King Saud University
College of Computer and Information Sciences
Information Technology department

IT 326: Data Mining
1st Semester 1446 H

Heart Attack

Group #4
Wednesday 2-4

Section #:	71162	
Group #:	Group #4 Wednesday 2-4	
Group Members	**NAME**	**ID**
	Aljawharah Alsubaie	444201140
	Fajer Alamro	444200800
	Rama Alomair	444200662
	Mashael Alammar	444200786
	Layal Alghamdi	444201202

Supervised By: L. Hanan Altamimi

Contents

1. Problem

Heart attacks are one of the leading causes of death worldwide, often due to delayed recognition of warning signs or subtle, unnoticed symptoms. In this project, we aim to study and analyse patient data to predict heart attack risks early, providing individuals and healthcare providers with the opportunity to take preventive action.

Early detection is crucial, it saves lives, reduces complications, and alleviates the burden on healthcare systems. By predicting the likelihood of a heart attack, this project seeks to empower individuals to take preventive measures and protect their health.[4]

2. Data Mining Task

In our project, we will employ two data mining tasks to help predict the likelihood of heart attacks: classification and clustering. For classification, we will train a model to determine whether an individual is at risk of experiencing a heart attack or not, based on a set of medical and lifestyle attributes such as age, cholesterol levels, blood pressure, heart rate, Diabetes, etc. Classification will be based on the "heart attack risk level" class, which could be binary (at risk or not).[2]

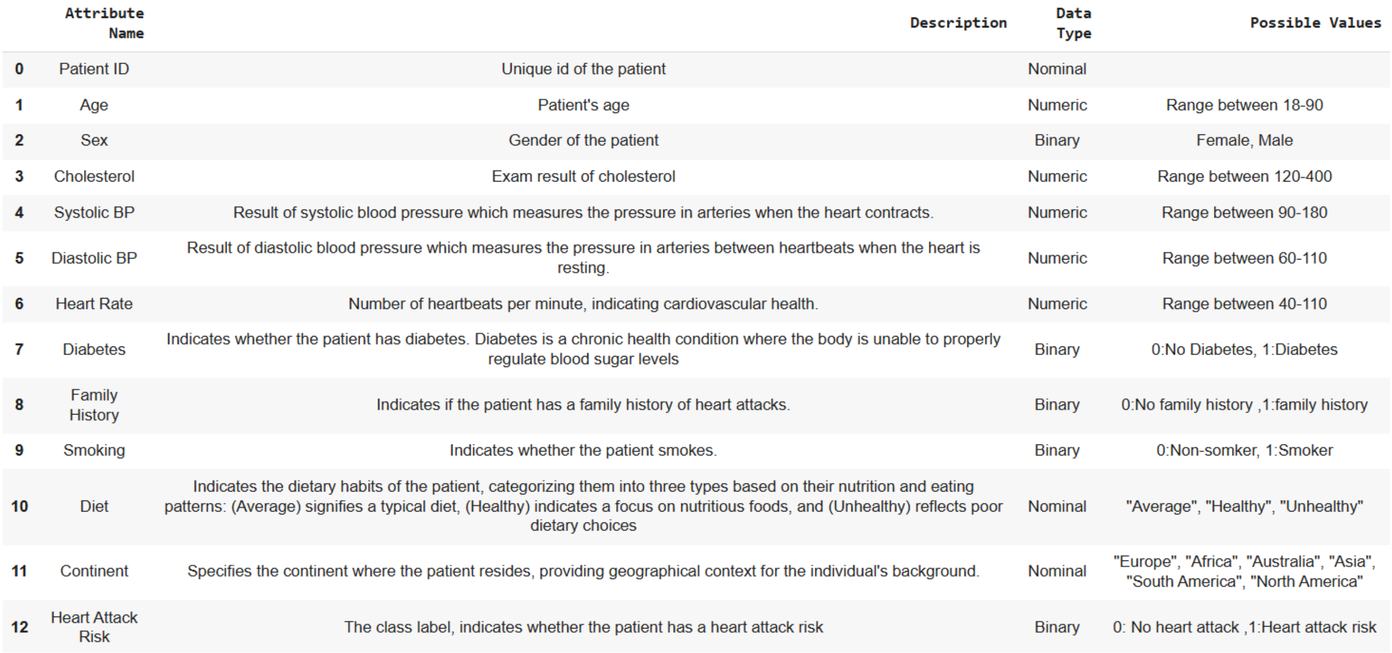
As for clustering, our model will group individuals with similar characteristics into clusters without considering the heart attack risk class. These clusters will help identify patterns and shared traits among individuals, offering deeper insights into risk factors and potential relationships between attributes. This approach may also uncover new insights to improve the understanding of heart attack risks and support targeted preventive strategies.

3. Data

The Source: https://www.kaggle.com/datasets/m1relly/heart-attack-prediction/data

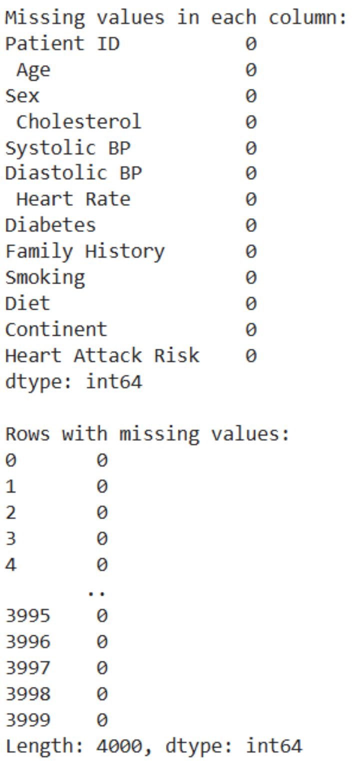
- Number of attributes: 12
- Number of objects: 4000
- Class label: Heart Attack Risk

- Attributes' description



	Attribute Name	Description	Data Type	Possible Values
0	Patient ID	Unique id of the patient	Nominal	
1	Age	Patient's age	Numeric	Range between 18-90
2	Sex	Gender of the patient	Binary	Female, Male
3	Cholesterol	Exam result of cholesterol	Numeric	Range between 120-400
4	Systolic BP	Result of systolic blood pressure which measures the pressure in arteries when the heart contracts.	Numeric	Range between 90-180
5	Diastolic BP	Result of diastolic blood pressure which measures the pressure in arteries between heartbeats when the heart is resting.	Numeric	Range between 60-110
6	Heart Rate	Number of heartbeats per minute, indicating cardiovascular health.	Numeric	Range between 40-110
7	Diabetes	Indicates whether the patient has diabetes. Diabetes is a chronic health condition where the body is unable to properly regulate blood sugar levels	Binary	0:No Diabetes, 1:Diabetes
8	Family History	Indicates if the patient has a family history of heart attacks.	Binary	0:No family history ,1:family history
9	Smoking	Indicates whether the patient smokes.	Binary	0:Non-somker, 1:Smoker
10	Diet	Indicates the dietary habits of the patient, categorizing them into three types based on their nutrition and eating patterns: (Average) signifies a typical diet, (Healthy) indicates a focus on nutritious foods, and (Unhealthy) reflects poor dietary choices	Nominal	"Average", "Healthy", "Unhealthy"
11	Continent	Specifies the continent where the patient resides, providing geographical context for the individual's background.	Nominal	"Europe", "Africa", "Australia", "Asia", "South America", "North America"
12	Heart Attack Risk	The class label, indicates whether the patient has a heart attack risk	Binary	0: No heart attack ,1:Heart attack risk

- Missing values



```
Missing values in each column:
Patient ID          0
 Age                0
Sex                 0
 Cholesterol        0
Systolic BP         0
Diastolic BP        0
 Heart Rate         0
Diabetes            0
Family History      0
Smoking             0
Diet                0
Continent           0
Heart Attack Risk   0
dtype: int64

Rows with missing values:
0       0
1       0
2       0
3       0
4       0
       ..
3995    0
3996    0
3997    0
3998    0
3999    0
Length: 4000, dtype: int64
```

We have no missing values. All columns are complete.

- Statical Measures for each numeric column:

Using the summary_stats () function, we observed several points from these summary statistics, such as:

- Age: The values ranging from 18 to 90 years, with an average of 53.75 years. This indicates that the risk of a heart attack is spread across individuals within a wide age range.

- Cholesterol: The values vary significantly, with a maximum of 400 and a minimum of 120 with a mean 260.71. It indicates considerable variability in cholesterol levels

- Heart Rate: The values range from 40 bpm to 110 bpm, with a mean of 74.9 bpm, which is within the normal heart rate range, which indicates variability in heart rates among individuals.

- Diabetes: The values are binary, limited to 0 and 1, with a mean of 0.65, indicating that a significant portion of the population may have diabetes.

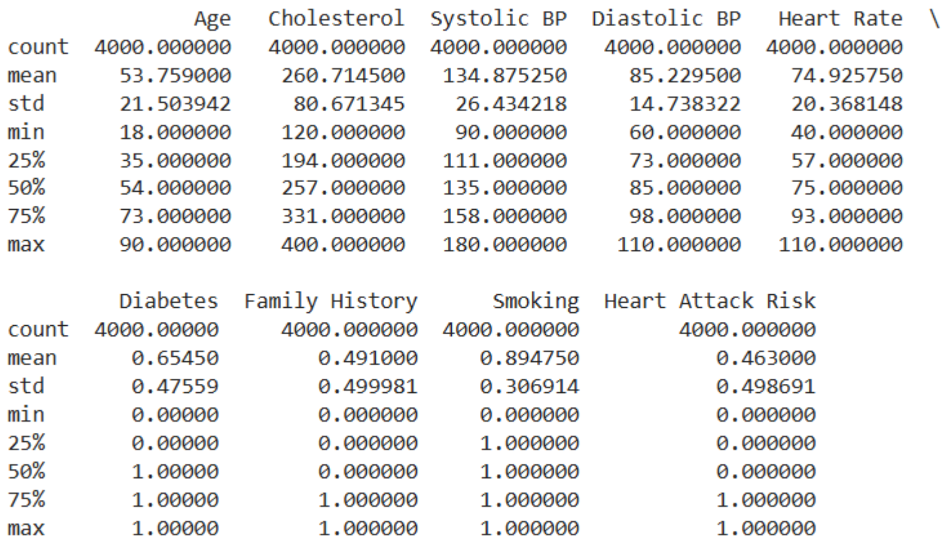
- Family History: The values are binary, limited to 0 and 1, with a mean of 0.491 indicating that nearly half of the observations report a family history of heart attack risk.

- Smoking: The values are binary, limited to 0 and 1, with a mean of 0.894, The mean indicates a high prevalence of smoking within the dataset. Since the mean is close to 1, it indicates that most individuals in the dataset are smokers.

- Systolic BP: The values vary, with a maximum of 180 and a minimum of 90, with the mean being 134.875, suggesting that the distribution is likely symmetrical which indicates a balanced spread of values around the center, indicating considerable variability in systolic BP observations (wide range of blood pressure levels).

- Diastolic BP: Diastolic BP values range from 60 to 110, with a mean of 85.23, indicating some variability in the diastolic BP observations.

- Heart Attack Risk: The values are binary, limited to 0 and 1, with a mean of 0.463 which suggests considerable variability in heart attack risk among the observations, with many individuals falling towards both ends of the risk spectrum.



```
               Age  Cholesterol  Systolic BP  Diastolic BP   Heart Rate  \
count  4000.000000  4000.000000  4000.000000   4000.000000  4000.000000   
mean     53.759000   260.714500   134.875250     85.229500    74.925750   
std      21.503942    80.671345    26.434218     14.738322    20.368148   
min      18.000000   120.000000    90.000000     60.000000    40.000000   
25%      35.000000   194.000000   111.000000     73.000000    57.000000   
50%      54.000000   257.000000   135.000000     85.000000    75.000000   
75%      73.000000   331.000000   158.000000     98.000000    93.000000   
max      90.000000   400.000000   180.000000    110.000000   110.000000   

         Diabetes  Family History      Smoking  Heart Attack Risk  
count  4000.00000     4000.000000  4000.000000        4000.000000  
mean      0.65450        0.491000     0.894750           0.463000  
std       0.47559        0.499981     0.306914           0.498691  
min       0.00000        0.000000     0.000000           0.000000  
25%       0.00000        0.000000     1.000000           0.000000  
50%       1.00000        0.000000     1.000000           0.000000  
75%       1.00000        1.000000     1.000000           1.000000  
max       1.00000        1.000000     1.000000           1.000000  
```

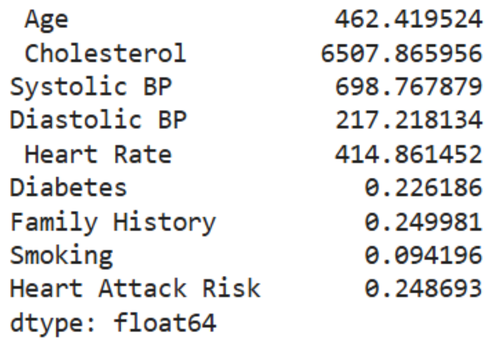
- Show the Variance:

Variance helps us understand the extent of dispersion or scatter of values in each column. As variance increases, it indicates that the values are more spread out from the mean; conversely, decreasing variance suggests that the values are more closely clustered around the mean. Therefore, our variance results indicate the following:

- Age, Systolic BP, Diastolic BP, Heart Rate: These columns exhibit high variance, indicating a significant level of dispersion and spread of values.

- Cholesterol: This column shows very high variance, reflecting an even greater level of dispersion.

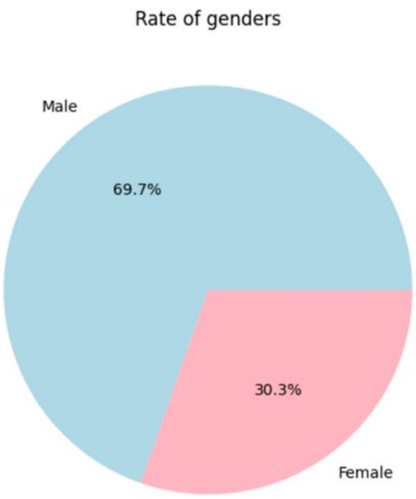
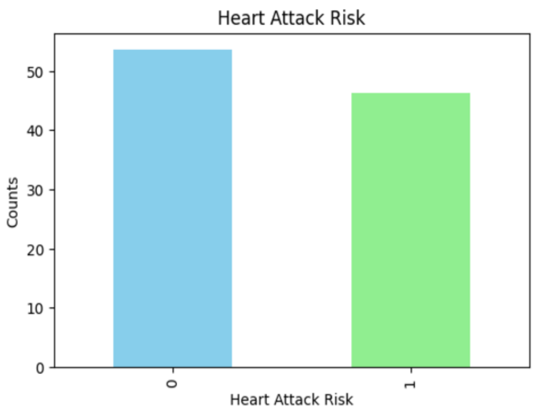
- Diabetes, Family History, Smoking, and Heart Attack Risk: These columns have moderate to low variance, suggesting a moderate to low degree of dispersion and value spread."

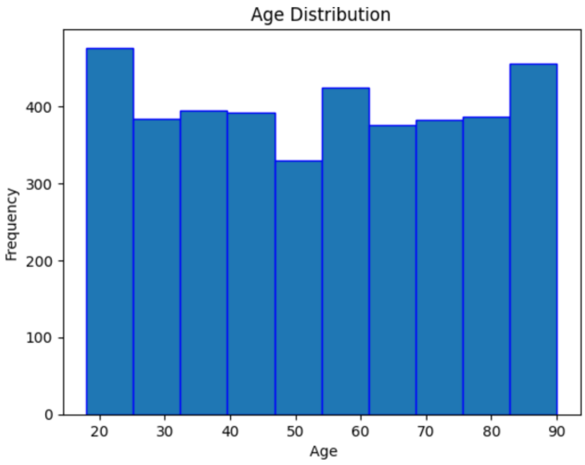
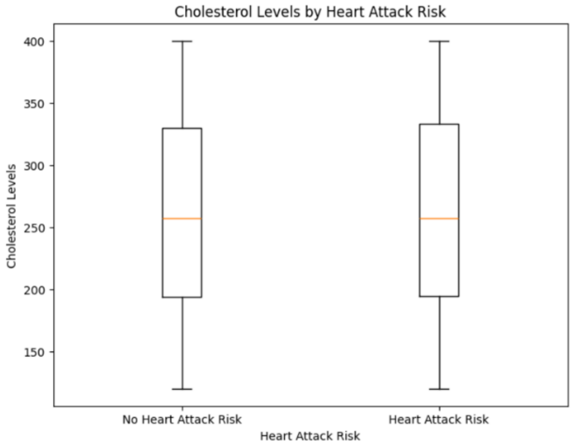


```
 Age                   462.419524
 Cholesterol          6507.865956
Systolic BP            698.767879
Diastolic BP           217.218134
 Heart Rate            414.861452
Diabetes                 0.226186
Family History           0.249981
Smoking                  0.094196
Heart Attack Risk        0.248693
dtype: float64
```

- Understanding the data through graph representations:

The "heart attack risk" class name was primarily utilized to comprehend the relationship between heart attack and all attributes, specifically how they are associated with the probability of having a heart attack. It indicates which individuals are affected and whether they are not by linking to every attribute in the data. It also helps determine the gender differences in heart attack risk and the relationship between age and heart attack risk. This makes it easier to understand the factors influencing this condition and to find signs that could aid in an early diagnosis.

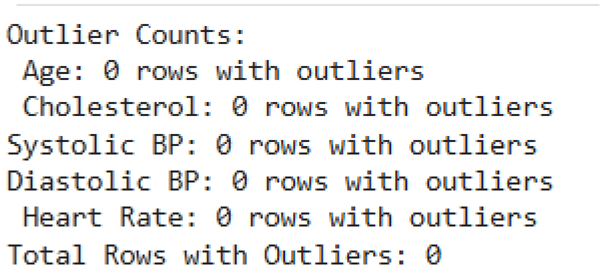
Name of Graph	Picture of Graph	Description
Pie Chart (nominal data)		People affected by heart attack were selected from the 'heart attack risk' attribute for both genders of the 'Gander' attribute for comparison between the ratio of affected women and men. As a result, it has been found that men are significantly more susceptible to liver disease than women.
Bar Plot (binary data)		The bar chart illustrates the numerical differences between class label, while the classification results reveal a balance in heart attack risk between individuals at risk and those not at risk.

Name of Graph	Picture of Graph	Description
Histogram (numeric data)		The histogram illustrates the age distribution of the study participants, indicating that individuals in their twenties are the most likely to participate.
Boxplot Chart (numeric data)		This is a Boxplot that compares the distribution of cholesterol levels between patients with and without a heart attack risk, allowing you to see the spread, median, and potential outliers in each group Median. Both groups have a similar median cholesterol level, indicating that the central tendency of cholesterol is approximately the same for patients with and without heart attack risk.

4. Data preprocessing:

[3]

- Detecting the outliers:



```
Outlier Counts:
 Age: 0 rows with outliers
 Cholesterol: 0 rows with outliers
Systolic BP: 0 rows with outliers
Diastolic BP: 0 rows with outliers
 Heart Rate: 0 rows with outliers
Total Rows with Outliers: 0
```

How to calculate outliers?

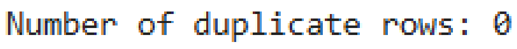
The IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

- An upper bound is defined as Q3+(1.5 x IQR)
- A lower bound is defined as Q1-(1.5 × IQR)

Data points falling outside these bounds are considered outliers

The results indicate that no outliers were detected for any of the attributes, with all counts reported as **0 rows with outliers**. This suggests a clean and well-structured dataset, free from anomalies or extreme values that fall outside the expected range for these variables. Consequently, the dataset is likely well-prepared for further analysis, eliminating the need for additional outlier handling or preprocessing.

- Show duplicates:



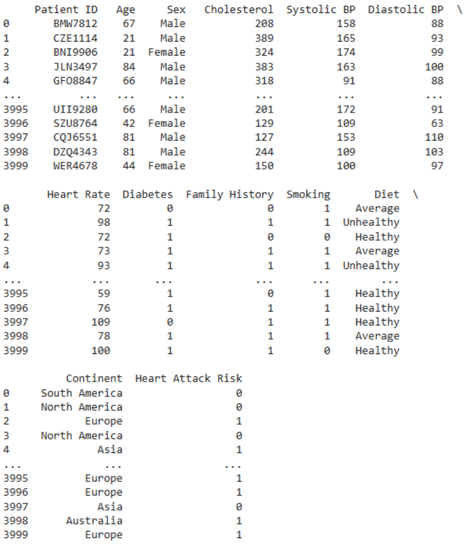
```
Number of duplicate rows: 0
```

This indicates that there are no duplicate rows in the dataset.

- Data Transformation:

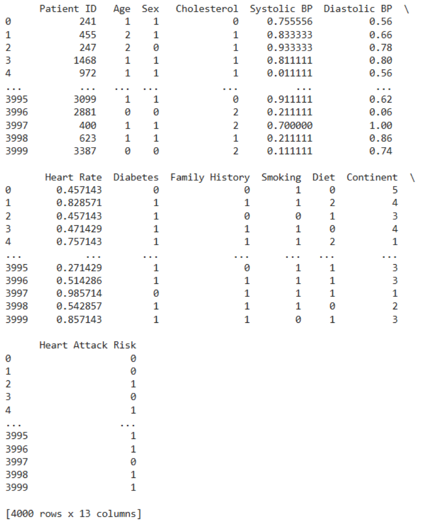
1. Encoding:

Before:



```
     Patient ID  Age     Sex  Cholesterol  Systolic BP  Diastolic BP  \
0       BMW7812   67    Male          208          158            88
1       CZE1114   21    Male          389          165            93
2       BNI9906   21  Female          324          174            99
3       JLN3497   84    Male          383          163           100
4       GFO8847   66    Male          318           91            88
...         ...  ...     ...          ...          ...           ...
3995    UII9280   66    Male          201          172            91
3996    SZU8764   42  Female          129          109            63
3997    CQJ6551   81    Male          127          153           110
3998    DZQ4343   81    Male          244          109           103
3999    WER4678   44  Female          150          100            97

      Heart Rate  Diabetes  Family History  Smoking       Diet  \
0             72         0               0        1    Average
1             98         1               1        1  Unhealthy
2             72         1               0        0    Healthy
3             73         1               1        1    Average
4             93         1               1        1  Unhealthy
...          ...       ...             ...      ...        ...
3995          59         1               0        1    Healthy
3996          76         1               1        1    Healthy
3997         109         0               1        1    Healthy
3998          78         1               1        1    Average
3999         100         1               1        0    Healthy

          Continent  Heart Attack Risk
0     South America                  0
1     North America                  0
2            Europe                  1
3     North America                  0
4              Asia                  1
...             ...                ...
3995         Europe                  1
3996         Europe                  1
3997           Asia                  0
3998      Australia                  1
3999         Europe                  1
```

After:

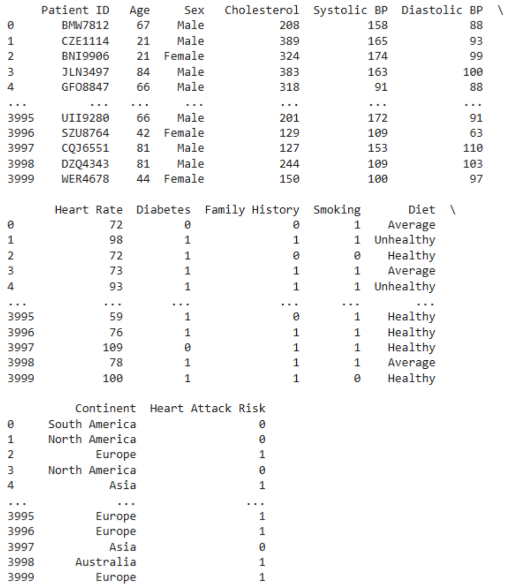


```
      Patient ID  Age  Sex  Cholesterol  Systolic BP  Diastolic BP  \
0            241    1    1            0     0.755556          0.56
1            455    2    1            1     0.833333          0.66
2            247    2    0            1     0.933333          0.78
3           1468    1    1            1     0.811111          0.80
4            972    1    1            1     0.011111          0.56
...          ...  ...  ...          ...          ...           ...
3995        3099    1    1            0     0.911111          0.62
3996        2881    0    0            2     0.211111          0.06
3997         400    1    1            2     0.700000          1.00
3998         623    1    1            1     0.211111          0.86
3999        3387    0    0            2     0.111111          0.74

      Heart Rate  Diabetes  Family History  Smoking  Diet  Continent  \
0       0.457143         0               0        1     0          5
1       0.828571         1               1        1     2          4
2       0.457143         1               0        0     1          3
3       0.471429         1               1        1     0          4
4       0.757143         1               1        1     2          1
...          ...       ...             ...      ...   ...        ...
3995    0.271429         1               0        1     1          3
3996    0.514286         1               1        1     1          3
3997    0.985714         0               1        1     1          1
3998    0.542857         1               1        1     0          2
3999    0.857143         1               1        0     1          3

      Heart Attack Risk
0                     0
1                     0
2                     1
3                     0
4                     1
...                 ...
3995                  1
3996                  1
3997                  0
3998                  1
3999                  1

[4000 rows x 13 columns]
```

This encoding method converts categorical variables in the dataset, such as gender, family history, smoking, and diet, into numerical values for computational purposes. For example, gender is encoded as 1 for male and 0 for female, while other categories like smoking and diet are transformed into 0 and 1 to represent different conditions. This standardization simplifies data processing and makes it compatible with machine learning models. Additionally, continuous variables like cholesterol and blood pressure are scaled to a range between 0 and 1, improving their use in predictive analysis.

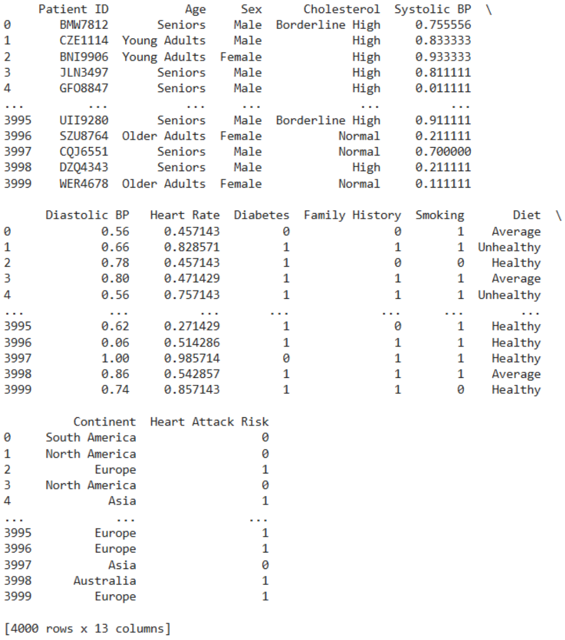
2. Normalization:

Before:



```
     Patient ID  Age     Sex  Cholesterol  Systolic BP  Diastolic BP  \
0       BMW7812   67    Male          208          158            88
1       CZE1114   21    Male          389          165            93
2       BNI9906   21  Female          324          174            99
3       JLN3497   84    Male          383          163           100
4       GFO8847   66    Male          318           91            88
...         ...  ...     ...          ...          ...           ...
3995    UII9280   66    Male          201          172            91
3996    SZU8764   42  Female          129          109            63
3997    CQJ6551   81    Male          127          153           110
3998    DZQ4343   81    Male          244          109           103
3999    WER4678   44  Female          150          100            97

      Heart Rate  Diabetes  Family History  Smoking       Diet  \
0             72         0               0        1    Average
1             98         1               1        1  Unhealthy
2             72         1               0        0    Healthy
3             73         1               1        1    Average
4             93         1               1        1  Unhealthy
...          ...       ...             ...      ...        ...
3995          59         1               0        1    Healthy
3996          76         1               1        1    Healthy
3997         109         0               1        1    Healthy
3998          78         1               1        1    Average
3999         100         1               1        0    Healthy

          Continent  Heart Attack Risk
0     South America                  0
1     North America                  0
2            Europe                  1
3     North America                  0
4              Asia                  1
...             ...                ...
3995         Europe                  1
3996         Europe                  1
3997           Asia                  0
3998      Australia                  1
3999         Europe                  1
```

After:

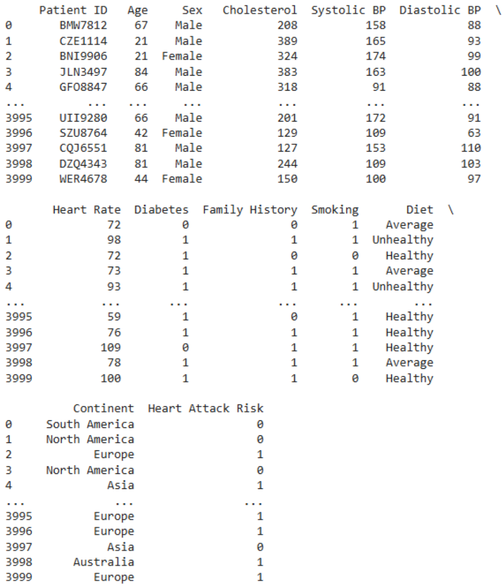


```
     Patient ID           Age     Sex      Cholesterol  Systolic BP  \
0       BMW7812       Seniors    Male  Borderline High     0.755556
1       CZE1114  Young Adults    Male             High     0.833333
2       BNI9906  Young Adults  Female             High     0.933333
3       JLN3497       Seniors    Male             High     0.811111
4       GFO8847       Seniors    Male             High     0.011111
...         ...           ...     ...              ...          ...
3995    UII9280       Seniors    Male  Borderline High     0.911111
3996    SZU8764  Older Adults  Female           Normal     0.211111
3997    CQJ6551       Seniors    Male           Normal     0.700000
3998    DZQ4343       Seniors    Male             High     0.211111
3999    WER4678  Older Adults  Female           Normal     0.111111

      Diastolic BP  Heart Rate  Diabetes  Family History  Smoking       Diet  \
0             0.56    0.457143         0               0        1    Average
1             0.66    0.828571         1               1        1  Unhealthy
2             0.78    0.457143         1               0        0    Healthy
3             0.80    0.471429         1               1        1    Average
4             0.56    0.757143         1               1        1  Unhealthy
...            ...         ...       ...             ...      ...        ...
3995          0.62    0.271429         1               0        1    Healthy
3996          0.06    0.514286         1               1        1    Healthy
3997          1.00    0.985714         0               1        1    Healthy
3998          0.86    0.542857         1               1        1    Average
3999          0.74    0.857143         1               1        0    Healthy

          Continent  Heart Attack Risk
0     South America                  0
1     North America                  0
2            Europe                  1
3     North America                  0
4              Asia                  1
...             ...                ...
3995         Europe                  1
3996         Europe                  1
3997           Asia                  0
3998      Australia                  1
3999         Europe                  1

[4000 rows x 13 columns]
```

Here in the Normalization method, we normalize the attributes and unify their scale since the range for each attribute is quite different, this method helps us to format all the values in the dataset and facilitates the analysis process.

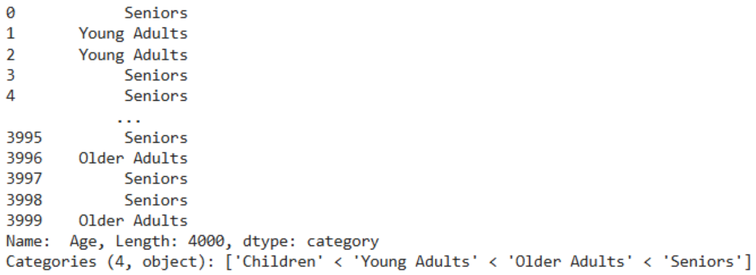
3. Discretization:

Before:

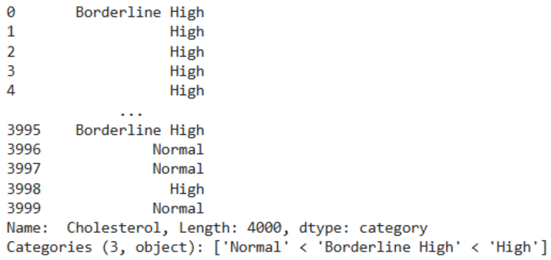


```
      Patient ID  Age     Sex  Cholesterol  Systolic BP  Diastolic BP  \
0        BMW7812   67    Male          208          158            88   
1        CZE1114   21    Male          389          165            93   
2        BNI9906   21  Female          324          174            99   
3        JLN3497   84    Male          383          163           100   
4        GFO8847   66    Male          318           91            88   
...          ...  ...     ...          ...          ...           ...   
3995     UII9280   66    Male          201          172            91   
3996     SZU8764   42  Female          129          109            63   
3997     CQJ6551   81    Male          127          153           110   
3998     DZQ4343   81    Male          244          109           103   
3999     WER4678   44  Female          150          100            97   

      Heart Rate  Diabetes  Family History  Smoking       Diet  \
0             72         0               0        1    Average   
1             98         1               1        1  Unhealthy   
2             72         1               0        0    Healthy   
3             73         1               1        1    Average   
4             93         1               1        1  Unhealthy   
...          ...       ...             ...      ...        ...   
3995          59         1               0        1    Healthy   
3996          76         1               1        1    Healthy   
3997         109         0               1        1    Healthy   
3998          78         1               1        1    Average   
3999         100         1               1        0    Healthy   

          Continent  Heart Attack Risk  
0     South America                  0  
1     North America                  0  
2            Europe                  1  
3     North America                  0  
4              Asia                  1  
...             ...                ...  
3995         Europe                  1  
3996         Europe                  1  
3997           Asia                  0  
3998      Australia                  1  
3999         Europe                  1  
```

After:



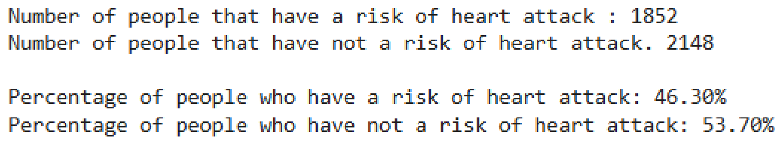
```
0            Seniors
1       Young Adults
2       Young Adults
3            Seniors
4            Seniors
            ...     
3995         Seniors
3996    Older Adults
3997         Seniors
3998         Seniors
3999    Older Adults
Name:  Age, Length: 4000, dtype: category
Categories (4, object): ['Children' < 'Young Adults' < 'Older Adults' < 'Seniors']
```



```
0       Borderline High
1                  High
2                  High
3                  High
4                  High
             ...       
3995    Borderline High
3996             Normal
3997             Normal
3998               High
3999             Normal
Name:  Cholesterol, Length: 4000, dtype: category
Categories (3, object): ['Normal' < 'Borderline High' < 'High']
```

In the discretization method, we categorize numerical age values into four groups: Children (0-17 years), Young Adults (18-34 years), Older Adults (35-65 years), and Seniors (65+ years). Also, we categorize Cholesterol into three groups: Normal (0-200), Borderline High (201-239), High (240-400). This simplifies data interpretation and analysis by grouping individuals into meaningful life stages. It enables clearer visualization and easier comparison of age demographics and enhances the interpretability of analytical results for stakeholders.

- Balance Data:



```
Number of people that have a risk of heart attack : 1852
Number of people that have not a risk of heart attack. 2148

Percentage of people who have a risk of heart attack: 46.30%
Percentage of people who have not a risk of heart attack: 53.70%
```

The number of people is 4000; we note that 1852 are at risk of having a heart attack, while 2148 are not at risk. In addition, we note that the data is balanced, as the percentage of those exposed to a risk is 46.3% and the percentage of those not exposed is 53.7%.

5. Data Mining Technique:

We utilized both supervised and unsupervised learning methods on our data through the use of classification and clustering techniques.

For our classification task, we used a decision tree. This recursive algorithm creates a tree structure where each leaf node corresponds to a final decision. Our model aims to predict whether a person is at risk of a heart attack, categorizing the results into ("1" meaning high risk) or ("0" meaning low risk). It makes predictions based on several attributes: Patient ID, Age, Sex, Cholesterol, Systolic BP, Diastolic BP, Heart Rate, Diabetes, Family History, Smoking, Diet, and Continent.

As we touched on before, classification is a type of supervised learning, so we need training data to train the model. We split our dataset into two subsets: training data and testing data. We tried three different sizes of training subsets: 70%, 60%, and 80%, and used two attribute selection measures (Information Gain (Entropy) and Gini Index). To evaluate our model and determine the best partitioning, we used accuracy_score to measure overall performance, and a confusion matrix to summarize basic performance evaluation measures such as sensitivity, specificity, precision, and error rate.

For implementing classification, we utilized the following Python libraries:

- pandas for data loading and preprocessing.

- sklearn for DecisionTreeClassifier, train_test_split, accuracy_score, and confusion_matrix.

- matplotlib.pyplot for trees visualization.

In the clustering process, which is a type of unsupervised learning, we omitted the “Heart Attack Risk” class label attribute since clustering does not use class labels. Instead, we utilized all other attributes such as: Patient ID, Age, Sex, Cholesterol, Systolic BP, Diastolic BP, Heart Rate, Diabetes, Family History, Smoking, Diet, and Continent. All of these attributes are numeric, or were converted to numeric values where needed, prior to clustering.

For creating the clusters, we employed the K-means algorithm. To determine the optimal number of clusters (k), we used the Elbow Method, which plots the within-cluster sum of squares (WSS) against the number of clusters. The “elbow point” is the value of k at which the WSS starts to level off, indicating diminishing returns in cluster compactness with increasing k. From the plot, we determined the optimal number of clusters to be 4.

We then fit the K-means algorithm with the optimal k (k=4), assigning each observation to its nearest cluster. After assigning cluster labels to each data point, we visualized the clustering results using 2D scatter plots (via seaborn) and 3D scatter plots (via matplotlib) for better clarity and interpretation.

For cluster validation, we calculated the average silhouette score of each cluster using the silhouette_score method from scikit-learn to assess the quality of separation and cohesion among the clusters. Additionally, the WSS values helped us evaluate the compactness of the clusters at different k values to confirm the stability of our choice for the optimal number of clusters.

For implementing clustering, we used:

- pandas for data handling.
- scikit-learn for KMeans and silhouette_score.
- matplotlib and seaborn for visualization.
- numpy for setting the random seed.
- yellowbrick.cluster for visualization SilhouetteVisualizer.

6. Evaluation and Comparison:

- Classification

- **Classification [70% training, 30% testing] Information Gain (Entropy) :**

Figure1.1:(Confusion Matrix):

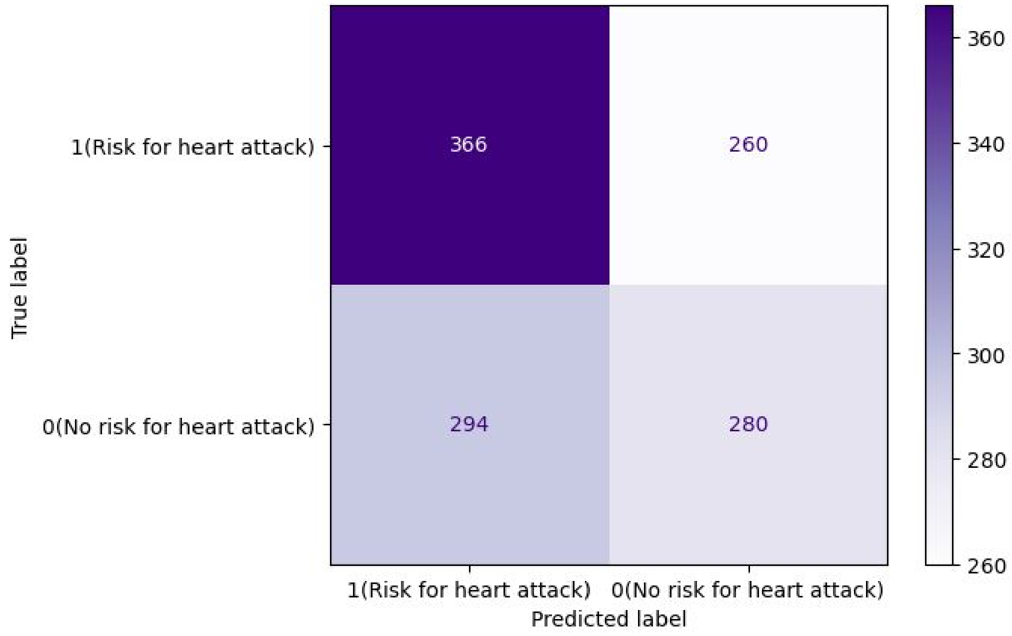
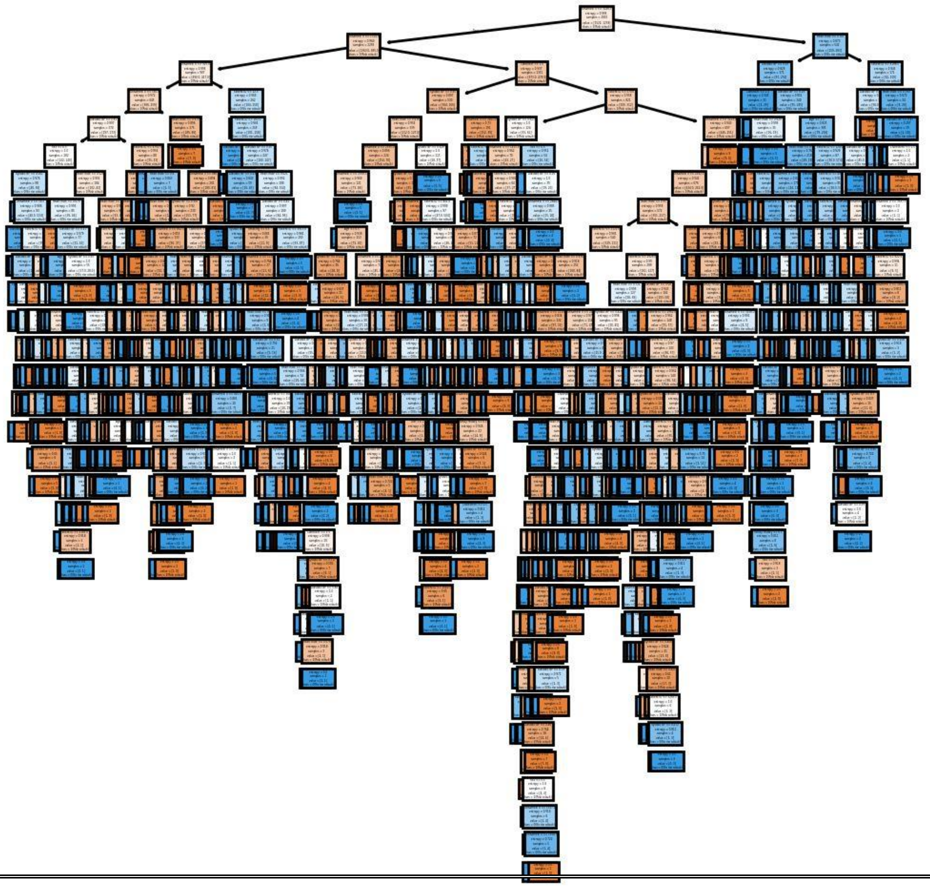


Figure2.1 (Decision Tree) :



- **Classification [60% training, 40% testing] Information Gain (Entropy) :**

Figure1.2:(Confusion Matrix):

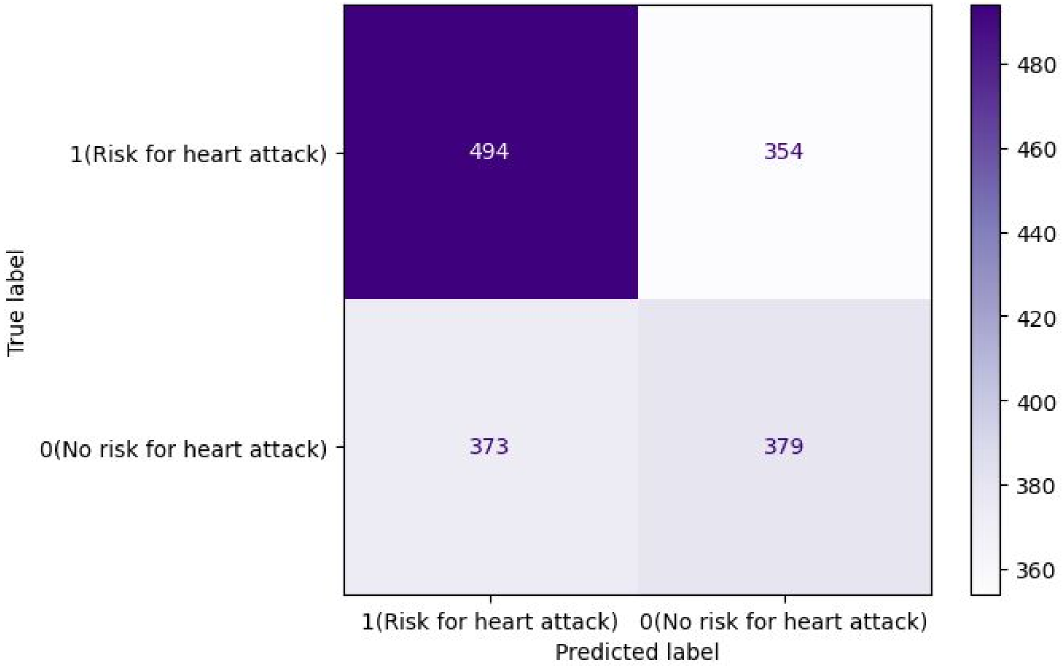
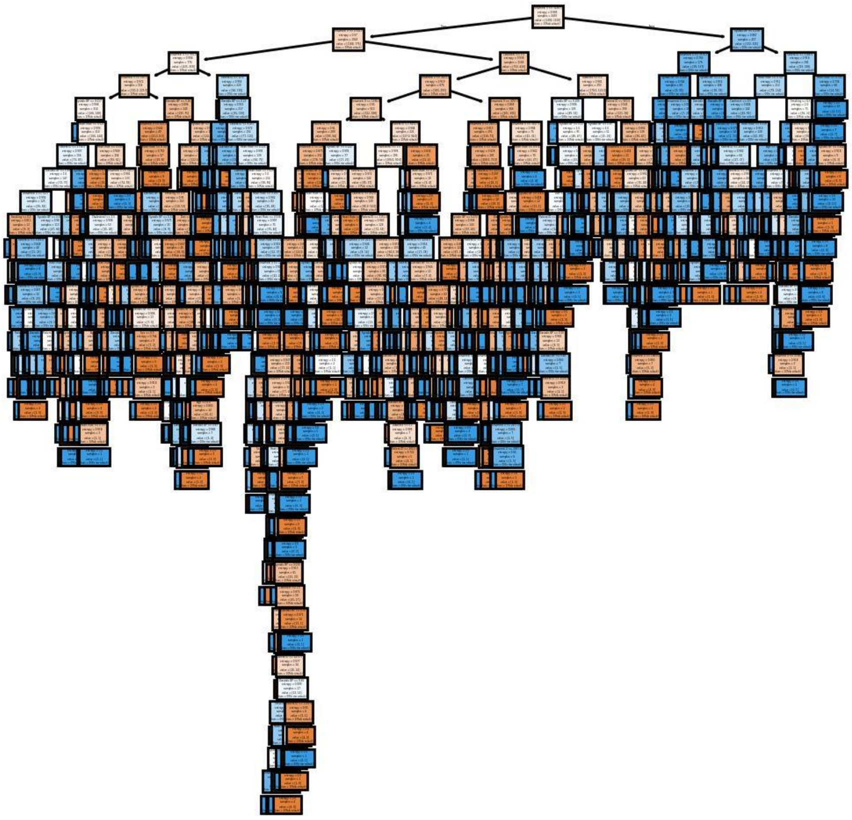


Figure2.2 (Decision Tree) :



- **Classification [80% training, 20% testing] Information Gain (Entropy) :**

Figure1.3:(Confusion Matrix):

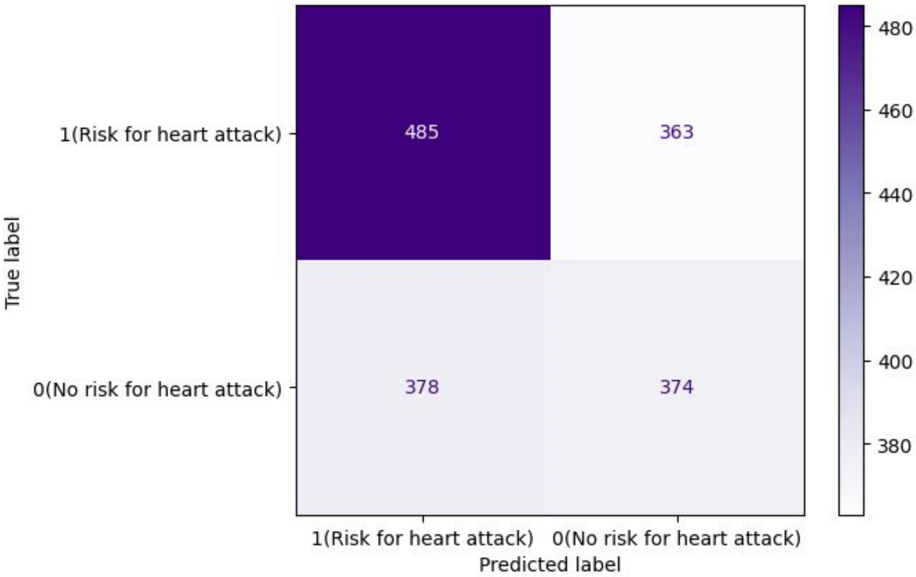
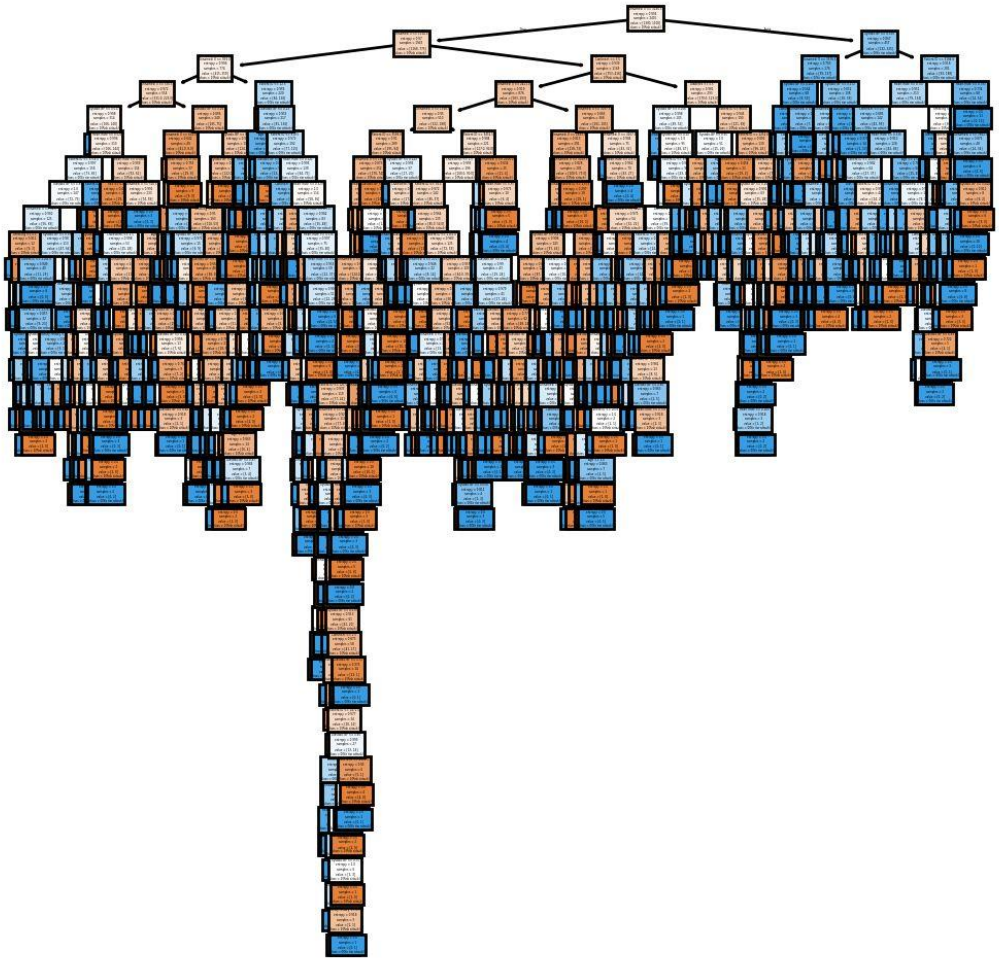


Figure2.3 (Decision Tree) :



- **Comparing the 3 different testing size for data splitting (Information Gain)**

	70% Training, 30% Testing	60% Training, 40% Testing	80% Training, 20% Testing
Accuracy	0.5383	0.5456	0.5369
Error Rate	0.4617	0.4544	0.4631
Sensitivity	0.4878	0.5040	0.4973
Specificity	0.5847	0.5825	0.5719
Precision	0.5185	0.5171	0.5075

- **Classification [70% training, 30% testing] Gini Index :**

Figure 2.1:(Confusion Matrix):

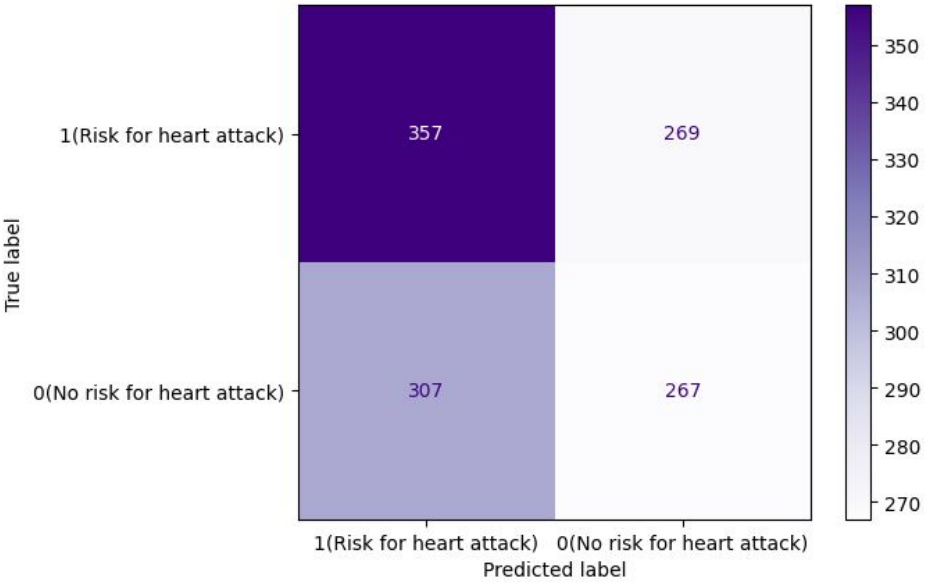
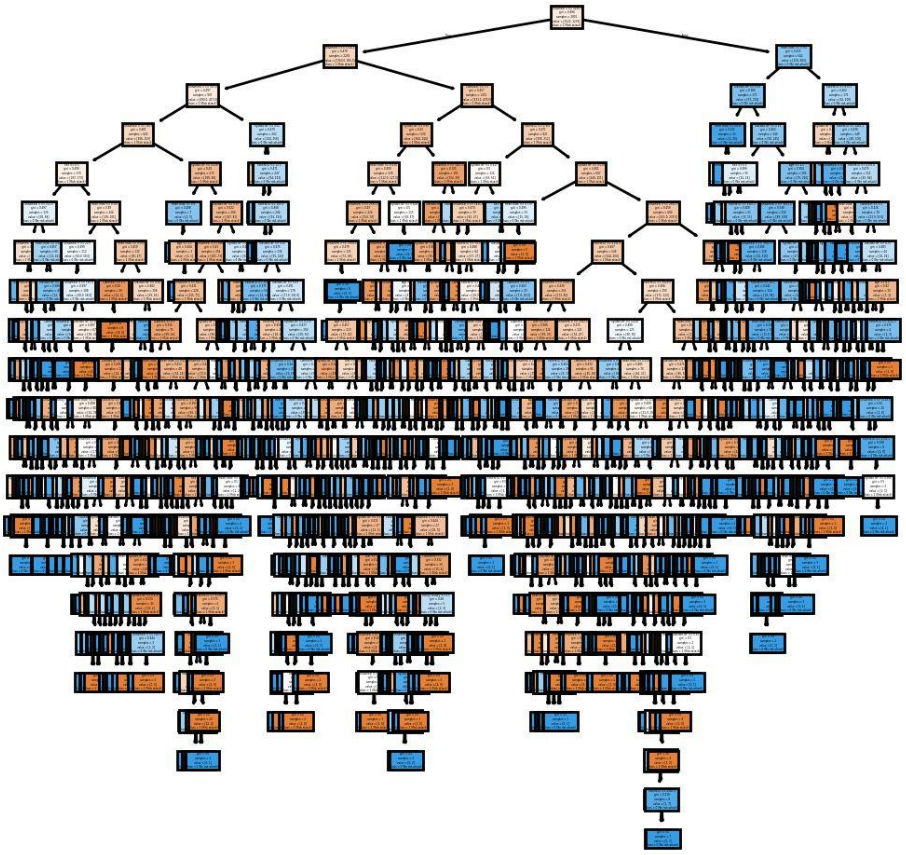


Figure 2.1 (Decision Tree) :



- **Classification [60% training, 40% testing] Gini Index :**

Figure 2.2:(Confusion Matrix):

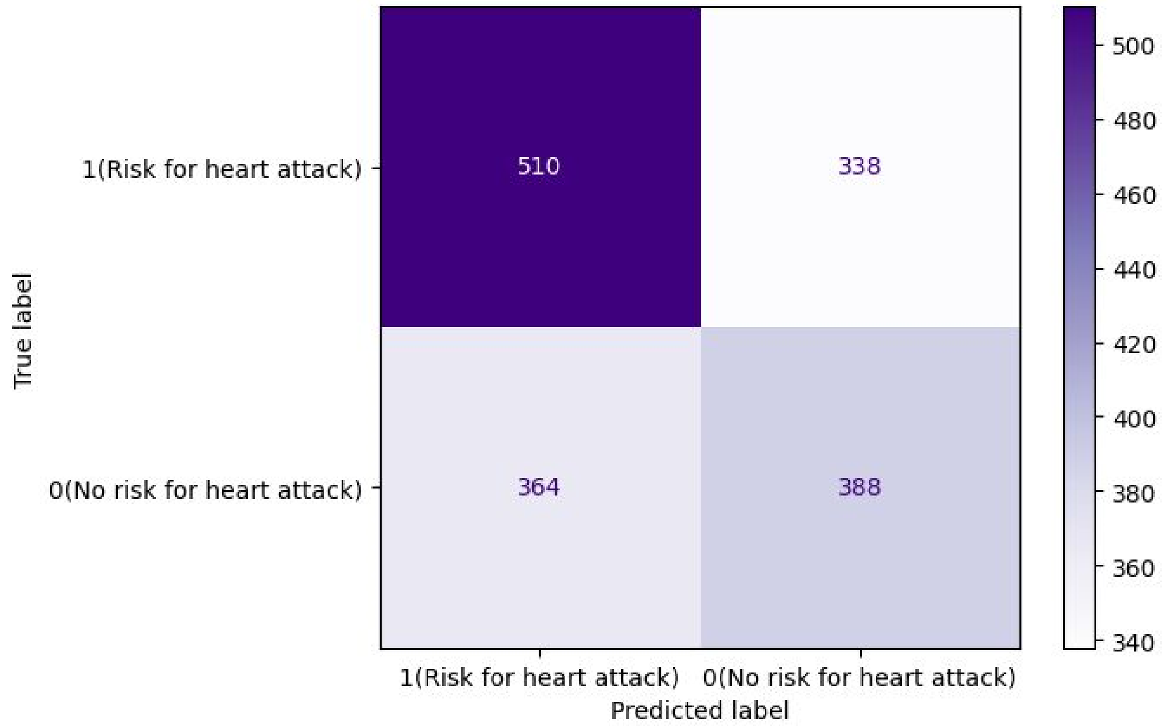
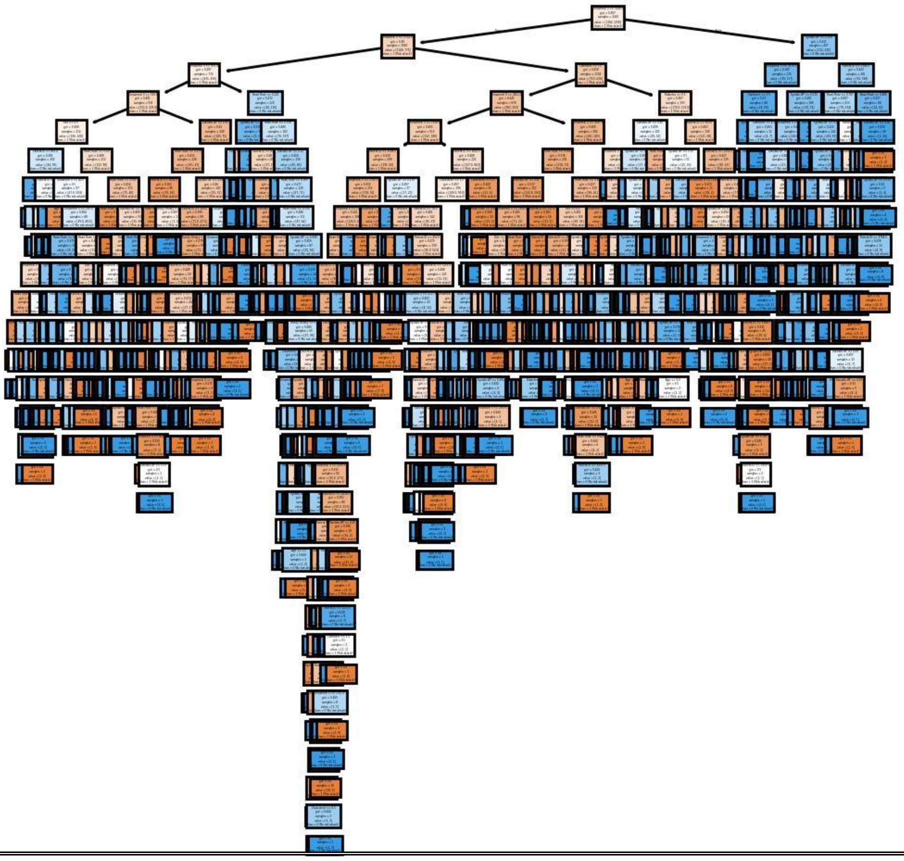


Figure 2.2 (Decision Tree) :



- **Classification [80% training, 20% testing] Gini Index :**

Figure 2.3:(Confusion Matrix):

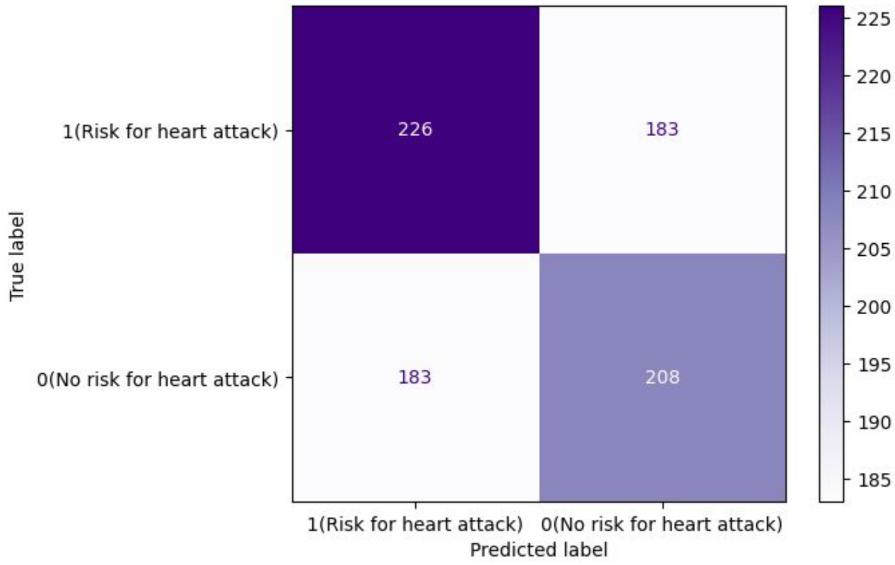
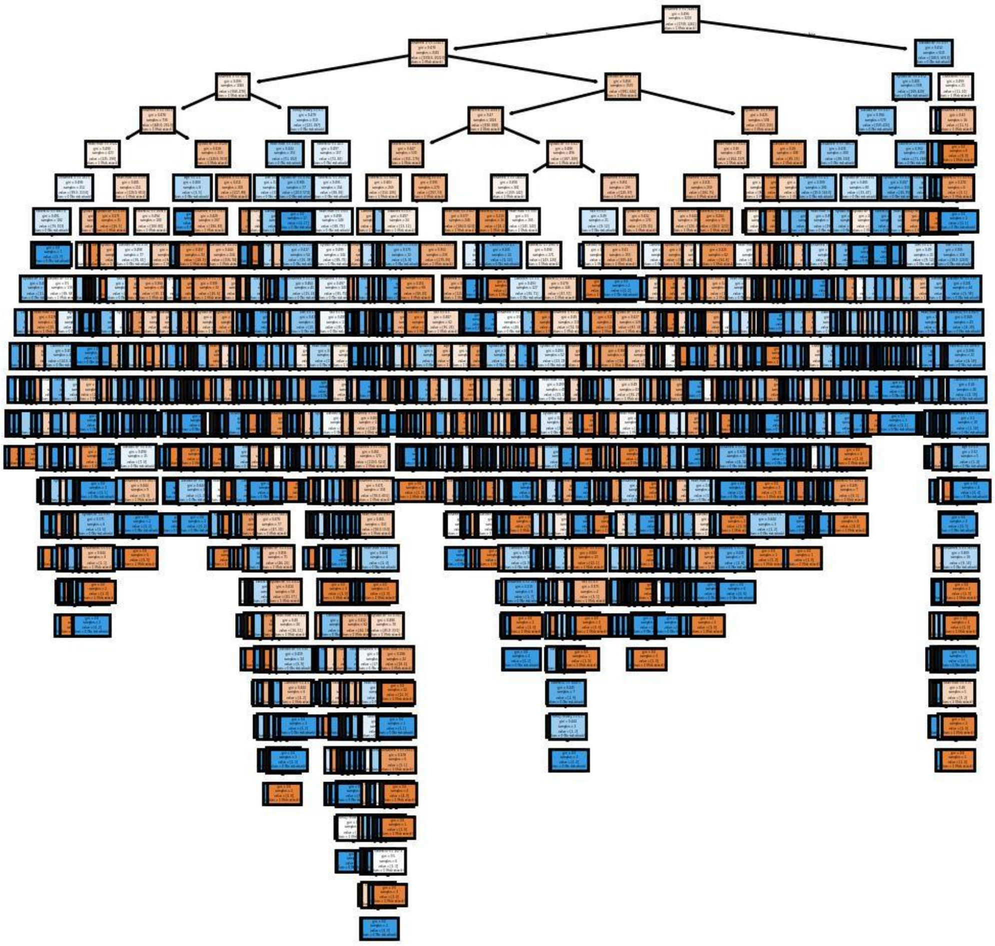
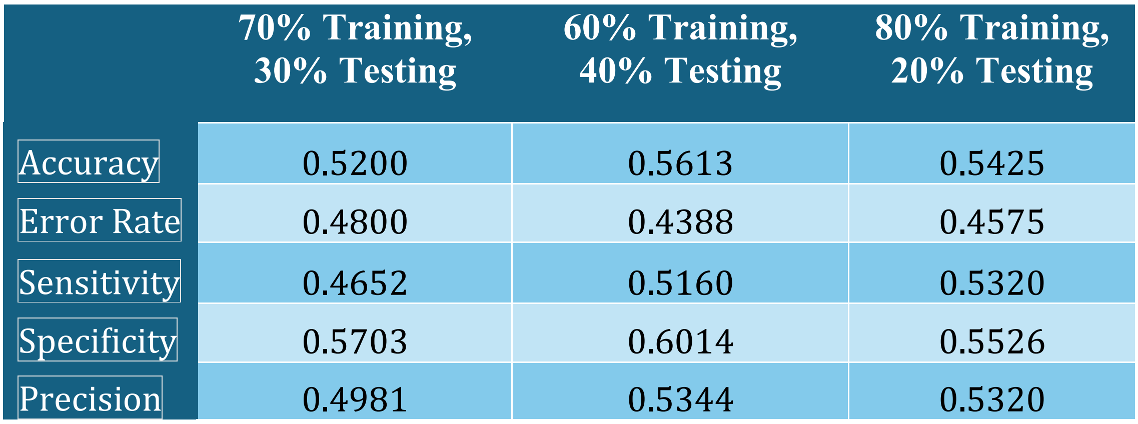


Figure 2.3 (Decision Tree):



- **Comparing the 3 different testing size for data splitting (Gini Index):**

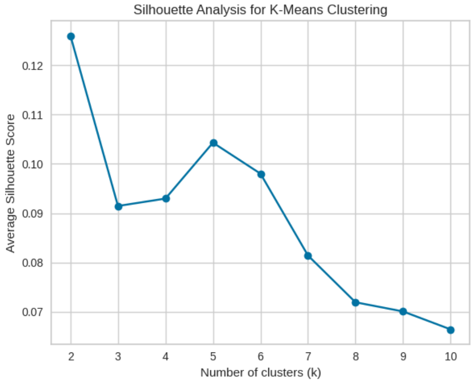


	70% Training, 30% Testing	60% Training, 40% Testing	80% Training, 20% Testing
Accuracy	0.5200	0.5613	0.5425
Error Rate	0.4800	0.4388	0.4575
Sensitivity	0.4652	0.5160	0.5320
Specificity	0.5703	0.6014	0.5526
Precision	0.4981	0.5344	0.5320

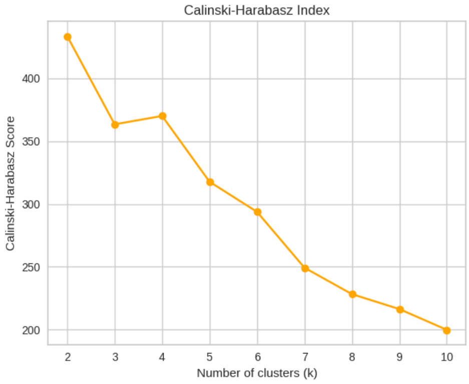
- Clustering

We chose 3 different sizes [2,3,5] based on the result of the validation methods that we used then we will use these sizes to perform the k-means clustering.

Silhouette Analysis: measures how well each data point fits within its assigned cluster compared to neighboring cluster



Calinski-Harabasz: considers both the within-cluster dispersion and the between-cluster dispersion to evaluate the clustering



.

Elbow Method: helps identify the optimal number of clusters for K-means by plotting inertia (cluster compactness) against the number of clusters.

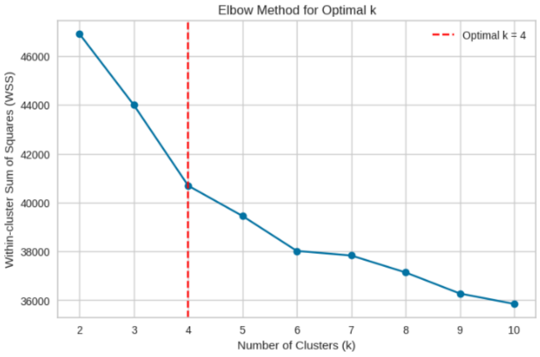


Figure 3.1 : silhouette scores [K=2]

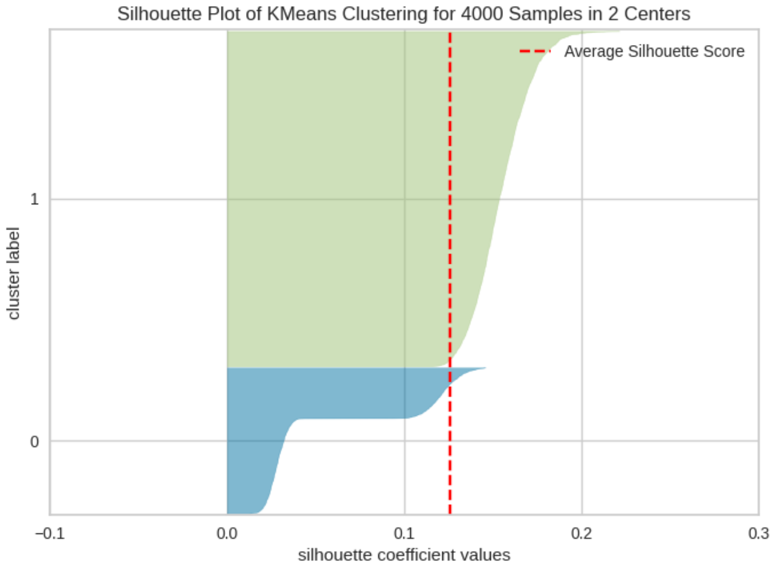


Figure 3.2 : silhouette scores [K=3]

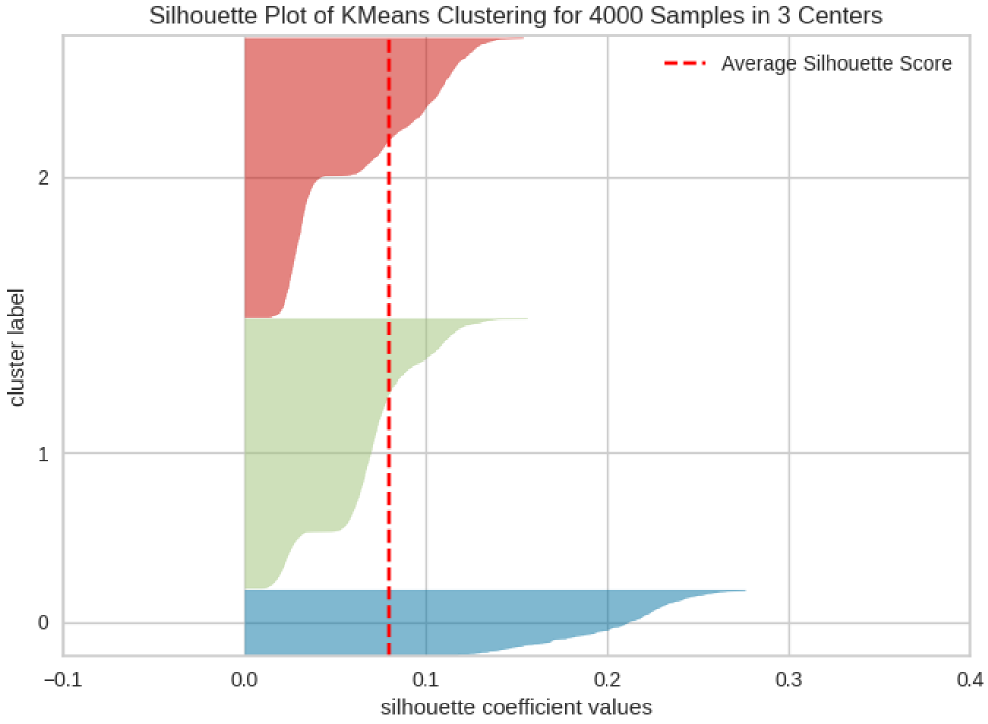
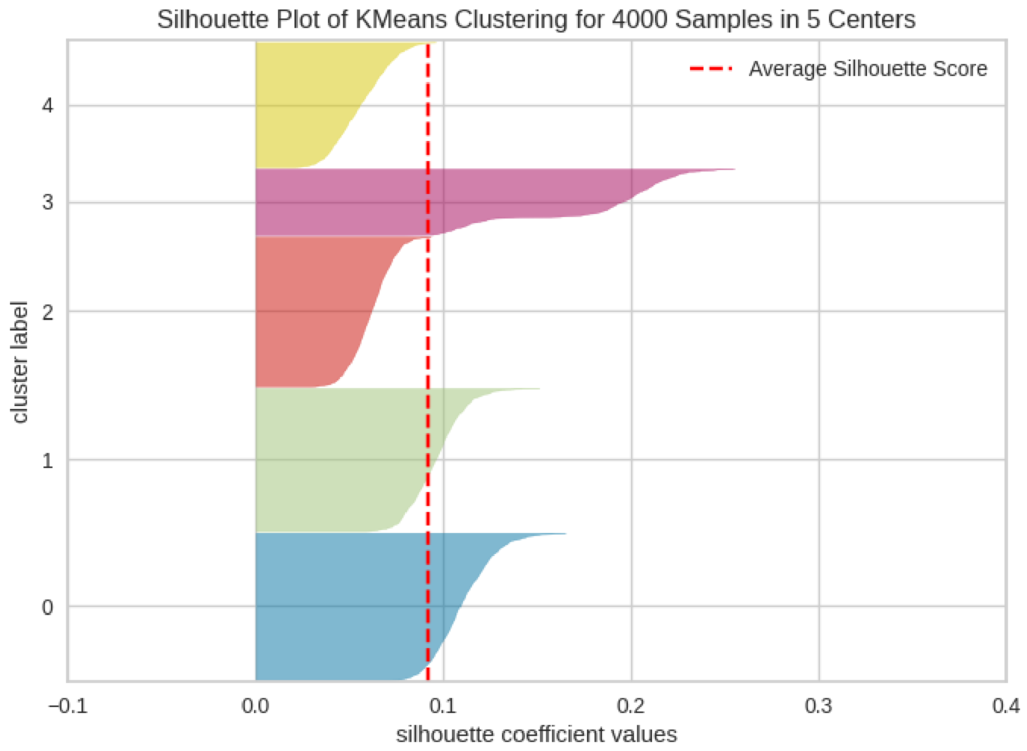


Figure 3.3 : silhouette scores [K=5]:



- **Comparing the 3 different testing size for data splitting (Clustering):**

No.Of Clusters	K=2	K=3	K=5
WSS	46911	43998	39450
Silhouette	0.125	0.091	0.104

7. Findings:

Classification:

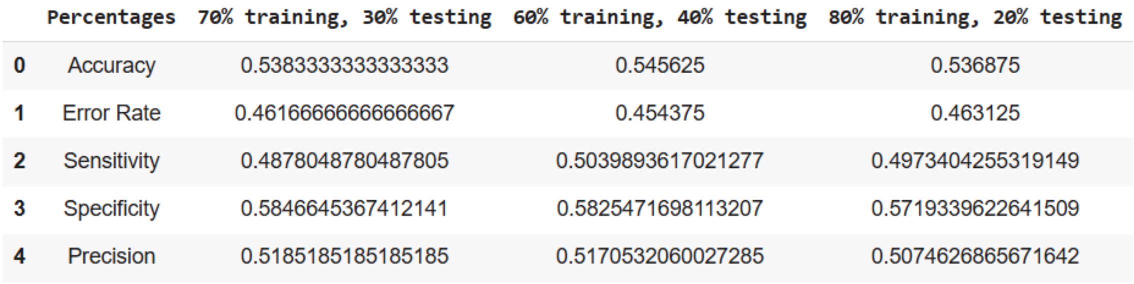
First, our team chose a dataset containing crucial health information about individuals. Our objective was to leverage this data to predict the likelihood of a person's risk of having a heart attack. By doing so, we hope to empower individuals with valuable insights and preventive strategies to enhance their health management and reduce the risk of heart-related issues.

To guarantee precise and trustworthy outcomes, we implemented several preprocessing methods to improve the quality of the dataset. By using these methods, we were able to prepare the data for additional analysis and optimize it. Furthermore, we utilized various visualization techniques to examine the dataset visually, which helped us gain a clearer insight into its features and identify the most suitable preprocessing steps.

After thoroughly examining the dataset visually, including the plots, we checked for any outliers or missing values. We found that there were neither outliers nor missing values that could negatively impact the accuracy of our predictions. With this confirmation, we proceeded to perform data transformation, including normalizing and discretizing certain attributes, ensuring that all features were treated equally, and making the data easier to understand for subsequent analysis. The goal of these actions was to create a predictive model that would be accurate and dependable so that people could make well-informed choices for better living.

After the preprocessing stage, we implemented various methods, including the Gini index, and information gain, in combination with different partitioning techniques. We carefully analyzed the results of each method to determine the best approach for our dataset.

- INFORMATION GAIN:



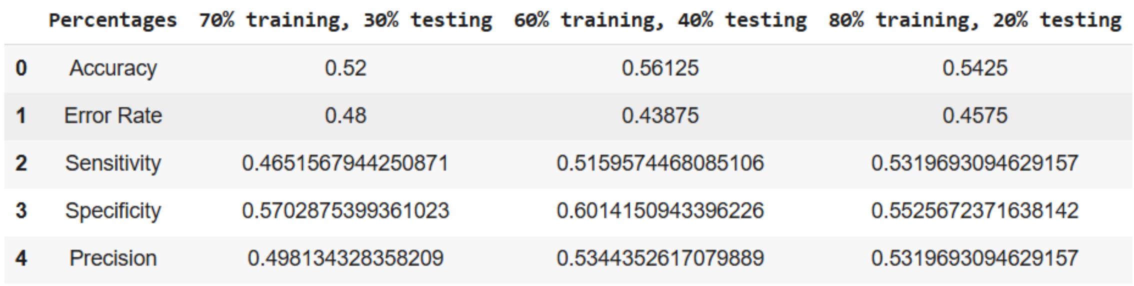
	Percentages	70% training, 30% testing	60% training, 40% testing	80% training, 20% testing
0	Accuracy	0.5383333333333333	0.545625	0.536875
1	Error Rate	0.46166666666666667	0.454375	0.463125
2	Sensitivity	0.4878048780487805	0.5039893617021277	0.4973404255319149
3	Specificity	0.5846645367412141	0.5825471698113207	0.5719339622641509
4	Precision	0.5185185185185185	0.5170532060027285	0.5074626865671642

The Information Gain results show the following model performance across different data splits for training and testing:

- Accuracy: The model trained on a 60% training set and 40% testing set achieved the highest accuracy at 0.545625 (or 54.56%), followed by the model trained on 70% training and 30% testing with an accuracy of 0.538333 (or 53.833%), and the model trained on 8% training and 20% testing with an accuracy of 0.5386 (or 53.68%).
- Error Rate: The model trained on a 80% training set and 20% testing set had the highest error rate at 0.463125 (or 46.31%), followed by the model trained on 70% training and 30% testing with an error rate of 0.46166 (or 46.16%), and the model trained on 60% training and 40% testing with an error rate of 0.454375 (or 45.43%).
- Sensitivity: The model trained on a 60% training set and 40% testing set achieved the highest sensitivity at 0.5039893617021277 (or 50.39%), followed by the model trained on 80% training and 20% testing with 0.4973404255 (or 49.73%), and the model trained on 70% training and 30% testing with 0.4878048780487 (or 48.78%).
- Specificity: The model trained on an 70% training set and 30% testing set obtained the highest specificity at 0.5846645367 (or 58.46%), followed by the model trained on 60% training and 40% testing with 0.582547169 (or 58.25%), and the model trained on 80% training and 20% testing with 0.57193396 (or 57.19%).
- Precision: The model trained on a 70% training set and 30% testing set achieved the highest precision at 0.5185185185185185 (or 51.85%), followed by the model trained on 60% training and 40% testing with 0.51705320 (or 51.70%), and the model trained on 80% training and 20% testing with 0.507462686 (or 50.74%).

Based on the values, the 60% training and 40% testing split appears to be the best choice for the model. This split achieves the highest accuracy (54.56%), which is typically a primary metric for assessing overall model performance. It also achieves the highest sensitivity (50.39%), Also second Specificity (58.25%) and precision (51.70%), meaning it is effective at correctly identifying positive cases and minimizing false positives. Additionally, the error rate for this split (45.43%) is the lowest among the three splits, indicating fewer incorrect predictions. Overall, this balance between sensitivity, specificity, and accuracy suggests it is well-suited for maintaining reliable predictions across different metrics.

- GINI INDEX:



	Percentages	70% training, 30% testing	60% training, 40% testing	80% training, 20% testing
0	Accuracy	0.52	0.56125	0.5425
1	Error Rate	0.48	0.43875	0.4575
2	Sensitivity	0.4651567944250871	0.5159574468085106	0.5319693094629157
3	Specificity	0.5702875399361023	0.6014150943396226	0.5525672371638142
4	Precision	0.498134328358209	0.5344352617079889	0.5319693094629157

Using these four metrics, we can determine which model is better suited for making predictions.

Accuracy: Measures the overall correctness of the model by calculating the ratio of correctly predicted observations to the total observations.

Error Rate: Represents the proportion of incorrect predictions made by the model. Sensitivity (Recall): Reflects the model's ability to correctly identify positive cases (true positives). It focuses on reducing false negatives.

Specificity: Measures the ability of the model to correctly identify negative cases (true negatives). It focuses on reducing false positives.

Precision: Indicates the proportion of positive identifications that were actually correct. It focuses on reducing false positives.

The results show the following model performance across different data splits for training and testing:

- Accuracy: The model trained on a 60% training set and 40% testing set achieved the highest accuracy at 0.56125 (or 56.12%), followed by the model trained on 80% training and 20% testing with an accuracy of 0.5425 (or 54.25%), and the model trained on 70% training and 30% testing with an accuracy of 0.52 (or 52%).
- Error Rate: The model trained on a 70% training set and 30% testing set had the highest error rate at 0.48 (or 48%), followed by the model trained on 80% training and 20% testing with an error rate of 0.4575 (or 45.75%), and the model trained on 60% training and 40% testing with an error rate of 0.43875 (or 43.87%).
- Sensitivity: The model trained on a 80% training set and 20% testing set achieved the highest sensitivity at 0.53196930 (or 53.19%), followed closely by the model trained on 60% training and 40% testing with a sensitivity of 0.515957446 (or 51.59 %), and the model trained on 70% training and 30% testing with a sensitivity of 0.465156794 (or 46.51%).
- Specificity: Specificity remained consistent across all splits, with the model trained on 60% training and 40% testing achieving 0.60141509433 (or 60.14%), followed by 70% training and 30% testing with 0.5702875399 (or 57.02%), and 80% training and 20% testing with 0.5525672371 (or 55.25%).
- Precision: The model trained on an 60% training set and 40% testing set obtained the highest precision at 0.53443526170 (or 53.44%), followed by the model trained on 80% training and 20% testing with 0.531969309 (or 53.19%), and the model trained on 70% training and 30% testing with 0.49813432 (or 49.81%).

Analysis: Based on these metrics, the 60% training and 40% testing split appears to be the best choice for the model. This split achieves the highest accuracy (56.12%), Specificity (60.14%) and Precision (or 53.44%) which are key metrics for assessing overall performance and the ability to correctly identify positive cases.

Furthermore, it has the lowest error rate (43.87%), ensuring fewer incorrect predictions. While the 80%-20% split performs slightly better in Sensitivity, the balance provided by the 60%-40% split across all metrics suggests it is better suited for reliable predictions.

-The best model between information gain and the Gini index:

After selecting the best model split from Information Gain, which was 60% training, 40% testing, and the best split from Gini Index, which was 60% training, 40% testing, we reviewed the values of each for comparison between Information Gain and Gini Index, and we reached the following conclusion:

	Information gain	Gini index
Accuracy	0.545625	0.56125
Error rate	0.454375	0.43875
Sensitivity	0.5039893617021277	0.5159574468085106
Specificity	0.5825471698113207	0.6014150943396226
Precision	0.5170532060027285	0.5344352617079889

Accuracy and Error Rate:

The **Gini Index** split provides a higher accuracy **(56%)** compared to **Information Gain (54%).** This leads to a **lower error rate** for the **Gini Index model (43%)** compared to **Information Gain (45%).** This indicates that the **Gini Index** model performs better in terms of correctly classifying cases, making it a more reliable model for this task.

Sensitivity and Specificity:

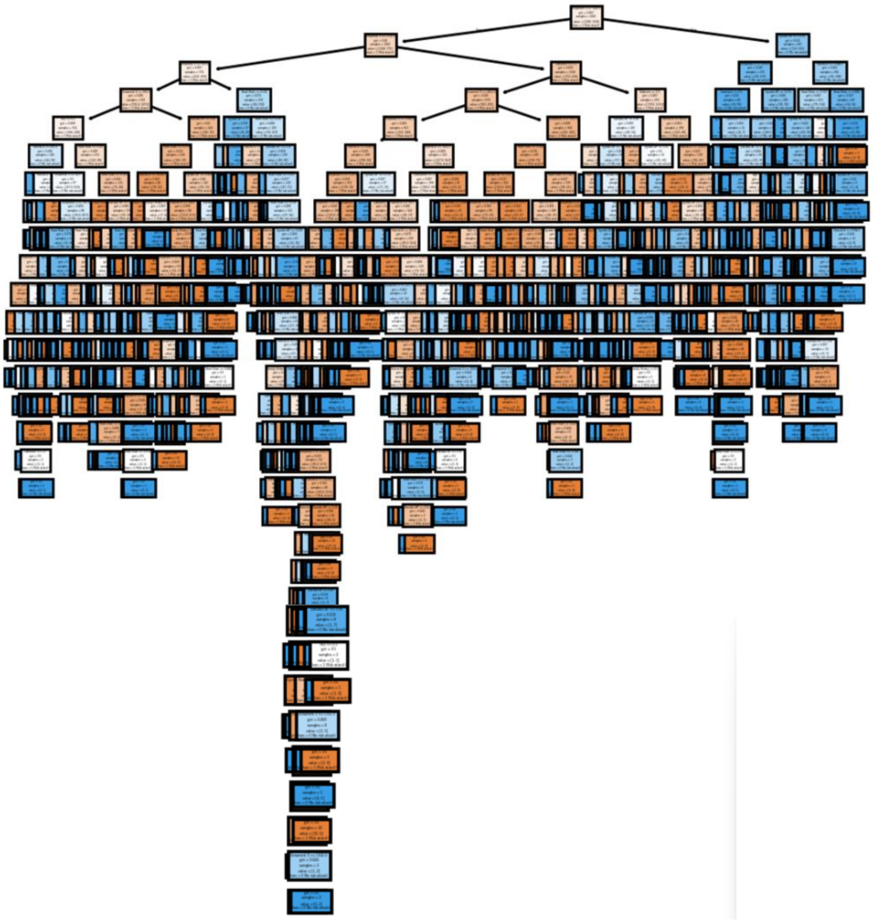
In terms of sensitivity, the Gini Index split slightly outperforms the **Information Gain** split with a sensitivity of **51%** compared to **50%** for I**nformation Gain.** Sensitivity reflects the model's ability to correctly identify positive cases. However, the difference is minimal.

For specificity, the **Gini Index** has a slightly better result **(60%)** compared to **Information Gain** split (58**%**). Specificity refers to the model's ability to correctly identify negative cases. So, **Gini Index** is slightly better at identifying negatives.

Precision:

The **Gini Index** provides a slightly higher **precision** (**53%**) compared to **Information Gain** (**51%**), meaning that when the model predicts a positive case, it is correct **53%** of the time for Gini Index, compared to **51%** for Information Gain.

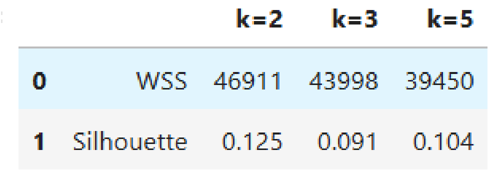
Based on these reasons, it can be concluded that the 60%-40% split using the Gini Index yields better overall performance, with high accuracy, low error rate, and high values for sensitivity, and precision. This was the decision tree associated with this division:



This decision tree classifier, trained using the **Gini index** as the splitting criterion, offers an intuitive way to predict heart attack risk based on various patient features. By using the class_label dictionary maps the class values 1 and 0 to descriptive labels: 1 stands for "Risk attack" and 0 stands for "No risk attack." The tree shows how the model classifies data, with each node representing a decision based on a feature. However, the tree is complex and difficult to read, making it challenging to interpret and use for decision-making.

Clustering:

From the analysis, we applied multiple clustering evaluation techniques to determine the optimal number of clusters (K) for our dataset. We calculated the average silhouette width for each **k,** and we concluded the following results:

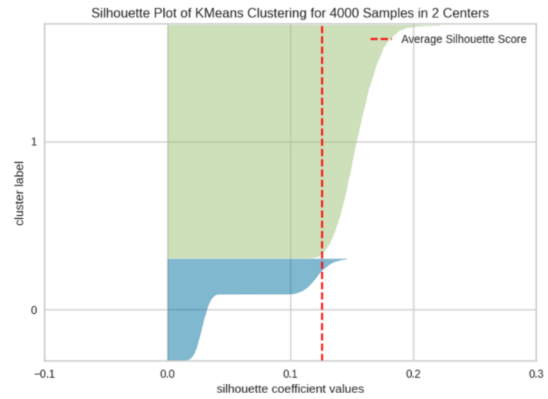


		k=2	k=3	k=5
0	WSS	46911	43998	39450
1	Silhouette	0.125	0.091	0.104

Based on these results, we've decided that K=2 is the best choice for our clustering model based on the metrics we've analyzed (WSS, Average Silhouette Score, Visualization of K-mean). due to K=2 being the highest silhouette width, also having the highest value of WSS.

Also, having a silhouette plot of k-means clustering of 400 samples of 2 centers was one of the most important criteria for choosing k=2 as the best k, indicating that it creates distinct and cohesive clusters.

And this was the corresponding chart:



The graph of K-Means Clustering for 400 samples in 2 centers shows that the majority of silhouette scores are positive, suggesting that the samples are well-aligned with their respective clusters and are sufficiently distant from other clusters. This indicates that the clustering solution has successfully separated the data points into distinct and well-defined clusters.

However, it's important to note that while most of the silhouette scores being positive is a good sign, it doesn't necessarily mean the clustering is "perfect" or without flaws. There may still be some overlap or uncertainty between clusters, especially for samples, where silhouette scores are close to 0 or even negative. This indicates that while clustering is generally effective, there could still be areas of ambiguity or misclassification.

Finally, both models play a valuable role in predicting the likelihood of a person experiencing a heart attack, aiding in our understanding of the contributing factors, such as high blood pressure, cholesterol levels, and lifestyle choices. However, since our dataset includes a "Selector" class (class label) that indicates whether a person is at risk of a heart attack or not, this makes supervised learning models (classification) more accurate and suitable for application than unsupervised learning models (clustering), where the expected outputs are known in advance using this class classification feature.

8. References:

[1] M1Relly, "Heart Attack Prediction," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/m1relly/heart-attack-prediction/data. [Accessed: Nov. 29, 2024].
[2] Fajer Alamro, "IT326-Project," GitHub. [Online]. Available: https://github.com/FajerAlamro/IT326-Project. [Accessed: Nov. 29, 2024].

[3] ”Labs and Lecture Slides”, College of Computer Science, Department of Information Technology, King Saud University. [Accessed: Nov. 29, 2024].

[4] World Health Organization(WHO), “Cardiovascular diseases”. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases. [Accessed: Nov. 29, 2024].