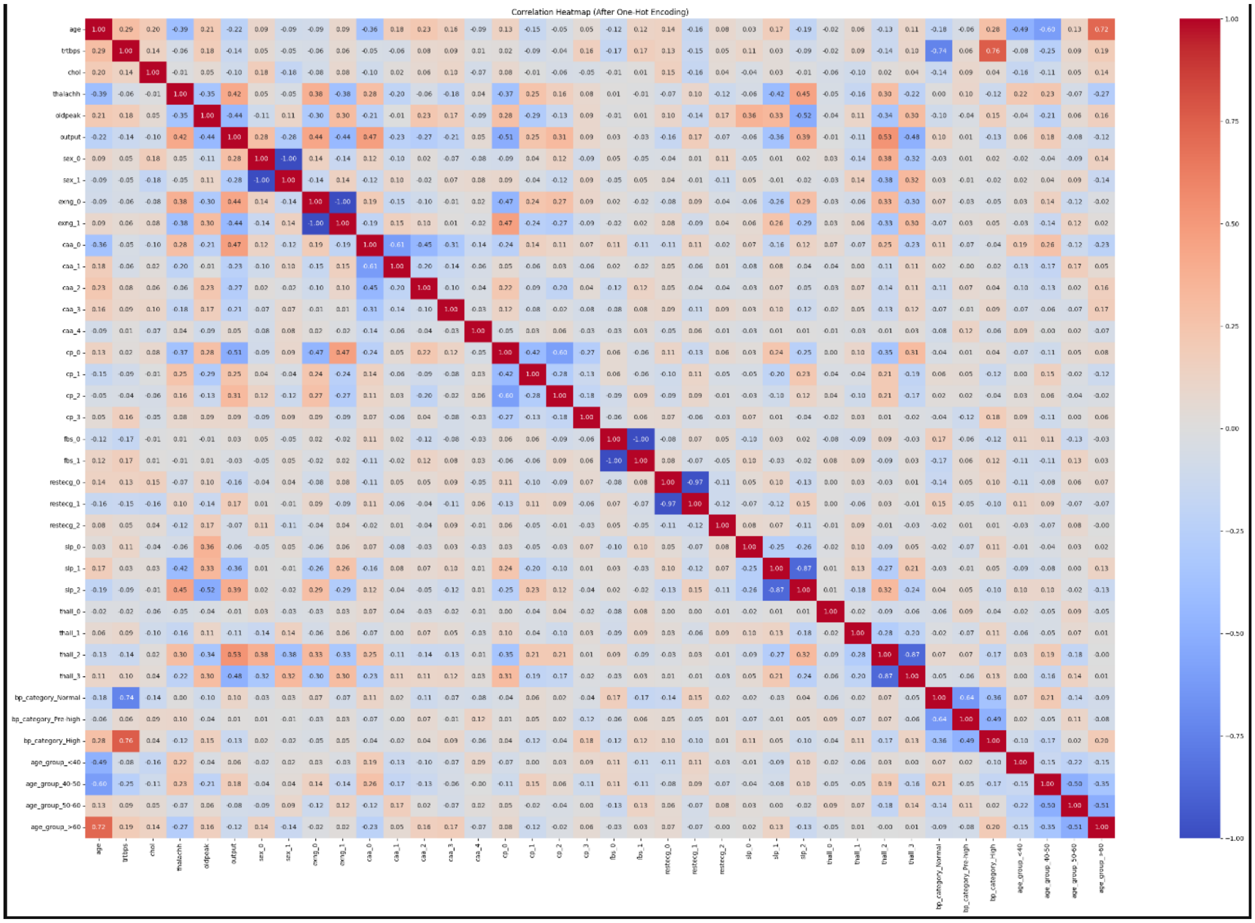
There are two features have been added to the dataset:

- age_group:
  Adding different age_groups with bins below 40, 40-50, 50-60, or above 60. To simplify the variable and reduce the noise of the data. Since the heart attack risk may not be a linear relationship between ages.
- bp_category (Resting blood pressure)
  Adding different categories with blood pressure, which is Normal (<120), Pre-high(120-140) and High(140-500). Adding more domain-specific Insights makes the feature and result more understandable.

3.2 One-hot encoding

Since machine learning requires numerical input. One-hot encoding converts category variables into binary format and avoids ordinal relationships from labelling encoding.

Here is the conclusion after adding the new feature:



The above figure shows the new correlation between new features and output. It shows that the 40-50 age group has a 0.18 relationship with having a high risk of a heart attack. Blood pressure high has a -0.13 relationship between output, which means having a lower risk of a heart attack.

4. Model Training

4.1 Performance Evaluation

The dataset has been divided into a training set (80%) and a test set(20%) to train the model. Then, using Grid Search Cross-Validation with 5-fold cross-validation, it optimizes for F1 score to test all the combinations of hyperparameters.

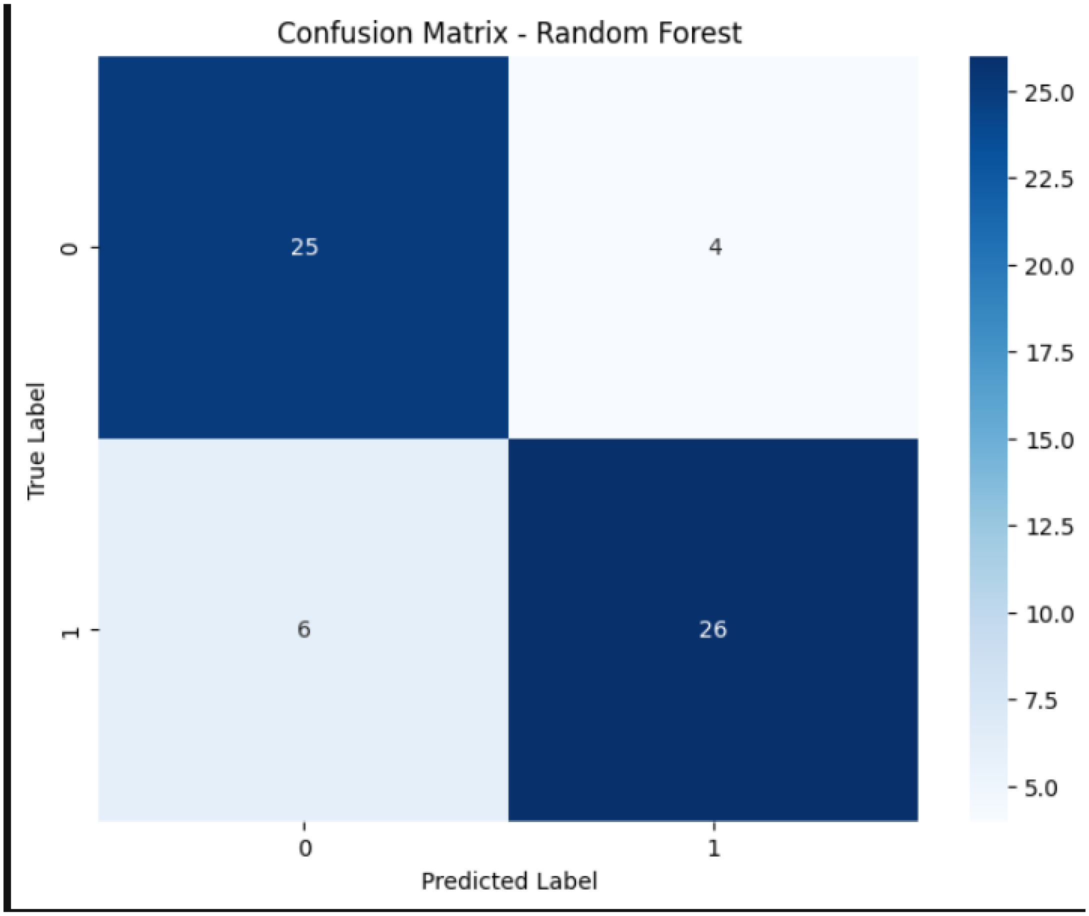
To compare the results from different models, all the best parameters for each model will be selected, and the Classification report (Precision, Recall, F1-score) and confusion matrix will be visualised to compare.

4.2 Classification model

All the model's random states have been set to 42 for reproducibility

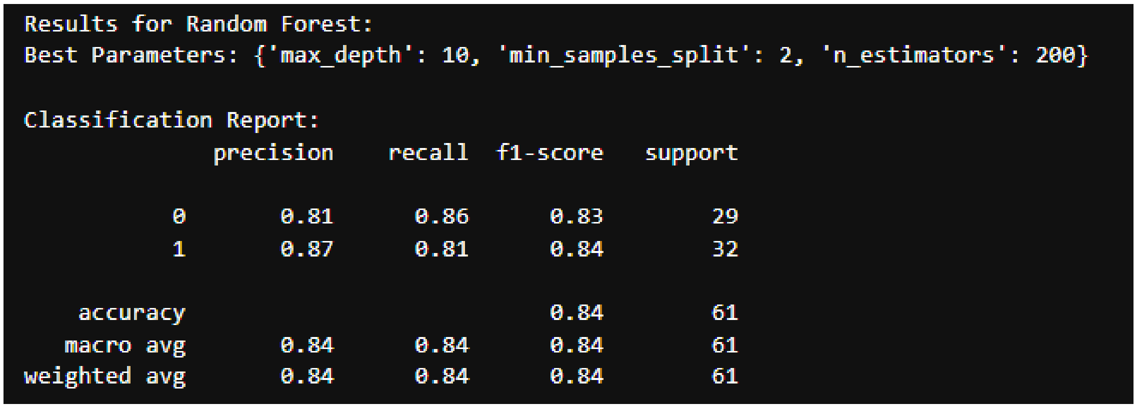
- Random Forest: Good for handling non-linear relationships and feature importance
- XGBoost: Powerful for structured/tabular data, handles imbalanced datasets well
- SVM: Effective in high-dimensional spaces, versatile through different kernel functions

1. Random Forest



The above diagram shows the true and predicted labels from the Random Forest model prediction.

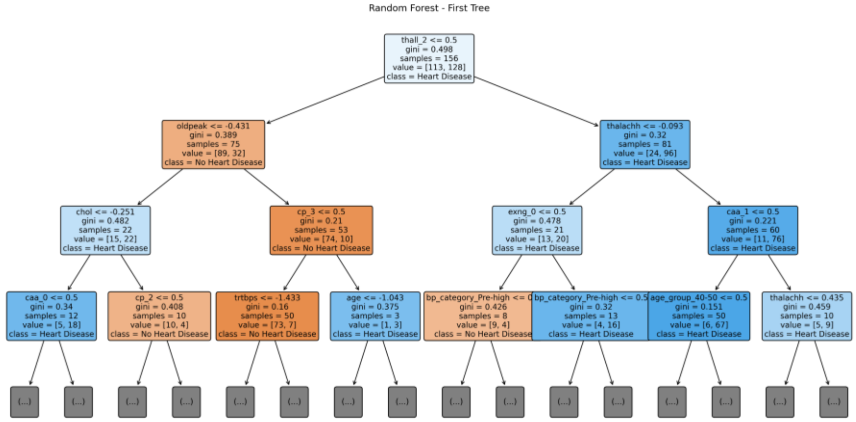
Here is the classification report:



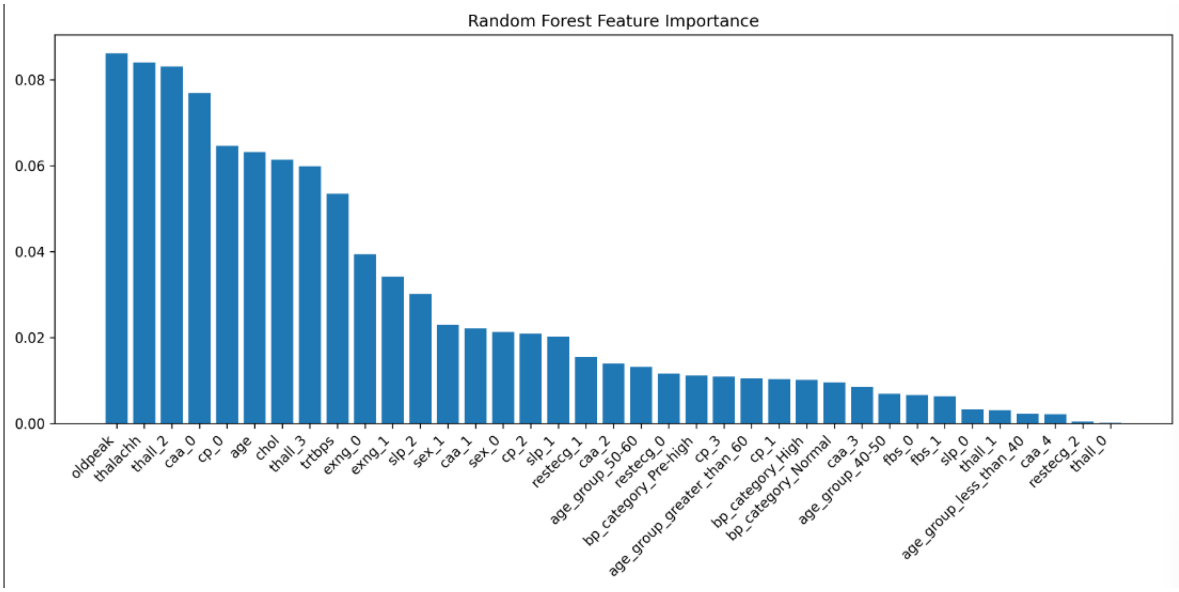
```
Results for Random Forest:
Best Parameters: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 200}

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.86      0.83        29
           1       0.87      0.81      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

The F1 score for heart attack risk is 0.84. The maximum depth of trees is 10, the Minimum sample required to split is 2, and the Number of trees is 200.

Here is the classification tree generated by the Random Forest:

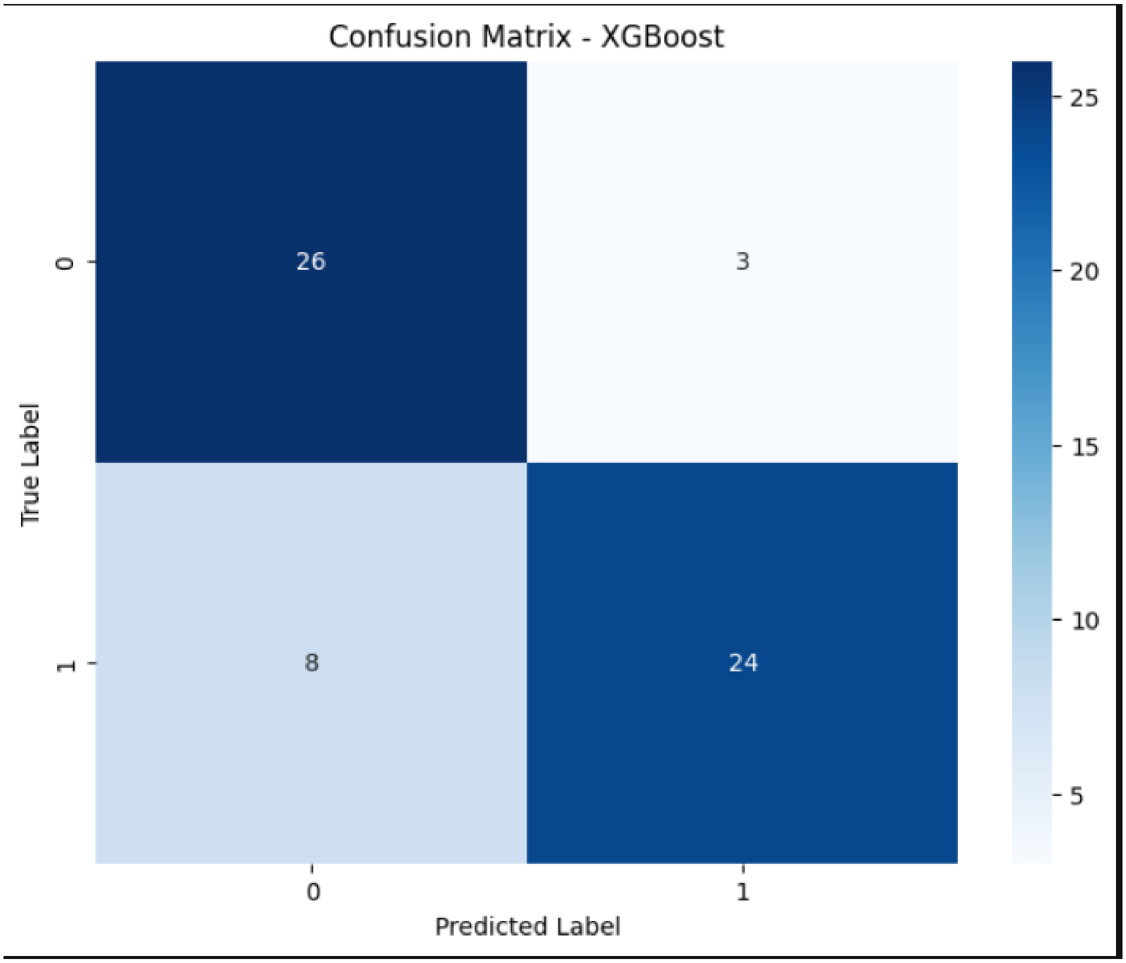


Here is the feature importance for better understanding:



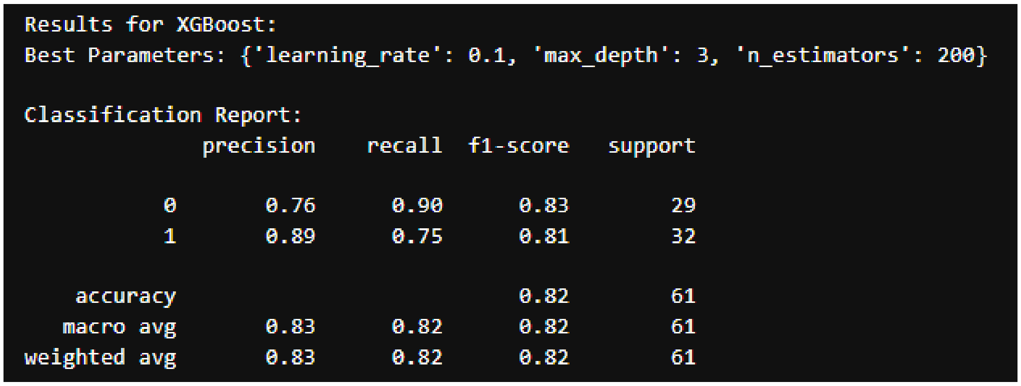
The above graph shows that the most important features in the random forest are the Previous peak (old peak) and the Maximum heart rate achieved(halacha).

2. XGBoost



The above diagram shows the true and predicted labels from the XGBoost model prediction.

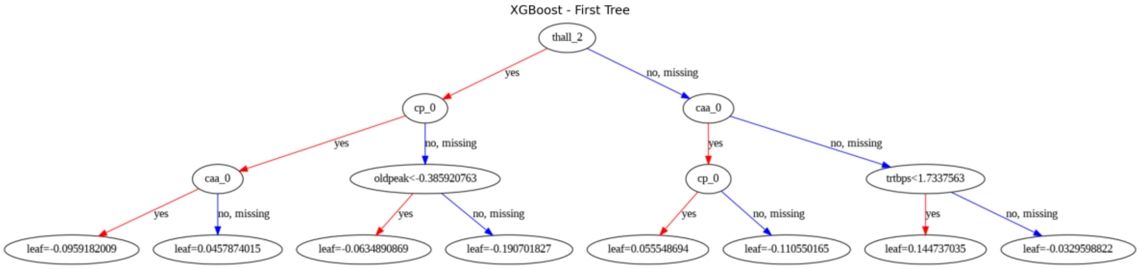
Here is the classification report:



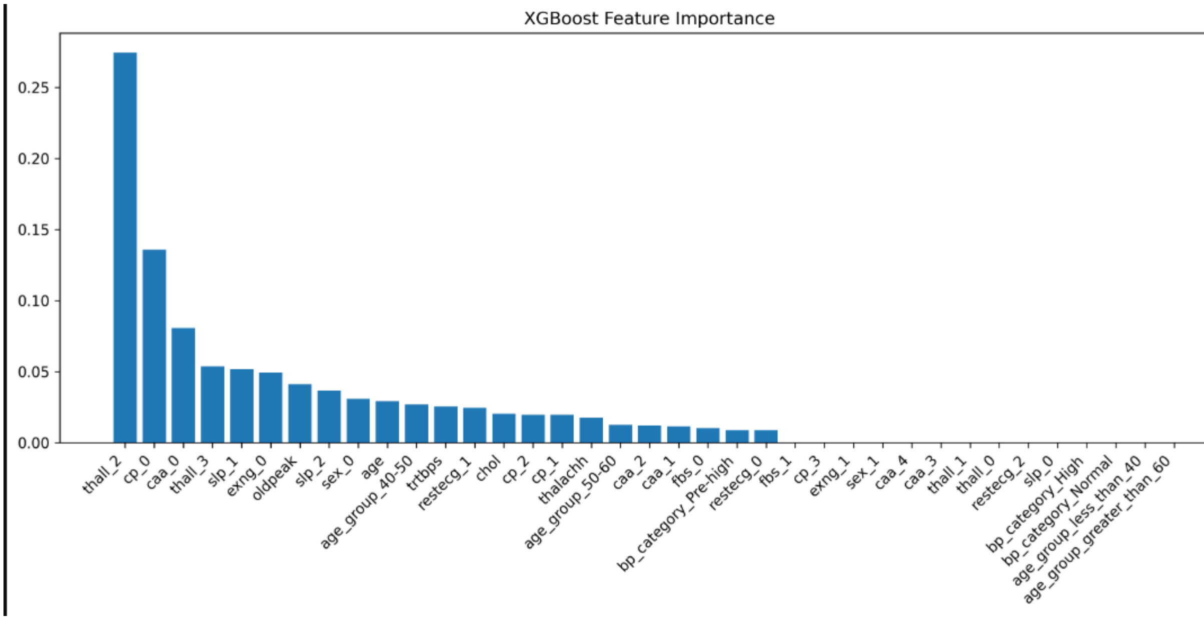
```
Results for XGBoost:
Best Parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}

Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.90      0.83        29
           1       0.89      0.75      0.81        32

    accuracy                           0.82        61
   macro avg       0.83      0.82      0.82        61
weighted avg       0.83      0.82      0.82        61
```

The F1 score of having heart attack risk is 0.82 with a maximum depth of trees of 3, Step size shrinkage is 0.1 and Number of boosting rounds is 200.

Here is the classification tree generated by the XGBoost:

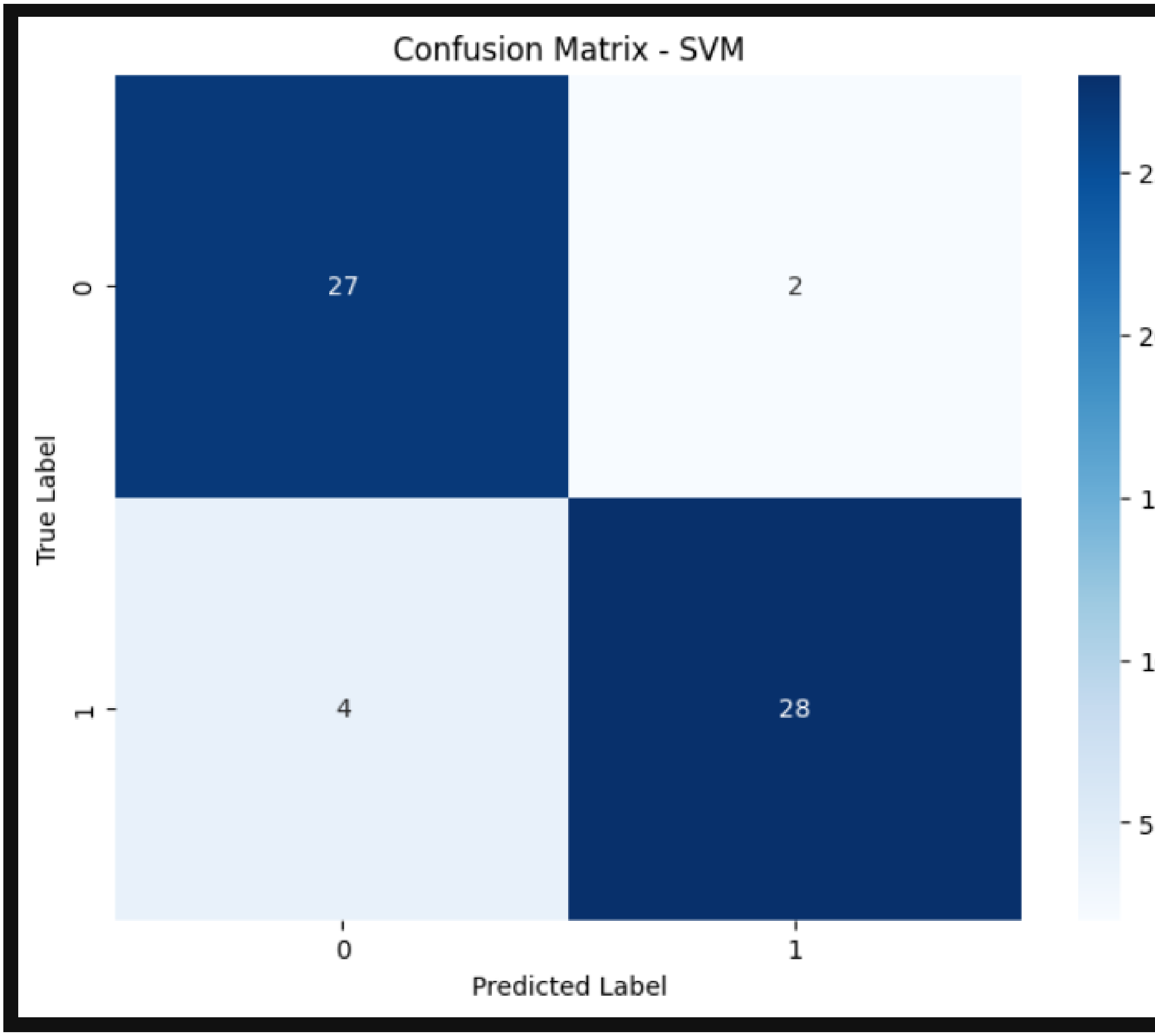


Here is the feature importance for better understanding:



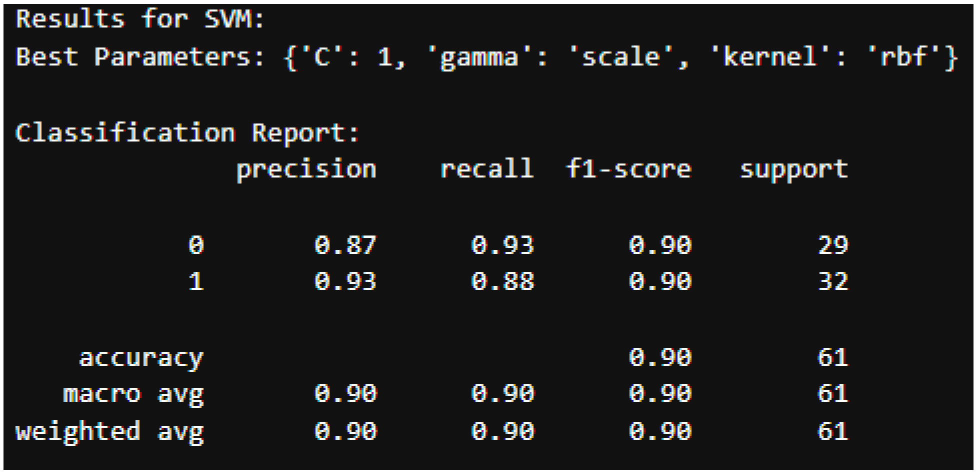
The above graph shows that the most crucial features in XGBoost are the Thalium Stress Test result (thall), which is 2, and the Chest pain type (cp), which is also 2.

3. SVM



The above diagram shows the true and predicted labels from the SVM model prediction.

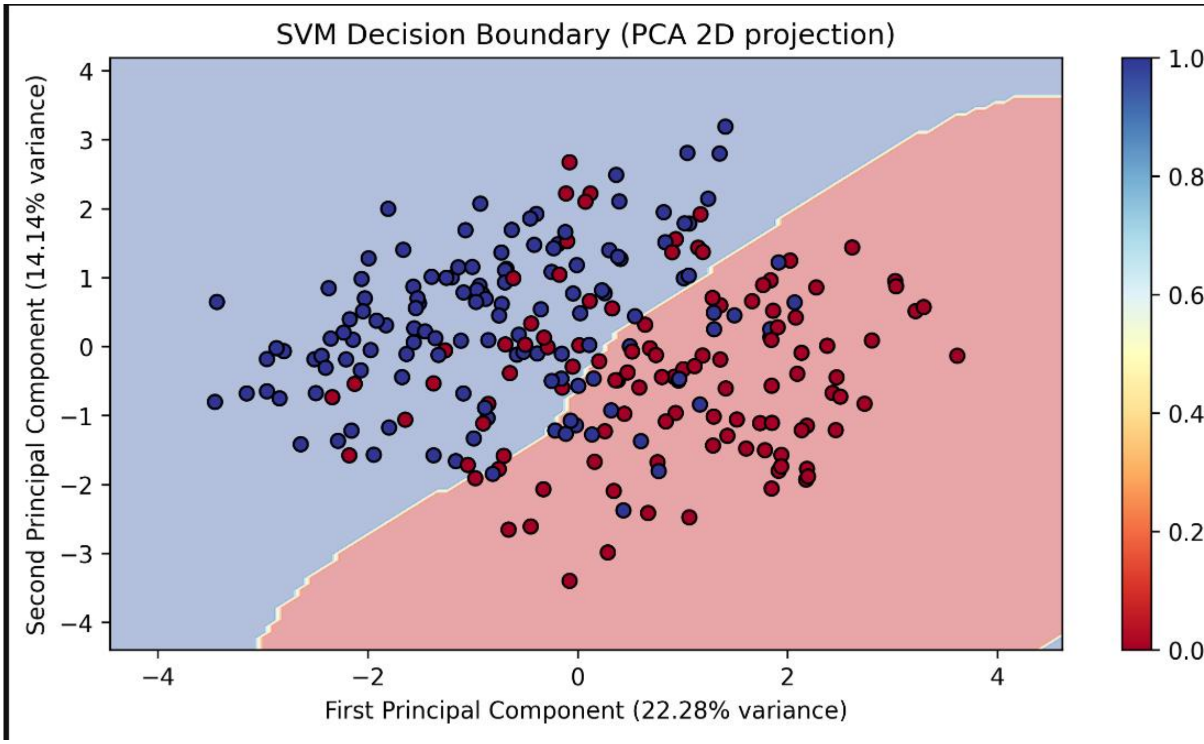
Here is the classification report:



```
Results for SVM:
Best Parameters: {'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.93      0.90        29
           1       0.93      0.88      0.90        32

    accuracy                           0.90        61
   macro avg       0.90      0.90      0.90        61
weighted avg       0.90      0.90      0.90        61
```

The F1 score for heart attack risk is 0.90, with a Regularization parameter of 0.1, Kernel type rbf, and Kernel coefficient scale.

Here is the graph with the SVM decision boundary with 2D only.



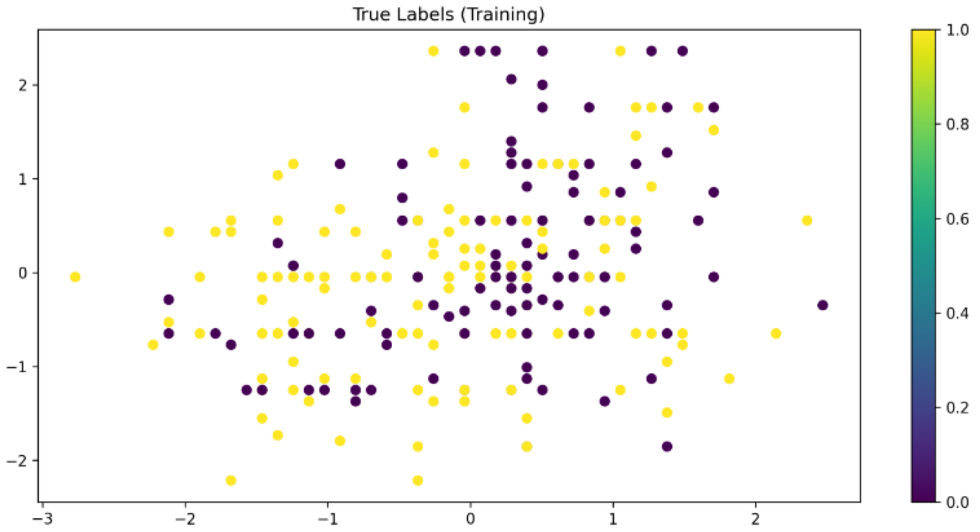
The graph seems strange with the decision boundary of SVM since much information is lost in this 2D projection, which is explained by PCA 36.42% total and 22.28% variance.

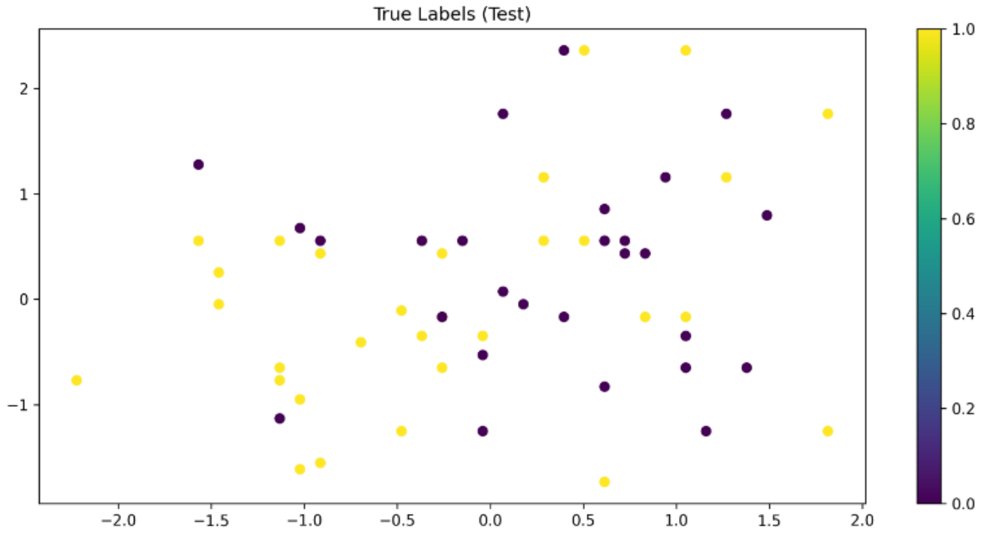
4.3 Compare all the classification model

Compare with the F1 score of all the classification models. SVM performs the best, but it is hard to understand and recognize in daily life. XGBoost has a simpler feature importance, which is Thalium Stress Test result (thall) with 2. Also, it has a simple classification tree for easier understanding.

4.4 Clustering

K-mean and DBSCAN clustering were used for clustering. Also, training data(80%) and test data (20%) are separated by comparing the results with different sample data sizes. Here is the plot of the true label:

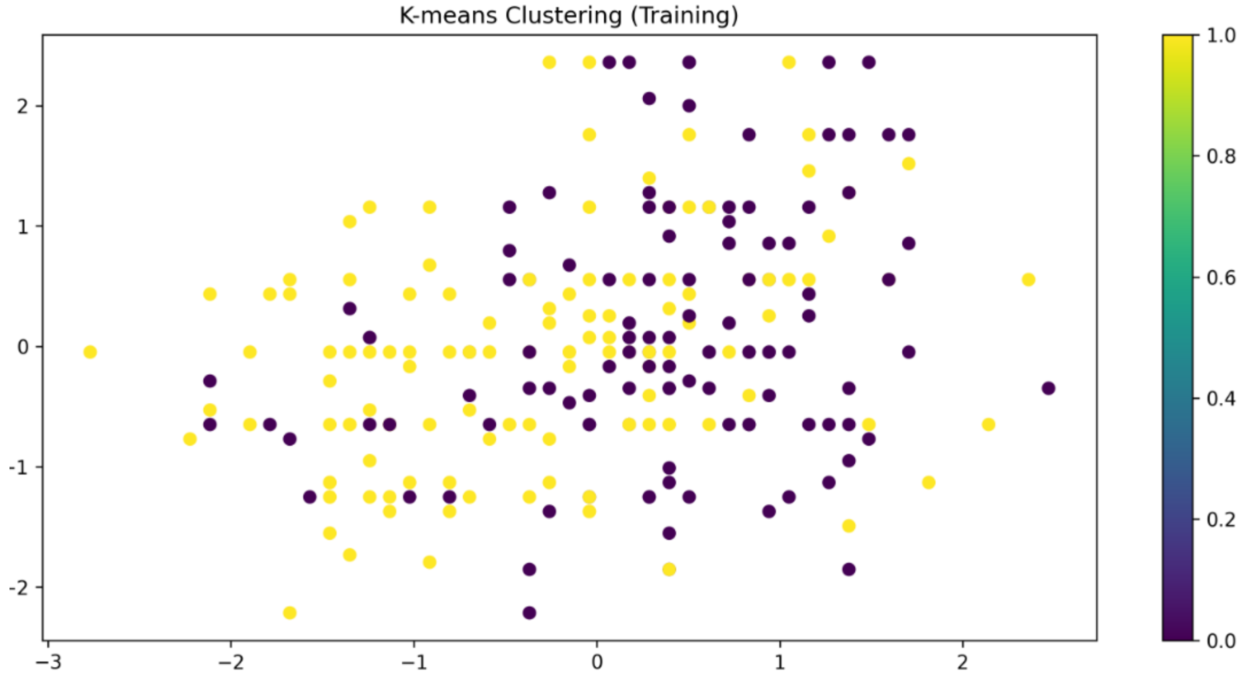




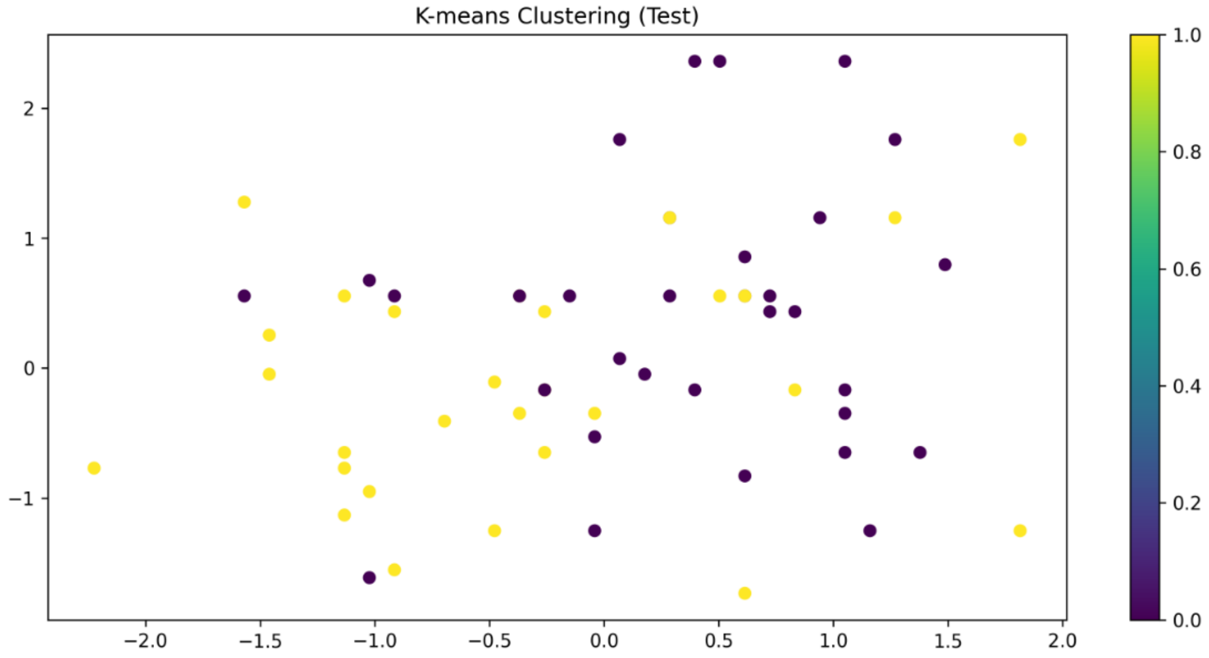
1. K-mean

The parameters that have been used are 2 number of clusters and a random state of 42 for reproducibility.

Here is the plot of the training set and test set with K-mean clustering:

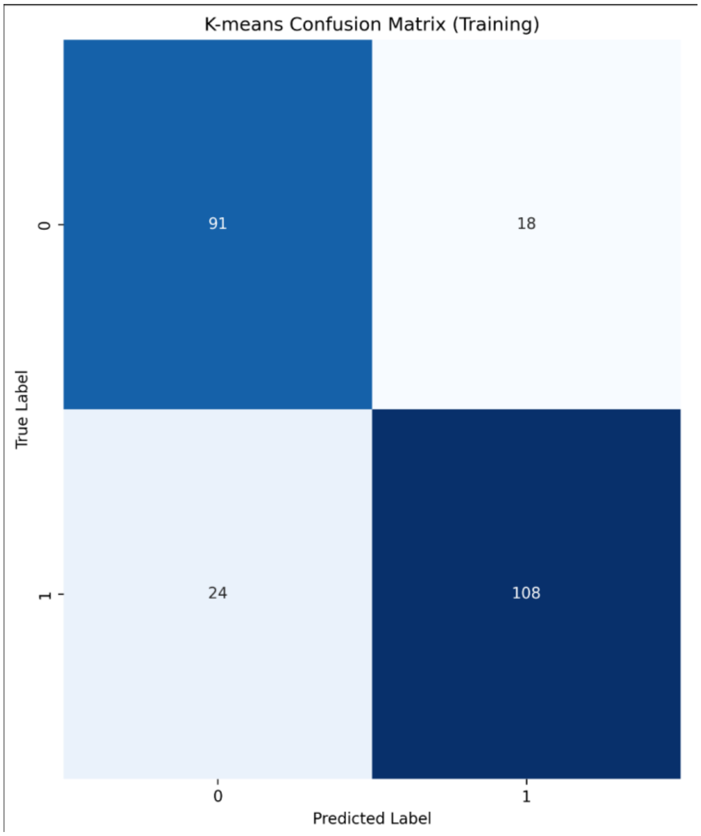


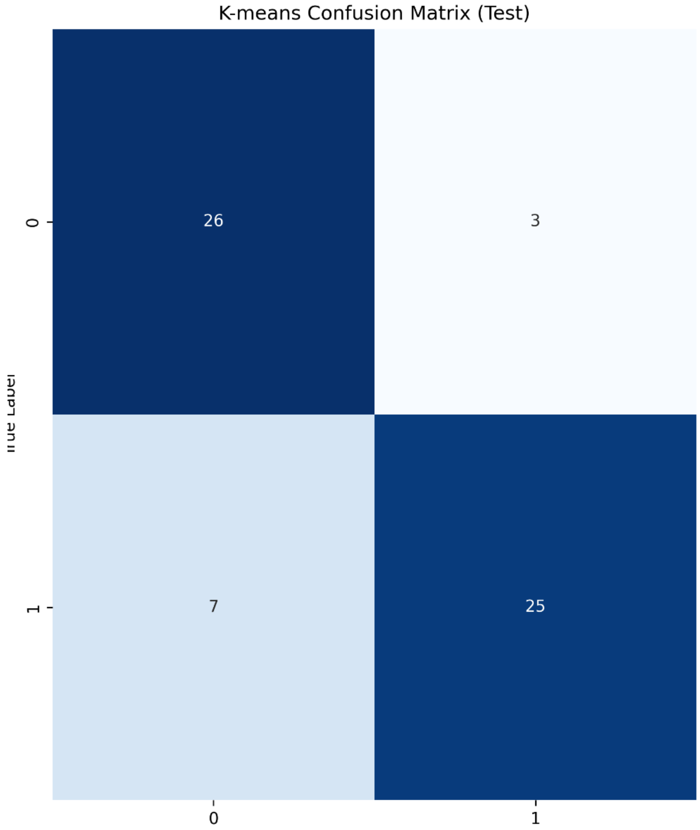
K-means Clustering (Training)

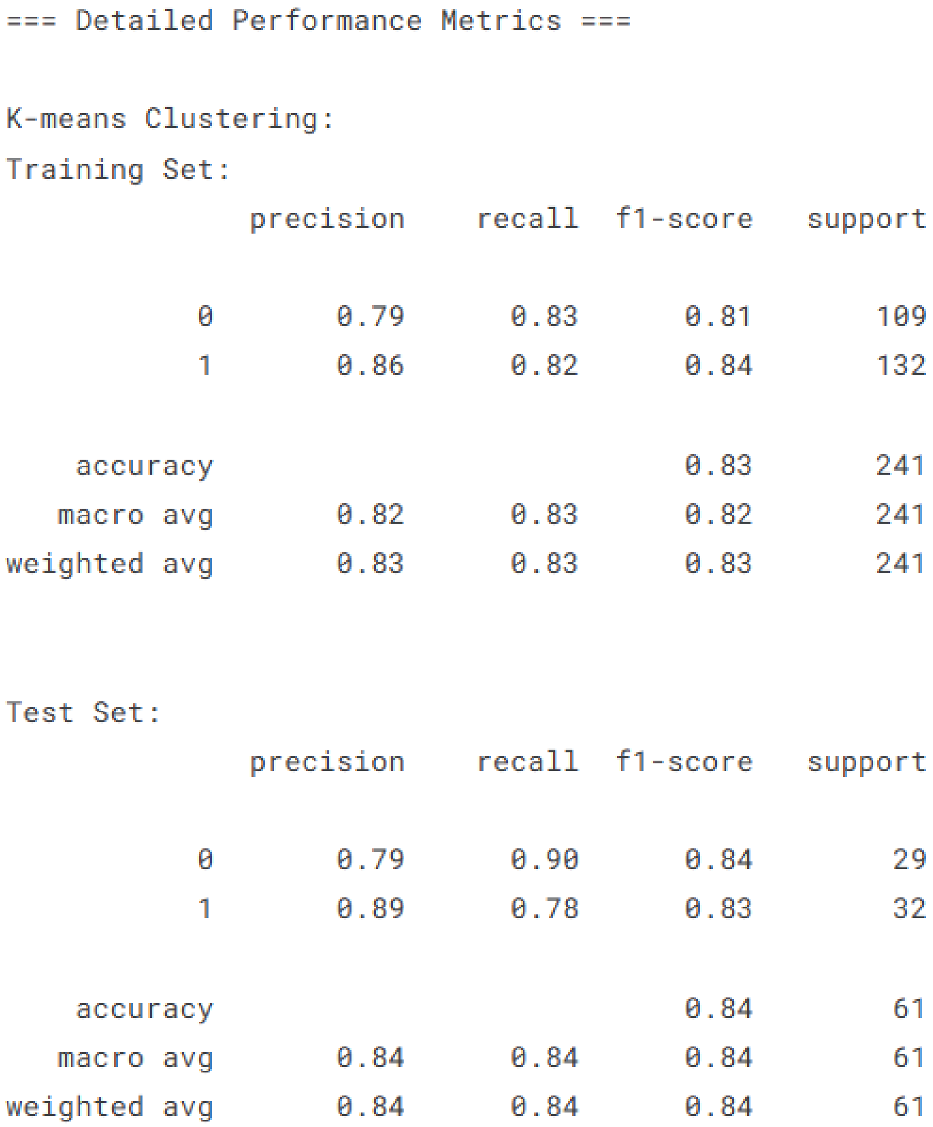


K-means Clustering (Test)

Here show the true label and the predicted label shown from k-mean clustering

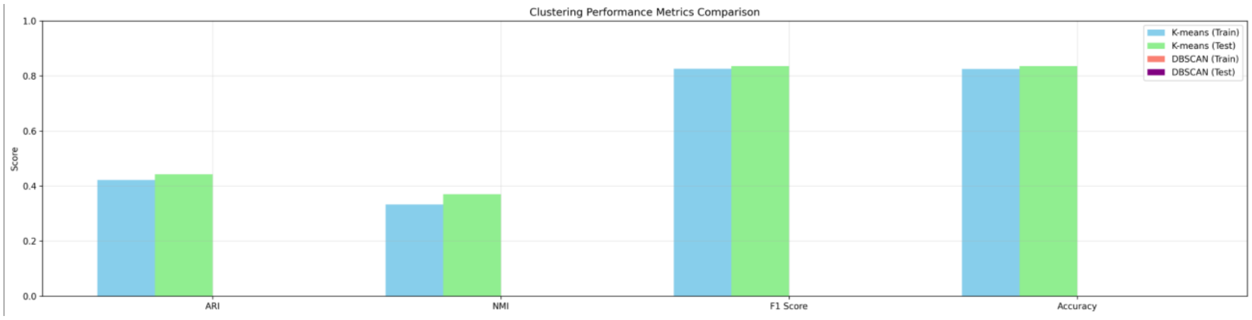

K-means Confusion Matrix (Training)


K-means Confusion Matrix (Test)



```
=== Detailed Performance Metrics ===

K-means Clustering:
Training Set:
              precision    recall  f1-score   support

           0       0.79      0.83      0.81       109
           1       0.86      0.82      0.84       132

    accuracy                           0.83       241
   macro avg       0.82      0.83      0.82       241
weighted avg       0.83      0.83      0.83       241



Test Set:
              precision    recall  f1-score   support

           0       0.79      0.90      0.84        29
           1       0.89      0.78      0.83        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

The above graph shows that the f1 score of k-mean clustering is 0. in the weighted average in the training set and 0.16 in the weighted average in the test set, which is quite a bad result for classifying the risk of heart attack prediction.

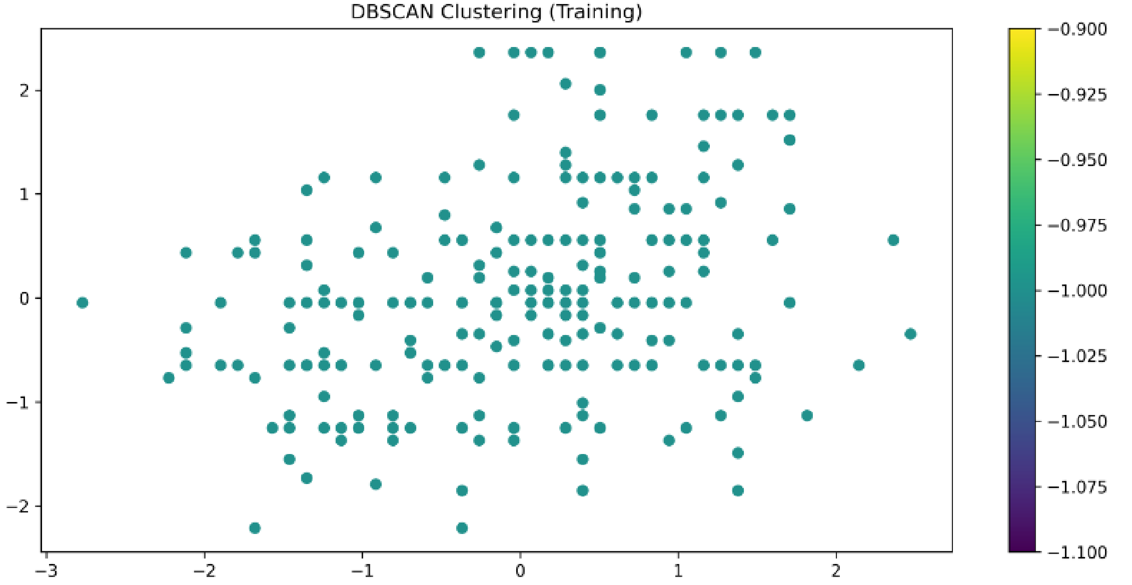


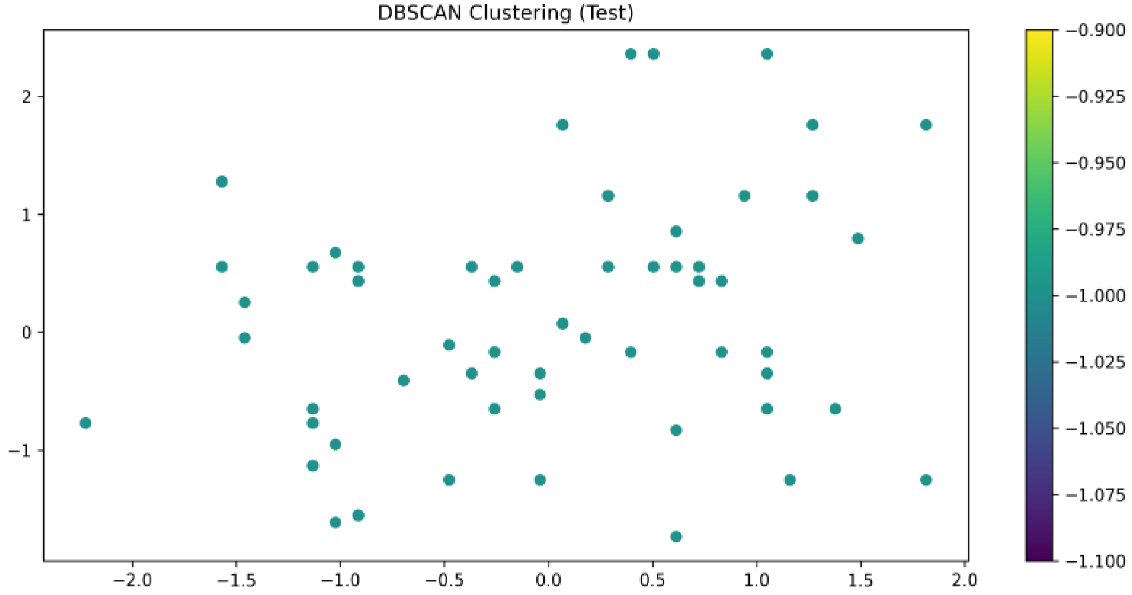
The above shows the performance compared in k-mean with the train dataset and test dataset. The average performance of the test set is slightly higher than that of the training set.

2. DBSCAN

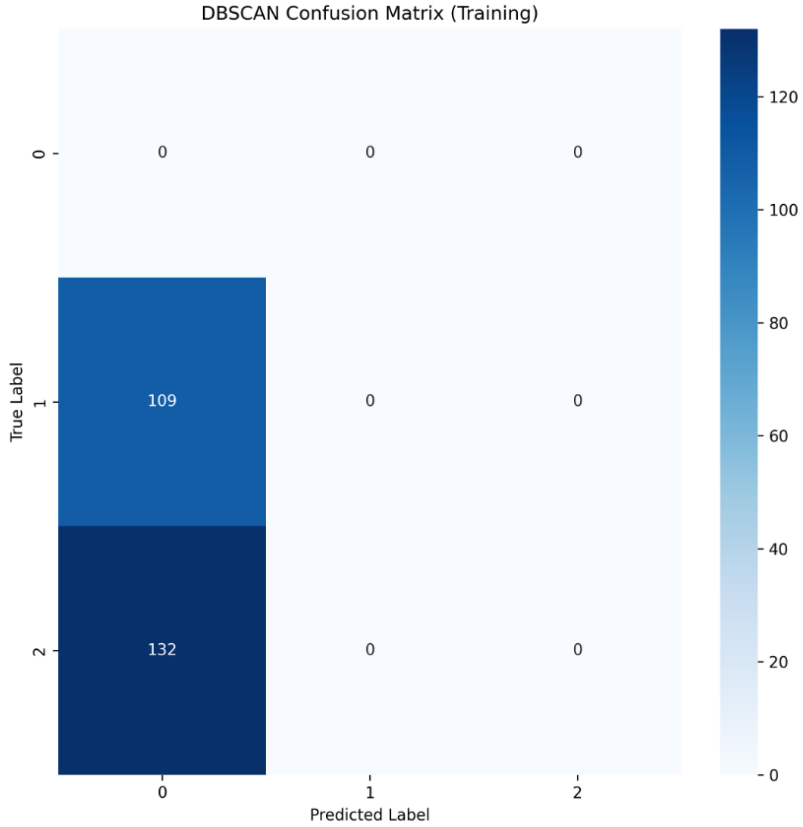
The parameters that have been used are 0.5 maximum distance between samples and 5 minimum points to form a dense region

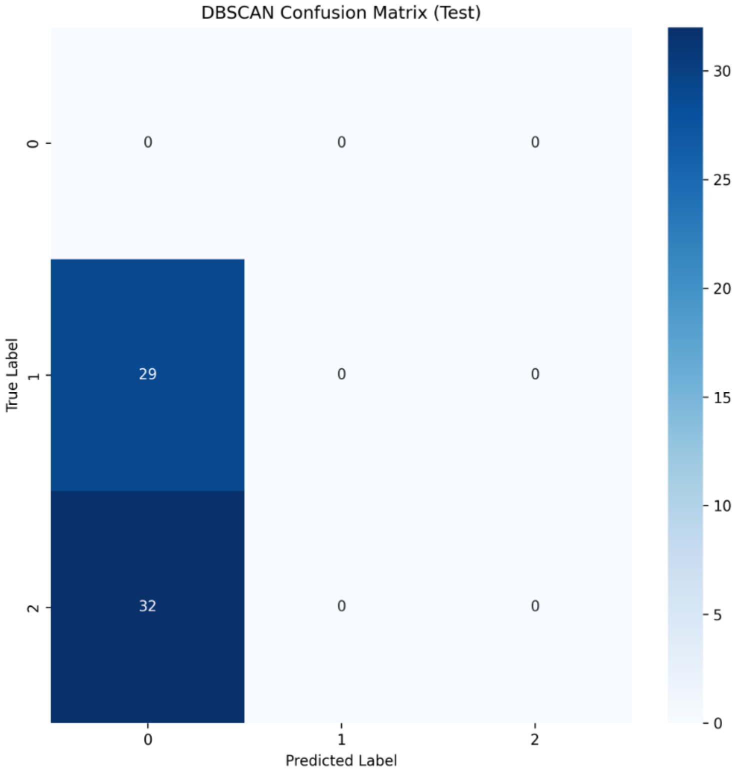
Here is the plot of the training set and test set with DBSCAN clustering:

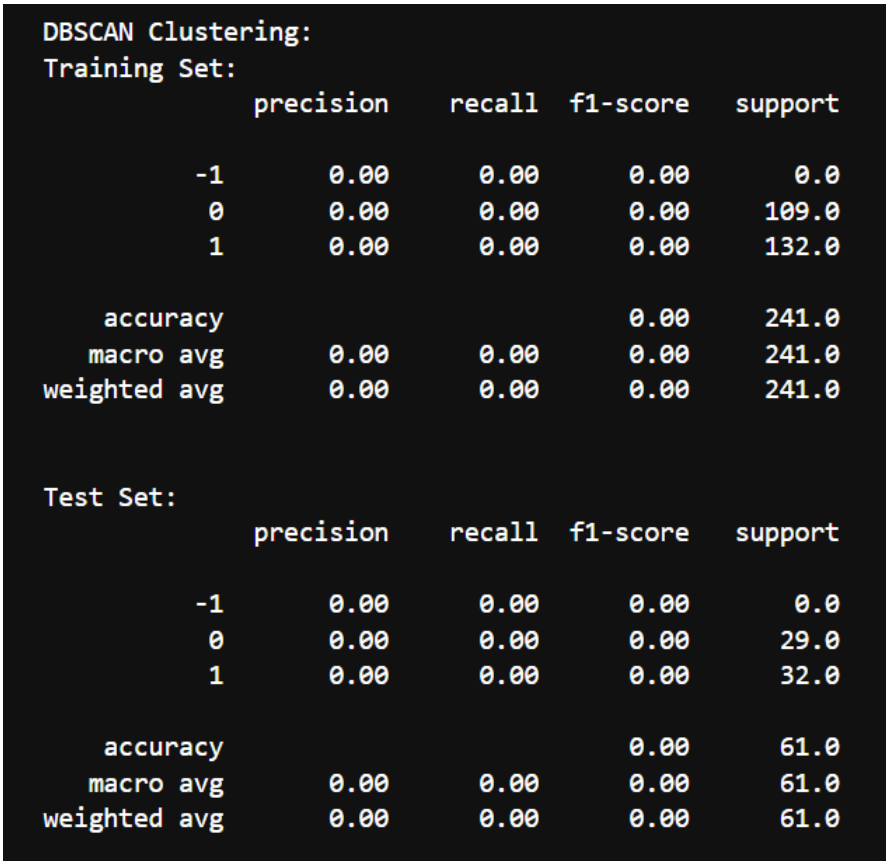




Here show the true label and the predicted label shown from the DBSCAN clustering


DBSCAN Confusion Matrix (Training)


DBSCAN Confusion Matrix (Test)



```
DBSCAN Clustering:
Training Set:
              precision    recall  f1-score   support

          -1       0.00      0.00      0.00       0.0
           0       0.00      0.00      0.00     109.0
           1       0.00      0.00      0.00     132.0

    accuracy                           0.00     241.0
   macro avg       0.00      0.00      0.00     241.0
weighted avg       0.00      0.00      0.00     241.0


Test Set:
              precision    recall  f1-score   support

          -1       0.00      0.00      0.00       0.0
           0       0.00      0.00      0.00      29.0
           1       0.00      0.00      0.00      32.0

    accuracy                           0.00      61.0
   macro avg       0.00      0.00      0.00      61.0
weighted avg       0.00      0.00      0.00      61.0
```

The above graph shows that the f1 score of DB clustering is 0 due to the failure of the clustering of the data.

4.5 Overall of Clustering

In the result, It is shown that k -mean does a great result accuracy in predicting

5. Association Rule Mining

Association Rule Mining determines the relationship between the different items in the data set. Here, we consider the relationship between the features and the output, which is the risk of heart attack.

5.1 Data Preprocessing:

The continuous variables, including age, blood pressure (trtbps), cholesterol (chol), maximum heart rate (thalachh), and ST depression (oldpeak), were binned into four categories to discretise the data. Simultaneously, categorical variables such as sex, exercise-induced angina (exng), number of major vessels (caa), chest pain type (cp), and others were transformed into labelled categories for better interpretation.

5.2 Transaction Formation:

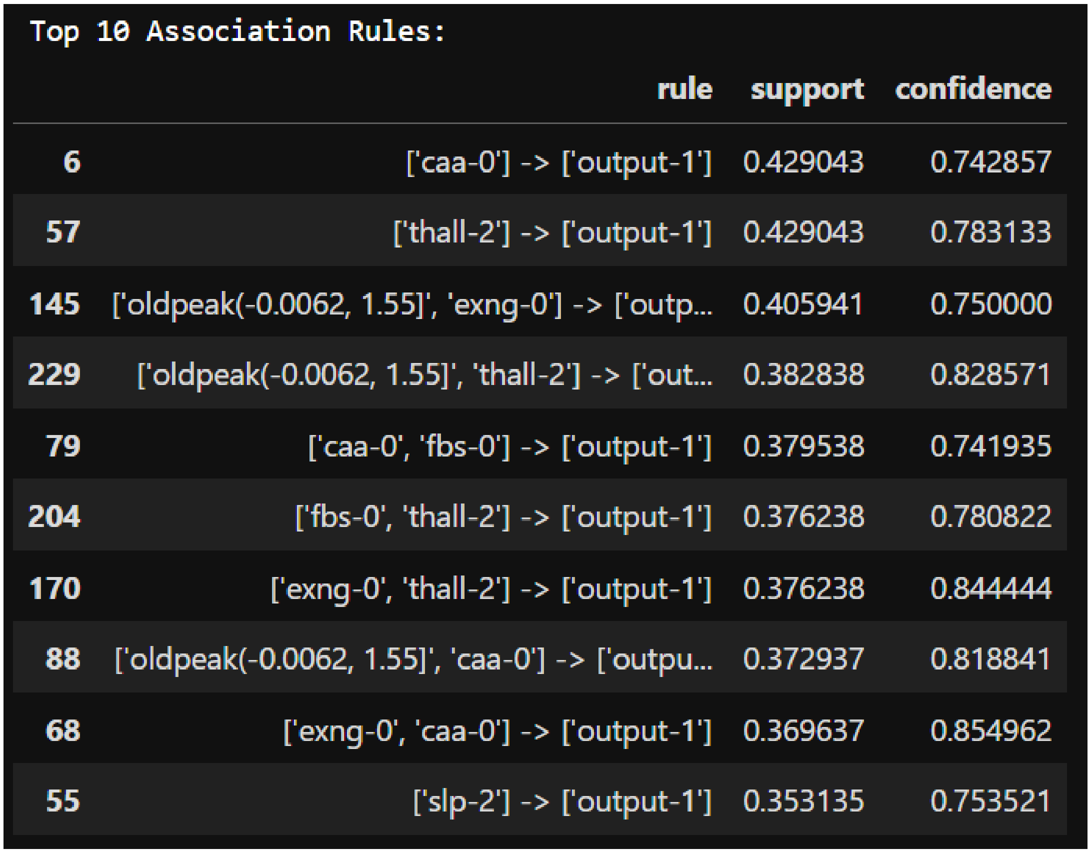
This was accomplished using the TransactionEncoder from the extended library, which converted the categorical data into a binary format that the Apriori algorithm could process.

5.3 Association Rule Mining:

The Apriori algorithm was then applied with a minimum support threshold of 0.3 to identify frequent itemsets within the data. These frequent itemsets formed the basis for generating association rules, with a minimum confidence threshold of 0.7 to ensure strong relationships between the antecedents and consequents.

5.4 Rule Analysis:

The Filtered rules only focus on those predicting 'output-1' (positive on heart attack risk) (Besides this 10 rule, other rules have been generated and exported into Excel)



Top 10 Association Rules:

	rule	support	confidence
6	['caa-0'] -> ['output-1']	0.429043	0.742857
57	['thall-2'] -> ['output-1']	0.429043	0.783133
145	['oldpeak(-0.0062, 1.55]', 'exng-0'] -> ['outp...	0.405941	0.750000
229	['oldpeak(-0.0062, 1.55]', 'thall-2'] -> ['out...	0.382838	0.828571
79	['caa-0', 'fbs-0'] -> ['output-1']	0.379538	0.741935
204	['fbs-0', 'thall-2'] -> ['output-1']	0.376238	0.780822
170	['exng-0', 'thall-2'] -> ['output-1']	0.376238	0.844444
88	['oldpeak(-0.0062, 1.55]', 'caa-0'] -> ['outpu...	0.372937	0.818841
68	['exng-0', 'caa-0'] -> ['output-1']	0.369637	0.854962
55	['slp-2'] -> ['output-1']	0.353135	0.753521

Rule 1:

If Number of major vessels(caa) = 0 → Then output = 1

(Support: 0.42904, Confidence: 0.74286)

Rule 2:

If Thalium Stress Test result (thall) = 2 → Then output = 1

(Support: 0.42904, Confidence: 0.78313)

Rule 3:

If Previous peak(oldpeak) is between -0.0062 and 1.55 AND exercise induced angina(exng) = 0 → Then output = 1

(Support: 0.40594, Confidence: 0.75000)

Rule 4:

If Previous peak(oldpeak) is between -0.0062 and 1.55 AND Thalium Stress Test result (thall) = 2 → Then output = 1

(Support: 0.38284, Confidence: 0.82857)

Rule 5:

If number of major vessels(caa) = 0 AND fasting blood sugar(fbs) = 0 → Then output = 1

(Support: 0.37954, Confidence: 0.74194)

Rule 6:

If fasting blood sugar(fbs) = 0 AND Thalium Stress Test result (thall) = 2 → Then output = 1

(Support: 0.37624, Confidence: 0.78082)

Rule 7:

If exercise induced angina(exng) = 0 AND Thalium Stress Test result (thall) = 2 → Then output = 1

(Support: 0.37624, Confidence: 0.84444)

Rule 8:

If oldpeak is between -0.0062 and 1.55 AND the number of major vessels(caa) = 0 → Then output = 1

(Support: 0.37294, Confidence: 0.81884)

Rule 9:

If exercise induced angina(exng) = 0 AND number of major vessels(caa) = 0 → Then output = 1

(Support: 0.36964, Confidence: 0.85496)

Rule 10:

If slope(slp) = 2 → Then output = 1

(Support: 0.35314, Confidence: 0.75352)

These are the rules that show different combinations of factors that are associated with heart attack risk. . The highest rule support for the rule with 1 condition is the number of major vessels(caa) with 0 and the Thalium Stress Test result (thall) with 2. This means people having 2 number vessels or 2 thalassemia will have a high risk of heart attack. For the rule with 2 conditions, the Previous peak(oldpeak) is between -0.0062 and 1.55 AND exercise-induced angina(exng) with 0. Also, the Previous peak(oldpeak) is between -0.0062 and 1.55 AND the Thalium Stress Test result (thall) with 2 will have a higher risk of a heart attack.

These rules show the number of major vessels with 0, thalassemia with 2, oldpeak is between -0.0062 and 1.55, and angina with 0 will have a higher heart attack rate.

6. Conclusion

After comparing the classification, clustering, and Association Rule Mining approach, SVM offers the highest predictive accuracy, but XGBoost provides the best understanding of the classification tree. The clustering method only k-mean is working with the data, and DBSCAN doesn’t align well with this heart attack risk dataset. Also, ARM analysis provides valuable complementary insights for understanding risk factor combinations.

These three methods show that the Thalium Stress Test result with 2 strongly correlates with heart attack risk.

Reference:

[1] Namanmanchanda, “Heart attack - eda + prediction (90% accuracy),” Kaggle, https://www.kaggle.com/code/namanmanchanda/heart-attack-eda-prediction-90-accuracy?scriptVersionId=63610771&cellId=4 (accessed Dec. 20, 2024).