جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

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

IT 326 – Data Mining
1st Semester 1446 H

Heart Failure Dataset

Section No.	Students Name	Student ID
Section #71161 Group #14	Aljawharah Alwabel	444200750
	Ruba Alshammari	444200470
	Norah Alwasil	444201009
	Reema Alotaibi	444200520

Supervised by: TA. Hannan Altamimi

1-Problem

Heart disease often remains undiagnosed until advanced stages, resulting in preventable complications and fatalities. This project addresses the critical need for early risk identification by analyzing key health factors, including age, cholesterol, blood pressure, and exercise-related symptoms. Through predictive modeling and pattern analysis, we aim to support healthcare providers with data-driven insights to enhance early detection, enable timely interventions, and reduce heart disease-related mortality.

2-Data Mining Task

In this project, we will utilize two key data mining techniques—classification and clustering—to predict the likelihood of heart disease.

Classification involves training a model to determine whether a patient has heart disease based on a range of medical factors, including cholesterol levels, blood pressure, age, gender, and others. The classification process will focus on the "heart disease" class to make accurate predictions.

Clustering, on the other hand, will group patients with similar characteristics without considering the heart disease classification. These clusters will help identify patterns and commonalities within the data, offering deeper insights into the factors associated with heart disease and potentially revealing previously unknown relationships.

3-Data

The source of dataset:

https://www.kaggle.com/code/parsalatifi/heart-failure-prediction-95-accuracy-score

-Number of attributes: 12

-No. of objects: 918
-Class label: HeartDisease

To try to understand our data, we reviewed:

- **General information:**

Number of attributes: 12

Number of objects: 918

Class lable: HeartDisease

Attribute types:

```
               Data Types
Age                 int64
Sex                object
ChestPainType      object
RestingBP           int64
Cholesterol         int64
FastingBS           int64
RestingECG         object
MaxHR               int64
ExerciseAngina     object
Oldpeak           float64
ST_Slope           object
HeartDisease        int64
```

- Missing values

```
Missing:

Age                 0
Sex                 0
ChestPainType       0
RestingBP           1
Cholesterol       172
FastingBS           0
RestingECG          0
MaxHR               0
ExerciseAngina      0
Oldpeak             0
ST_Slope            0
HeartDisease        0
dtype: int64

Rows with missing values:
     Age Sex ChestPainType RestingBP Cholesterol  FastingBS RestingECG  MaxHR  \
293   65   M           ASY       115        None          0     Normal     93
294   32   M            TA        95        None          1     Normal    127
295   61   M           ASY       105        None          1     Normal    110
296   50   M           ASY       145        None          1     Normal    139
297   57   M           ASY       110        None          1         ST    131
..   ...  ..           ...       ...         ...        ...        ...    ...
514   43   M           ASY       122        None          0     Normal    120
515   63   M           NAP       130        None          1         ST    160
518   48   M           NAP       102        None          1         ST    110
535   56   M           ASY       130        None          0        LVH    122
536   62   M           NAP       133        None          1         ST    119
```

```
    ExerciseAngina  Oldpeak ST_Slope  HeartDisease
293              Y      0.0     Flat             1
294              N      0.7       Up             1
295              Y      1.5       Up             1
296              Y      0.7     Flat             1
297              Y      1.4       Up             1
..             ...      ...      ...           ...
514              N      0.5       Up             1
515              N      3.0     Flat             0
518              Y      1.0     Down             1
535              Y      1.0     Flat             1
536              Y      1.2     Flat             1
[172 rows x 12 columns]
```

We have 173 values , 172 in Cholesterol and 1 in RestingBP.

- Statical Measures for each numeric column:

-Show Five Number Summary: using summary_stats() function. From these summary statistics, several key observations can be made:

- Age: Patients' ages range from 28 to 77, with a median of 54 and a mean of 53.5, indicating a middle-aged population. Most patients are in the middle-aged to elderly range, as 50% of the patients are aged 54 or older, while 25% are younger than 47.
- RestingBP: The values have a mean of 132.4 and a median of 130. The interquartile range is 120 to 140, showing that most patients have normal blood pressure levels. However, a minimum value of 0 suggests data errors, which will be handled later.
- Cholesterol: The majority of patients (IQR: 173 to 267) fall within normal(<200) to borderline high(>240) cholesterol levels. However, a minimum value of 0 suggests data errors, which will be handled later.
- FastingBS: Most patients (76.7%) have normal fasting blood sugar levels (0), while a small portion (23.3%) have elevated levels (1).
- MaxHR: Maximum heart rates vary between 60 and 202, with a median of 138, indicating diverse cardiovascular performance.
- Oldpeak: The IQR (0 to 1.5) means most patients experience mild to moderate ST depression during exercise, which is a common indicator of ischemia (reduced blood flow due to a blockage or narrowing of blood vessels) or heart disease.
- HeartDisease: The target variable indicates that 55.3% of patients have heart disease, while 44.7% do not, showing a relatively balanced distribution with a slight majority having heart disease.

Calculation used: IQR = Q3 - Q1. It measures the spread of the middle 50% of data, helping to identify inconsistencies and detect outliers. It's also useful for understanding the data's concentration.

```
              Age   RestingBP  Cholesterol   FastingBS       MaxHR  \
count  918.000000  918.000000   918.000000  918.000000  918.000000
mean    53.510893  132.396514   198.799564    0.233115  136.809368
std      9.432617   18.514154   109.384145    0.423046   25.460334
min     28.000000    0.000000     0.000000    0.000000   60.000000
25%     47.000000  120.000000   173.250000    0.000000  120.000000
50%     54.000000  130.000000   223.000000    0.000000  138.000000
75%     60.000000  140.000000   267.000000    0.000000  156.000000
max     77.000000  200.000000   603.000000    1.000000  202.000000

          Oldpeak  HeartDisease
count  918.000000    918.000000
mean     0.887364      0.553377
std      1.066570      0.497414
min     -2.600000      0.000000
25%      0.000000      0.000000
50%      0.600000      1.000000
75%      1.500000      1.000000
max      6.200000      1.000000
```

-Show the Variance:

Variance helps understand the extent of dispersion or scatter of values in each column. As the variance increases, it indicates that the values are more spread out and scattered away from the mean, whereas decreasing variance suggests that the values are less scattered and closer to the mean value. Therefore, our variance results indicate:

- Age: Moderate Variance is 88.97, indicating moderate variability in patient ages, which aligns with the wide age range (28–77).
- RestingBP: A high variance of 342.77 shows high variability in the readings, indicating diverse patient conditions. An outlier of value 0 may inflate this variance.
- Cholesterol: Very high variance (11964.89) which suggests extreme differences in cholesterol levels, mostly due to the outliers (0 values) and the wide range of values (0–603).
- FastingBS: A low variance of 0.18 reflects the imbalanced distribution, where the majority of patients have normal fasting blood sugar levels (0).
- MaxHR: A high variance of 648.23 indicates significant differences in maximum heart rates.
- Oldpeak: Low variance (1.14) means that there's little variation in ST depression during exercise, with most values clustered close together.
- HeartDisease: A variance of 0.25 shows a fairly balanced split between heart disease and no heart disease cases.

When calculating variance for a binary variable, the result will generally be low because the possible values (0 and 1) are close to each other.

Age 88.974254

RestingBP 342.773903

Cholesterol 11964.891079

FastingBS 0.178968

MaxHR 648.228614

Oldpeak 1.137572

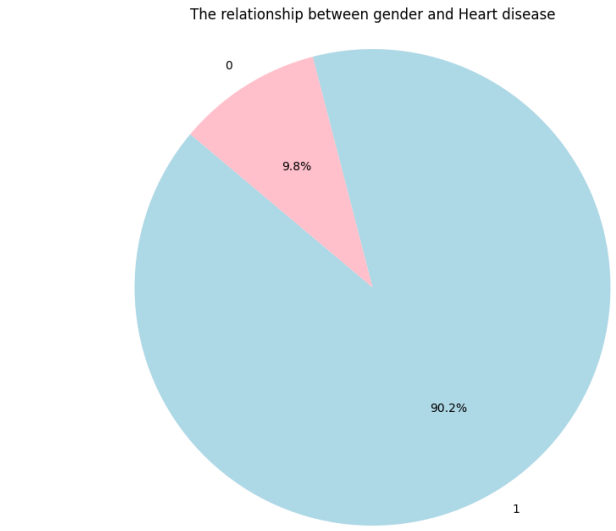
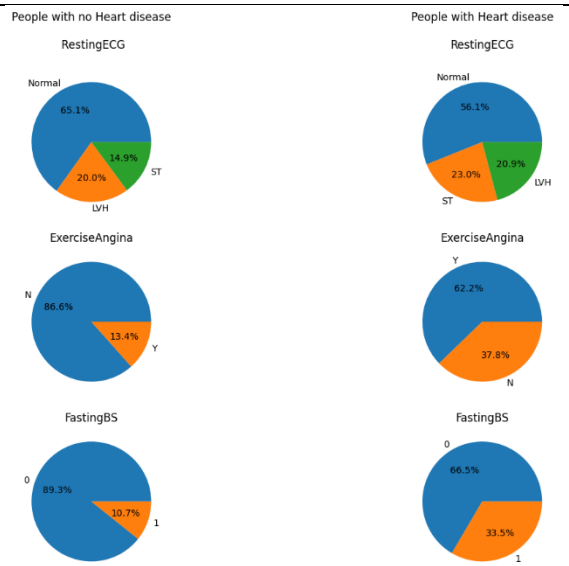
HeartDisease 0.247420

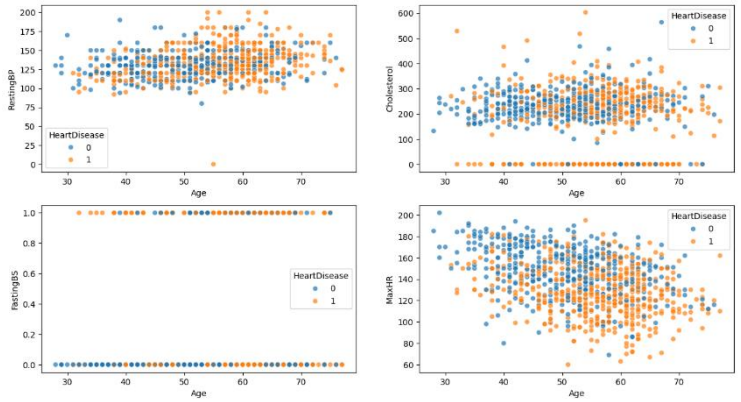
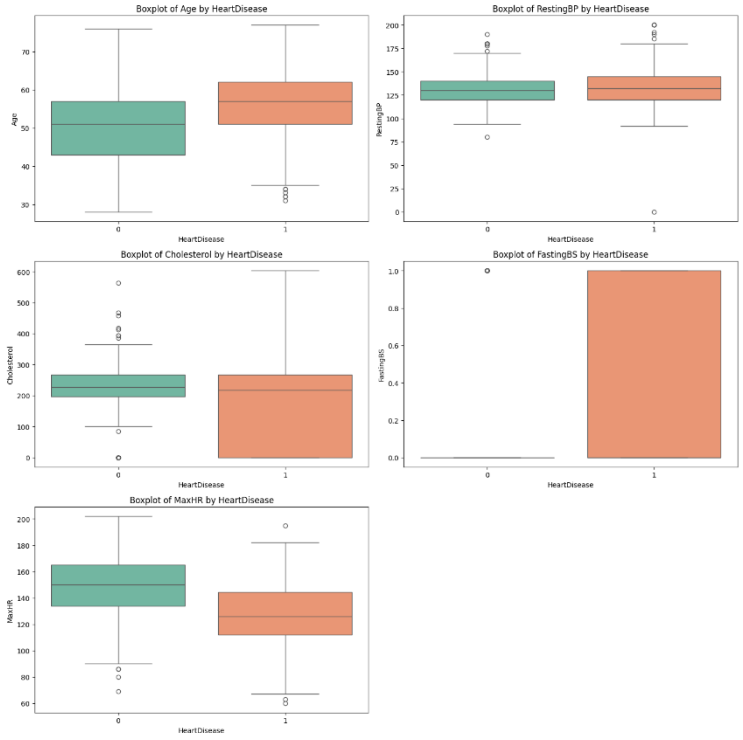
dtype: float64

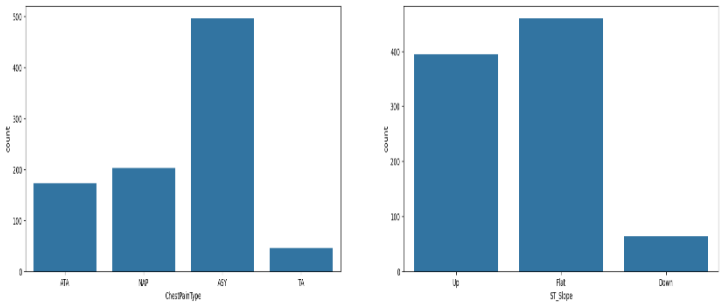
- Understanding the data through graph representations:

Analyzing the Data through Graphical Representations:

To explore the relationship between heart disease and various attributes, visual representations such as graphs are crucial. The "HeartDisease" label, which classifies individuals as affected or unaffected by heart disease, serves as the primary target variable. By examining how this label is associated with other attributes in the dataset, we can extract meaningful relationships and assess whether an increase in certain attributes correlates with a higher likelihood of heart disease. Furthermore, these visual analyses can help reveal if factors like gender and age show significant trends—whether heart disease is more common in men or women and whether age has a positive or negative correlation with the disease. This approach helps identify key risk factors and potential indicators that may contribute to early diagnosis of heart disease

Name of the Graph	Picture of Gragh	Description
Pie Chart		**Analyzing Heart Disease Through Graphs:** Graphs are essential to examine the relationship between heart disease ("HeartDisease" label) and other attributes. By visualizing trends, we can identify correlations, such as how certain factors increase the likelihood of heart disease. Insights into gender and age trends—like prevalence in men vs. women or age correlation—can help highlight key risk factors for early diagnosis.
Pie chart		**Key Patterns in Heart Disease Factors:** • **Fasting Blood Sugar:** Elevated levels are more common in those with heart disease (33.5%) than without (10.7%). • **Resting ECG:** Normal ECG is less frequent in heart disease cases (56.1% vs. 65.1%), with abnormalities like ST changes more prevalent. • **Exercise Angina:** Heart disease patients experience exercise-induced angina more often (62.2% vs. 13.4%). These graphs reveal significant links between heart disease and these health factors.

scatter plot		**Key Indicators for Heart Disease:** • **MaxHR & FastingBS:** Lower MaxHR and elevated FastingBS are strong predictors across all ages. • **Cholesterol & RestingBP:** High cholesterol and RestingBP are more critical in older age groups, especially 50–70. • **FastingBS:** Consistently linked to heart disease, regardless of age or other factors.
boxplot		**Heart Disease Insights:** 1. **Age:** Heart disease patients are older (median 58) vs. non-patients (median 52). 2. **RestingBP:** Wider range and more outliers in heart disease cases, linked to risk. 3. **Cholesterol:** Higher levels in heart disease patients suggest it as a risk factor. 4. **Fasting Blood Sugar:** Elevated in heart disease, low in non-patients. 5. **MaxHR:** Lower median (140 bpm) and more variability in heart disease vs. higher, concentrated values (155 bpm) in non-patients.

barchart		This graph displays two count plots, representing the distribution of ChestPainType **and** ST_Slope **from** the heart disease dataset. The ChestPainType plot shows that most people have the ASY (asymptomatic) type, which means they do **not** experience noticeable symptoms of chest pain. For the ST_Slope plot, the majority of people have a Flat ST segment, which could indicate a risk of heart issues. These findings suggest that asymptomatic chest pain **and** a flat ST slope are common among the individuals **in** the dataset, both of which could be associated **with** heart disease.

4-Data Preprocessing

Missing Values

The dataset contains 173 missing value.

```
Missing:  Age                 0
Sex                 0
ChestPainType       0
RestingBP           1
Cholesterol       172
FastingBS           0
RestingECG          0
MaxHR               0
ExerciseAngina      0
Oldpeak             0
ST_Slope            0
HeartDisease        0
dtype: int64
```

```
Missing:

Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
```

```
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

Description:

Null and missing values can badly affect the efficiency of the dataset and the information that can be extracted from the data later, thus we checked if our data contained missing or null values and we handled these missing values by calculating the mean value for the target columns which is Cholesterol and RestingBP column, and then wereplace the missing values with the resulting mean. to get more efficient dataset.

● Detecting and removing the outliers:

```
Outlier Counts:
Age: 0 rows with outliers
RestingBP: 28 rows with outliers
Cholesterol: 183 rows with outliers
MaxHR: 2 rows with outliers
Oldpeak: 16 rows with outliers
Total Rows with Outliers: 229
```

```
Outlier Counts:
Age: 0 rows with outliers
RestingBP: 0 rows with outliers
Cholesterol: 0 rows with outliers
MaxHR: 0 rows with outliers
Oldpeak: 0 rows with outliers
Total Rows with Outliers: 0
```

Description:

As shown by the previous code, the total number of outliers is 229, indicating a significant presence of extreme values. To handle this, we chose to **cap** the outliers instead of removing them, **replacing them with the nearest non-outlier values**. This approach retains the full dataset while reducing the impact of extreme values, preserving valuable information for our analysis, The large number of outliers in the dataset suggests the existence of extreme values that differ considerably from the main cluster of data points. These outliers represent data points that fall far outside the expected range of values, indicating substantial deviations from the typical patterns observed. Their presence may be due to various factors, such as rare occurrences, data entry errors, or inherent variability in the population being studied.

● Data Transformation

1. Encoding:

```
Data after Label Encoding:
   Age  Sex  ChestPainType  RestingBP  Cholesterol  FastingBS  RestingECG  \
0   40    1              1        140        289.0          0           1
1   49    0              2        160        180.0          0           1
2   37    1              1        130        283.0          0           2
3   48    0              0        138        214.0          0           1
4   54    1              2        150        195.0          0           1

   MaxHR  ExerciseAngina  Oldpeak  ST_Slope  HeartDisease
0    172               0      0.0         2             0
1    156               0      1.0         1             1
2     98               0      0.0         2             0
3    108               1      1.5         1             1
4    122               0      0.0         2             0
```

Description:

we encode the columns as following for better handling:

1. **Sex** (Gender of the patient):
    a. Before encoding: Categories like "Male", "Female".
    b. After encoding:
        i. 0 = Female
        ii. 1 = Male
2. **ChestPainType** (Type of chest pain the patient experiences):
    a. Before encoding: Categories like "TA" (Typical Angina), "ATA" (Atypical Angina), "NAP" (Non-Anginal Pain), "ASY" (Asymptomatic).
    b. After encoding:
        i. 0 = ASY
        ii. 1 = ATA
        iii. 2 = NAP
        iv. 3 = TA
3. **RestingECG** (Resting electrocardiogram results):
    a. Before encoding: Categories like "Normal", "ST" (ST-T wave abnormality), "LVH" (Left Ventricular Hypertrophy).
    b. After encoding:
        i. 0 = LVH
        ii. 1 = Normal
        iii. 2 = ST
4. **ST_Slope** (Slope of the peak exercise ST segment):
    a. Before encoding: Categories like "Up", "Flat", "Down".
    b. After encoding:
        i. 0 = Down
        ii. 1 = Flat
        iii. 2 = Up
5. **ExerciseAngina** (Angina during exercise):
    a. Before encoding: Categories like "Yes", "No".

b. After encoding:
   i. 0 = No
   ii. 1 = Yes

2. Normalization:

```
DataFrame after Decimal Scaling Normalization:
      Age  Sex  ChestPainType  RestingBP  Cholesterol  FastingBS  RestingECG  \
0    0.40    1              1      0.140        0.289          0           1
1    0.49    0              2      0.160        0.180          0           1
2    0.37    1              1      0.130        0.283          0           2
3    0.48    0              0      0.138        0.214          0           1
4    0.54    1              2      0.150        0.195          0           1
..    ...  ...            ...        ...          ...        ...         ...
913  0.45    1              3      0.110        0.264          0           1
914  0.68    1              0      0.144        0.193          1           1
915  0.57    1              0      0.130        0.131          0           1
916  0.57    0              1      0.130        0.236          0           0
917  0.38    1              2      0.138        0.175          0           1

     MaxHR  ExerciseAngina  Oldpeak  ST_Slope  HeartDisease
0    0.172               0     0.00         2             0
1    0.156               0     0.10         1             1
2    0.098               0     0.00         2             0
3    0.108               1     0.15         1             1
4    0.122               0     0.00         2             0
..     ...             ...      ...       ...           ...
913  0.132               0     0.12         1             1
914  0.141               0     0.34         1             1
915  0.115               1     0.12         1             1
916  0.174               0     0.00         1             1
917  0.173               0     0.00         2             0
[918 rows x 12 columns]
```

Description:

Here in the Normalization process we chose Decimal scaling, we normalize the attributes and unify their scale since the range for each attribute is quite different. This ensures that all the features have comparable ranges, preventing attributes with larger values from dominating others. This method helps us to format all the values in the dataset, making them more consistent and easier for identify and assess heart disease risk factors effectively.

3. Aggregation

Sex	HeartDisease	Age	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope
0	0	0.512028	1.237762	0.128580	0.246087	0.069930	0.916084	0.149049	0.111888	0.043566	1.720280
	1	0.561800	0.380000	0.140160	0.229838	0.320000	0.860000	0.137820	0.540000	0.127700	1.020000
1	0	0.502022	1.168539	0.130738	0.218356	0.127341	0.966292	0.147670	0.146067	0.039157	1.749064
	1	0.558690	0.456332	0.133164	0.179639	0.336245	1.039301	0.126566	0.631004	0.125568	1.061135

In the aggregation method, we grouped the "sex" and "HeartDisease" columns and applied an aggregation function (in this case, "mean") to the data. This step helps us to analyze how the mean values of different attributes vary between male and female patients who are either have heartDisease or not . By aggregating the data in this way, we can identify any patterns or differences in attribute means based on sex and hearDisease status. This analysis can provide valuable insights into potential correlations or associations between these variables and help in making informed decisions or drawing conclusions in subsequent analyses

4. Discretization:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS
0	Young	1	1	Normal	High	0
1	Young	0	2	High	Low	0
2	Young	1	1	Low	High	0
3	Young	0	0	Normal	Moderate	0
4	Middle-Aged	1	2	High	Low	0
..	...	...	...	...	...	...
913	Young	1	3	Very Low	High	0
914	Old	1	0	High	Low	1
915	Middle-Aged	1	0	Low	Low	0
916	Middle-Aged	0	1	Low	Moderate	0
917	Young	1	2	Normal	Low	0

	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope
0	1	High	0	No Depression	2
1	1	High	0	Mild Depression	1
2	2	Low	0	No Depression	2
3	1	Low	1	Severe Depression	1
4	1	Low	0	No Depression	2
..	...	...	...	...	...
913	1	Moderate	0	Mild Depression	1
914	1	Moderate	0	Severe Depression	1
915	1	Low	1	Mild Depression	1
916	0	High	0	No Depression	1
917	1	High	0	No Depression	2

```
   HeartDisease
0               0
1               1
2               0
3               1
4               0
..            ...
913             1
914             1
915             1
916             1
917             0

[918 rows x 12 columns]
```

In this discretization process, we categorized the attributes as follows:

1. Age:
    a. Categories:
        i. 28-50 years: Young
        ii. 50-58 years: Middle-Aged 58+years: Old
    b. Reason for Discretization:
    To simplify the analysis of age's impact on heart disease risk.
2. Resting Blood Pressure:
    a. Categories:
        i. Very Low: 0-120 mmHg
        ii. Low: 120-130 mmHg
        iii. Normal: 130-140 mmHg
        iv. High: 140-200 mmHg
    b. Reason for Discretization:
    To differentiate between blood pressure levels and their effects on heart health.
3. Cholesterol Levels:
    a. Categories:
        i. Low: 0-197 mg/dL
        ii. Moderate: 197-250 mg/dL
        iii. High: 250-603 mg/dL
    b. Reason for Discretization:
    To facilitate comparisons of cholesterol levels and their relation to heart disease risk.
4. Maximum Heart Rate:
    a. Categories:
        i. Low: 60-125 bpm
        ii. Moderate: 125-150 bpm
        iii. High: 150-202 bpm
    b. Reason for Discretization:
    To identify the impact of fitness levels on heart health.
5. Oldpeak (Depression Level):
    a. Categories:
        i. No Depression: -2.6 to 0
        ii. Mild Depression: 0 to 1.2
        iii. Severe Depression: 1.2 to 6.2

b. Reason for Discretization:
To distinguish between cardiac depression levels based on exercise performance

We utilized these simplifications to enhance the interpretability and analysis of the data, making it easier to identify and assess heart disease risk factors effectively

5. Feature selection:

Chi-Square:

We need to focus on categorical attributes to apply the Chi-square test and determine uncorrelated attributes with respect to the class label "HeartDisease". The categorical attribute in our dataset after applying all data transformation methods is "Age".

```
Column: Age
Chi-square Statistic: 65.08090035638334
Degrees of Freedom: 2
Expected:
[[140.23965142 173.76034858]
 [128.62745098 159.37254902]
 [141.1328976  174.8671024 ]]

Column: RestingBP
Chi-square Statistic: 20.97630037410108
Degrees of Freedom: 3
Expected:
[[ 98.25708061 121.74291939]
 [ 89.32461874 110.67538126]
 [ 91.5577342  113.4422658 ]
 [130.86056645 162.13943355]]

Column: Cholesterol
Chi-square Statistic: 30.782274715683553
Degrees of Freedom: 2
Expected:
[[136.22004357 168.77995643]
 [138.00653595 170.99346405]
 [135.77342048 168.22657952]]

Column: MaxHR
Chi-square Statistic: 146.58661140651304
Degrees of Freedom: 2
Expected:
[[124.16122004 153.83877996]
 [144.25925926 178.74074074]
 [141.5795207  175.4204793 ]]

Column: Oldpeak
Chi-square Statistic: 148.50725792086038
Degrees of Freedom: 2
Expected:
```

```
[[105.8496732  131.1503268 ]
 [170.16339869 210.83660131]
 [133.9869281  166.0130719 ]
```

Correlation Coefficient:

```
Correlation Coefficient:
Sex: 0.30544491596313866
ChestPainType: -0.38682769426256153
FastingBS: 0.2672911861103007
ExerciseAngina: 0.4942819918242627
ST_Slope : -0.5587707148497031
RestingECG  : 0.057384357013450675
```

Feature selection:

Correlation Coefficient:

- ST_Slope (-0.56): The strongest negative correlation with heart disease, suggests that a steeper decline in the ST segment, is linked to a higher likelihood of heart disease.
- ExerciseAngina (0.49): The strongest positive correlation, suggesting that patients who experience angina during exercise are associated with the likelihood of heart disease.
- ChestPainType (-0.39): Moderate negative correlation, meaning a certain types of chest pain are inversely associated with the likelihood of heart disease.
- Sex (0.31): Moderate positive correlation, suggesting that one of the sexes is more likely to have heart disease compared to the other.
- FastingBS (0.27): Weak positive correlation, indicating that patients with elevated fasting blood sugar are slightly more likely to have heart disease.
- RestingECG (0.06): Very weak correlation, meaning that the resting electrocardiogram results have almost no linear relationship with the presence of heart disease.

In conclusion, and based on the results, we decided to **delete the RestingECG** column due to its weak correlation (0.06) with heart disease.

Chi-square:
After reviewing the probability (alpha) table, we chose the significance level of 0.05; therefore, the critical value is 5.991. When comparing the Chi-square statistics to the critical value, we see that the Chi-square statistics for every attribute in the test is greater than the critical value (65 > 5.9917), (21 > 5.9917), (31 > 5.9917), (147 > 5.9917), (149 > 5.9917).

This means that all the categorical attributes provide valuable information for predicting whether or not a patient has heart disease. Therefore, the correlation is stronger in this case, leading us to **keep all of the categorical attributes**

5-Data Mining Technique

We applied both supervised and unsupervised learning methods to our dataset using classification and clustering techniques.

Classification: For classification, we utilized a supervised learning approach to predict whether an individual has heart disease. The dataset was divided into training and testing subsets, enabling us to train the model on one subset and evaluate its performance on the other using metrics such as accuracy, sensitivity, specificity, and precision.

To enhance interpretability, we implemented a decision tree using the Python library `scikit-learn`. This method was chosen for its simplicity in visualizing and interpreting decisions. Each leaf node of the decision tree represents whether an individual is likely to have heart disease based on attributes such as sex, age, chest pain type, cholesterol level, resting blood pressure, maximum heart rate, exercise-induced angina, ST depression (Oldpeak), and ST segment slope (ST_Slope).

To optimize the model, we experimented with two attribute selection measures—Entropy and Gini Index—and tested three data partitioning schemes (70/30, 80/20, and 60/40). Comparing performance results across these configurations allowed us to determine the optimal combination for our dataset.

Clustering: For clustering, we adopted an unsupervised learning approach using the `K-means` algorithm to group similar data points. This algorithm iteratively assigns data points to the nearest cluster centroids and refines the centroids with each iteration. We selected `K-means` due to its efficiency and effectiveness, even with large datasets.

The `KMeans` class from the `scikit-learn` library was used to implement the algorithm. Excluding the target variable "heart disease," all other attributes were included in the clustering process to explore patterns without being influenced by the classification label.

To evaluate the quality of clustering, we calculated the average silhouette score for each configuration. Additionally, we applied the Within-Cluster Sum of Squares (WSS) method to compare cluster sizes for three configurations (2, 3, and 4 clusters). This analysis helped us identify the optimal number of clusters by balancing separation and compactness.

6-Evaluation and Comparison

- Classification

- Classification [70% training, 30% testing] Information Gain:

Figure (1) (decision tree):

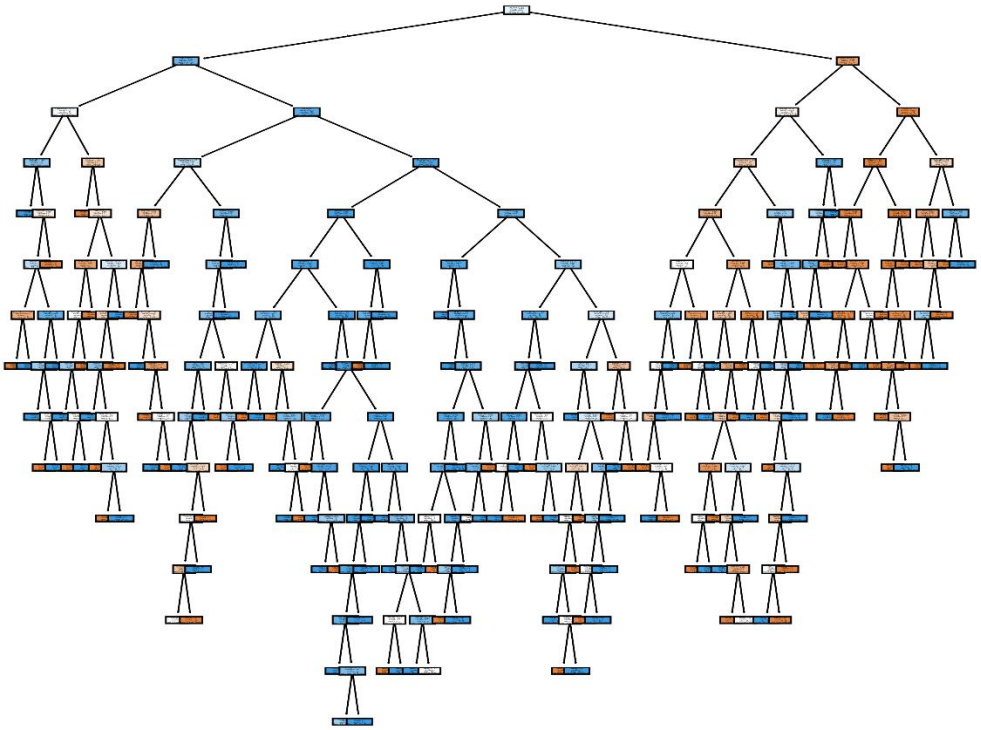
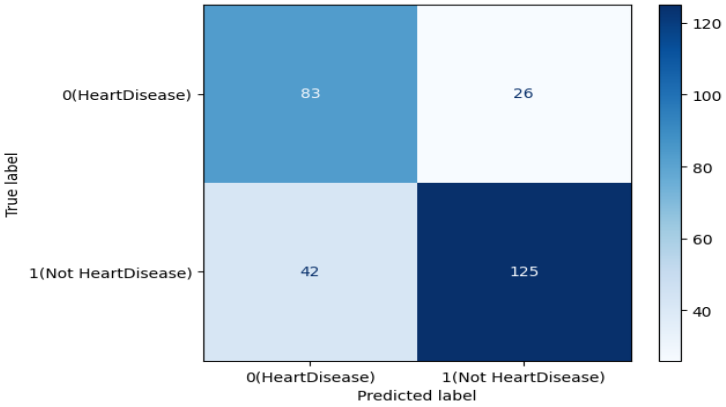


Figure (2) (confusion matrix):



- Classification [60% training, 40% testing] Information Gain:

Figure (1) (decision tree):

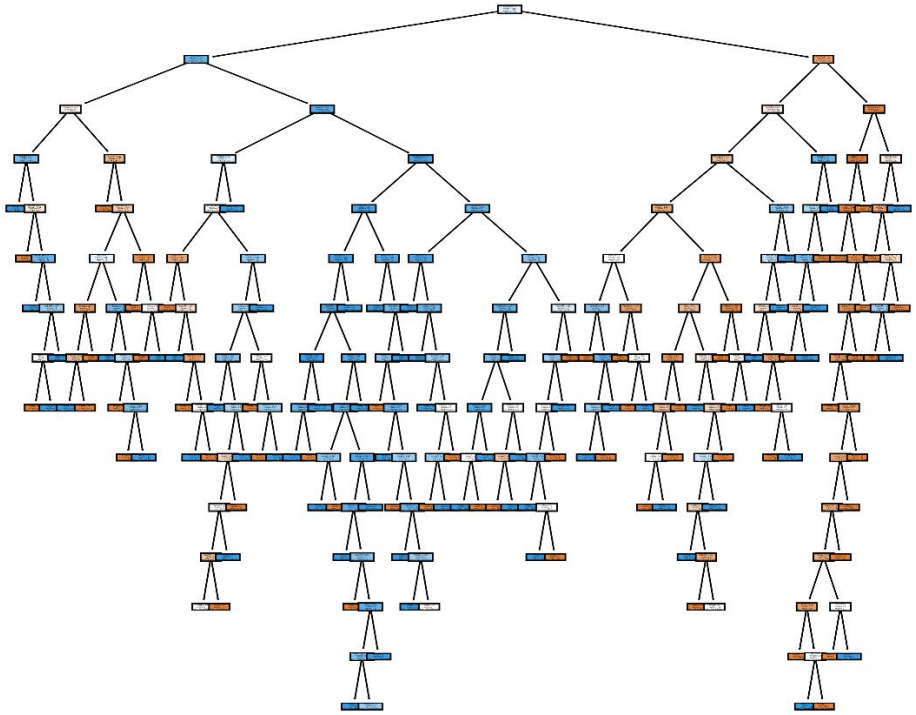
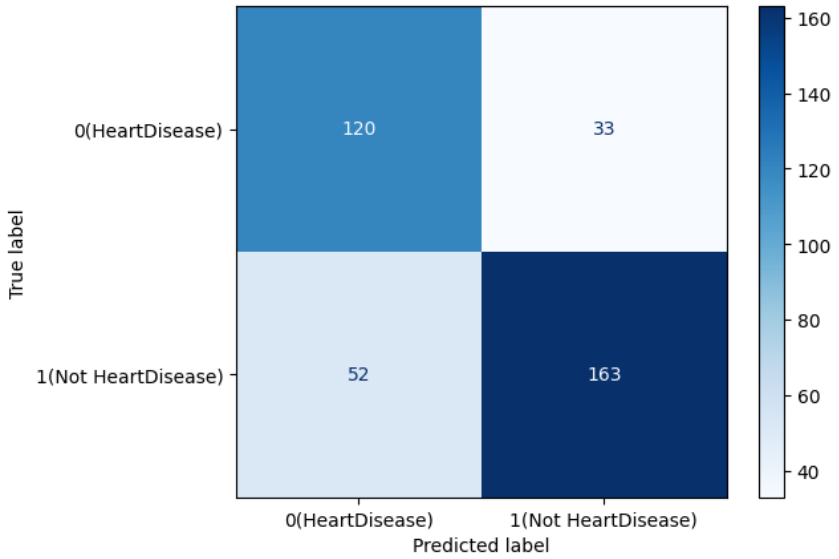


Figure (2) (confusion matrix):



- Classification [80% training, 20% testing] Information Gain:

Figure (1) (decision tree):

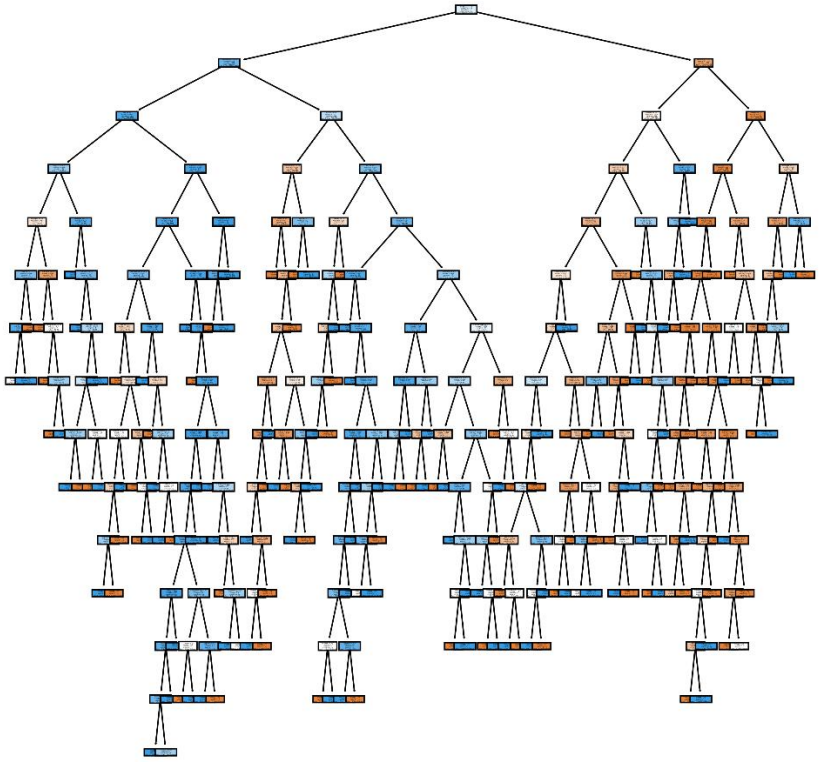
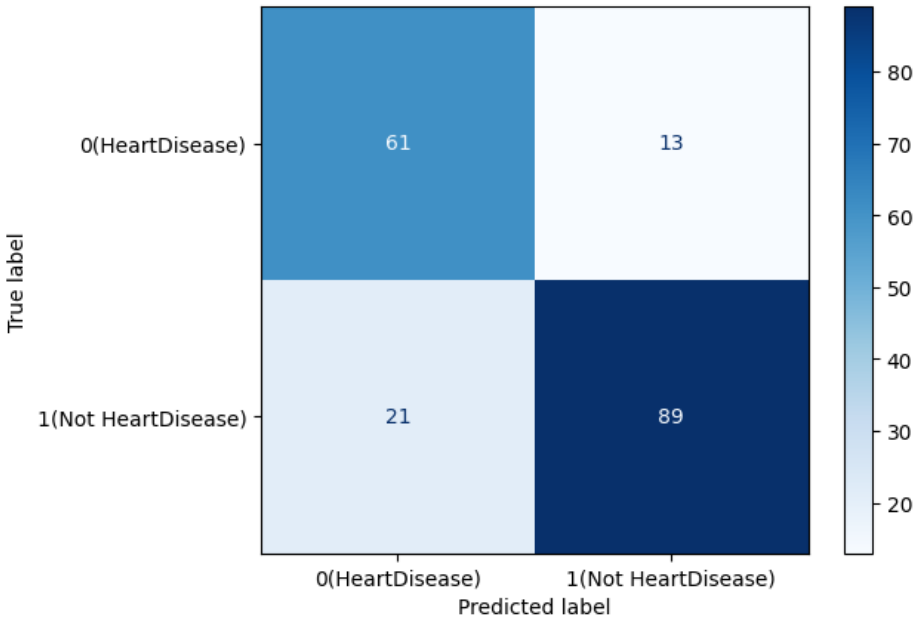


Figure (2) (confusion matrix):



Mining task	Comparison Criteria

Classification for Information Gain

- 70% Training data, 30% Test data.

| Accuracy |	**75.36%**
| Error Rate |	**24.64%**
| Sensitivity |	**74.85%**
| Specificity |	**76.15%**
| Precision |	**82.78%**

- 60% Training data, 40% Test data.

| Accuracy |	**76.90%**
| Error Rate |	**23.10%**
| Sensitivity |	**75.81%**
| Specificity |	**78.43%**
| Precision |	**83.16%**

- 80% Training data, 20% Test data.

| Accuracy |	**81.52%**
| Error Rate |	**18.48%**
| Sensitivity |	**80.91%**
| Specificity |	**82.43%**
| Precision |	**87.25%**

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

Figure (1) (decision tree):

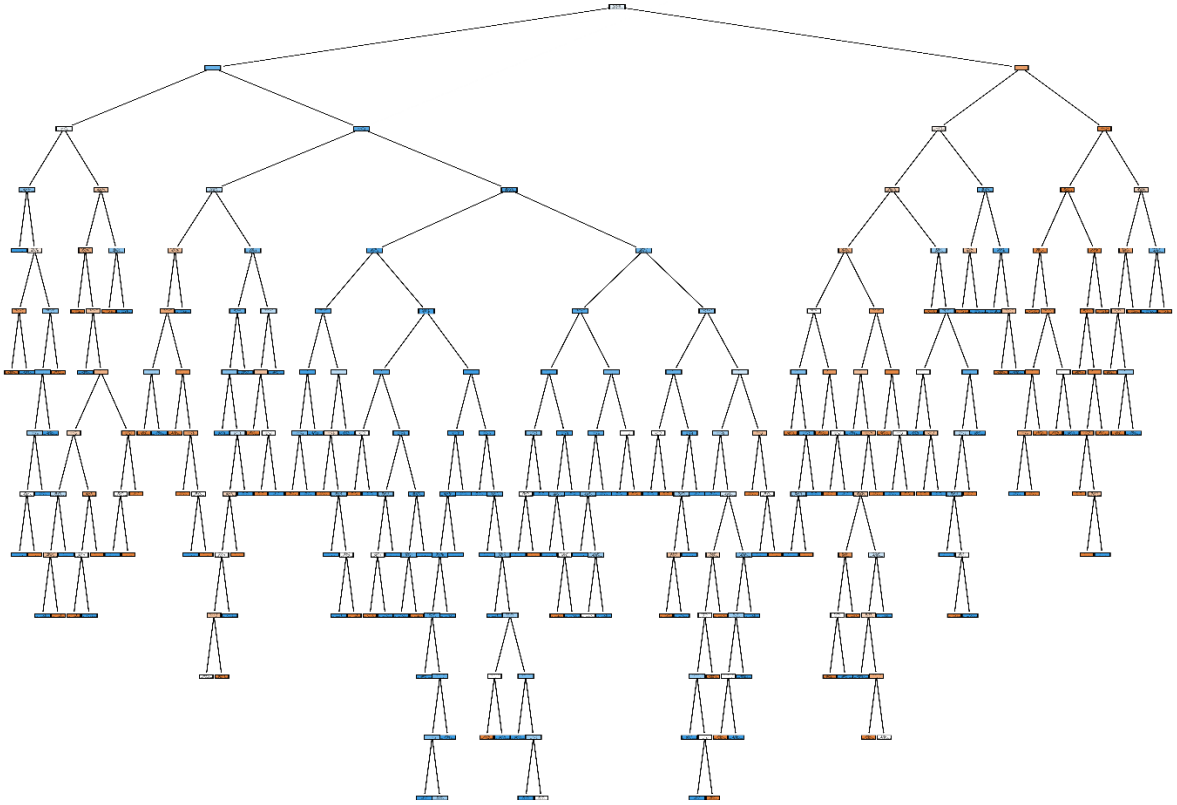
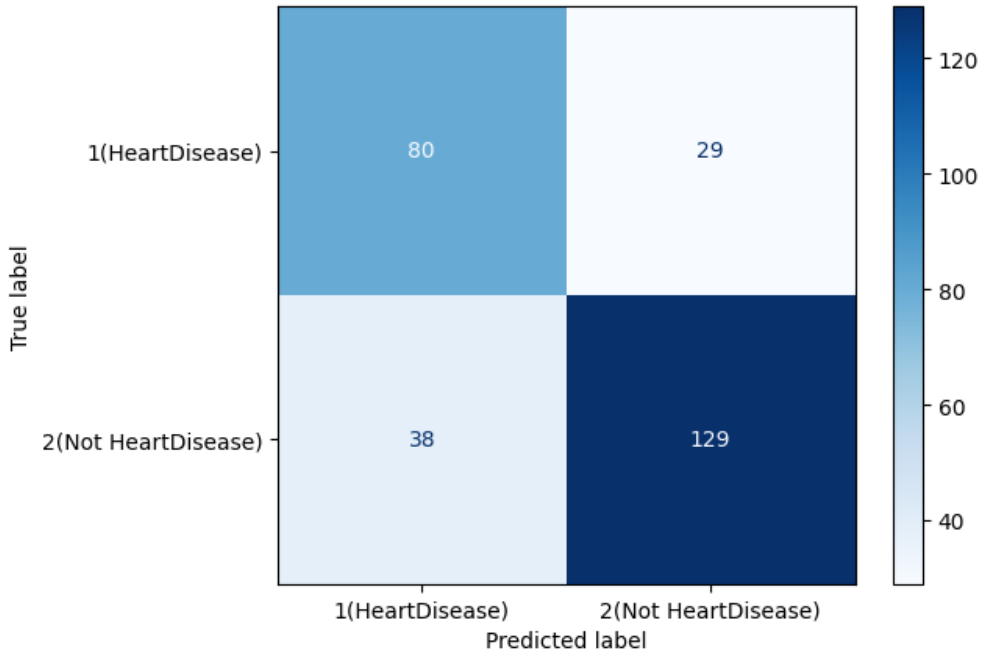


Figure (2) (confusion matrix):



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

Figure (1) (decision tree):

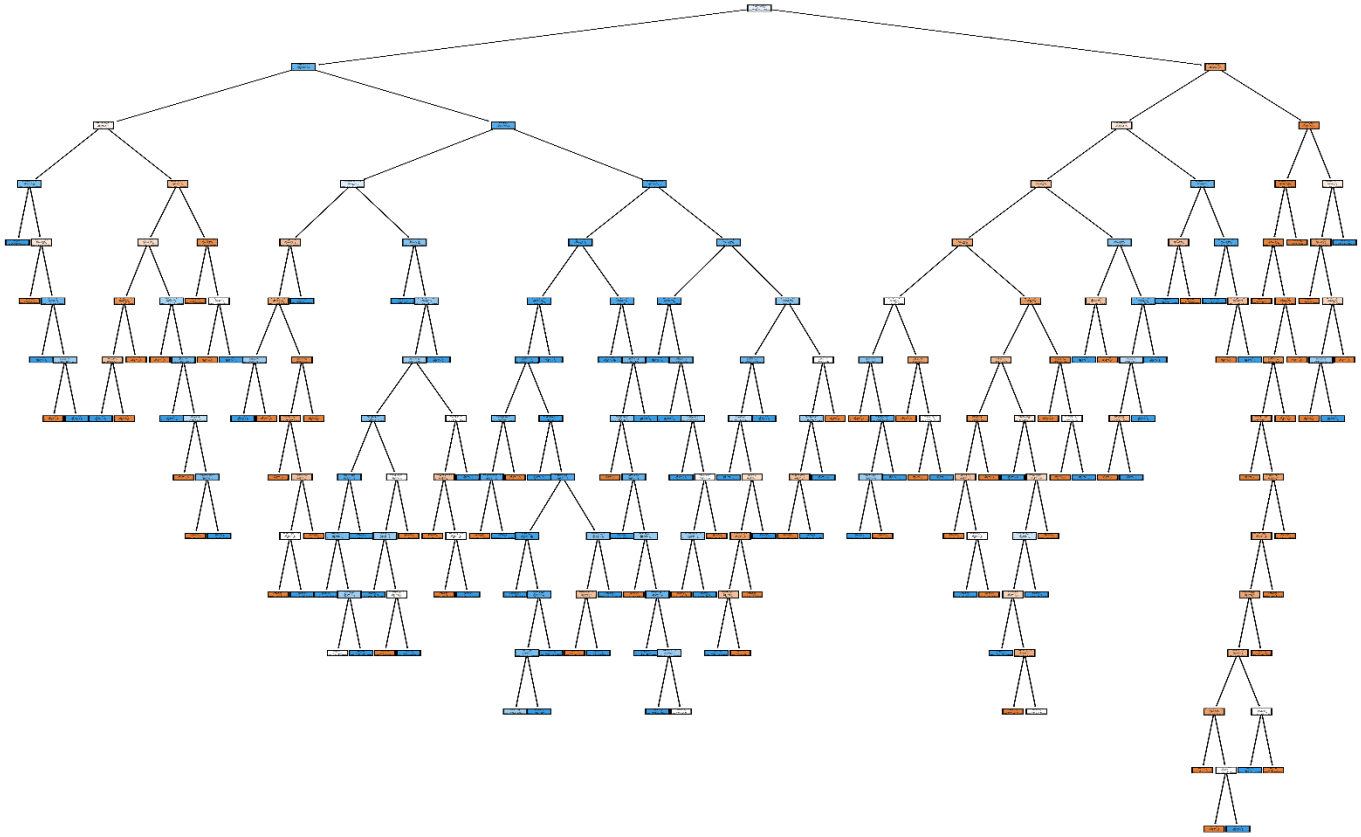
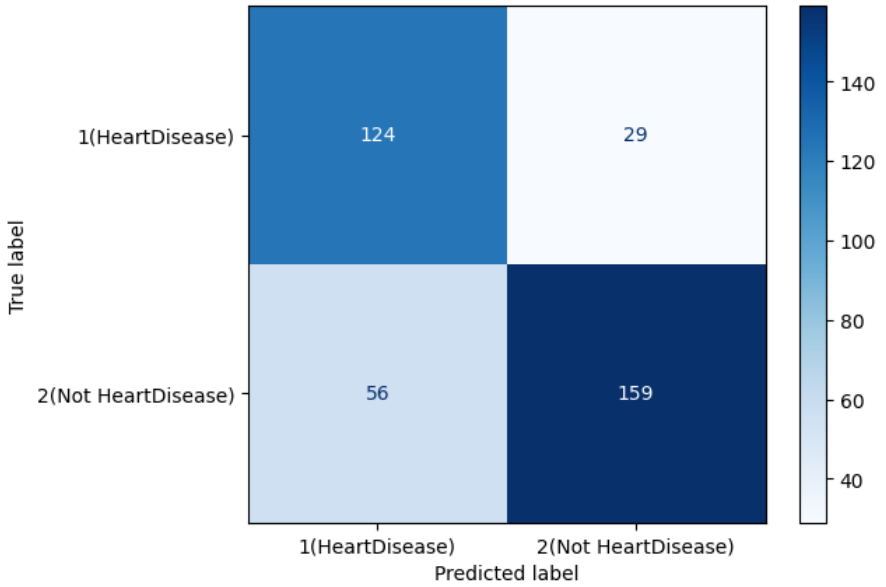


Figure (2) (confusion matrix):



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

Figure (1) (decision tree):

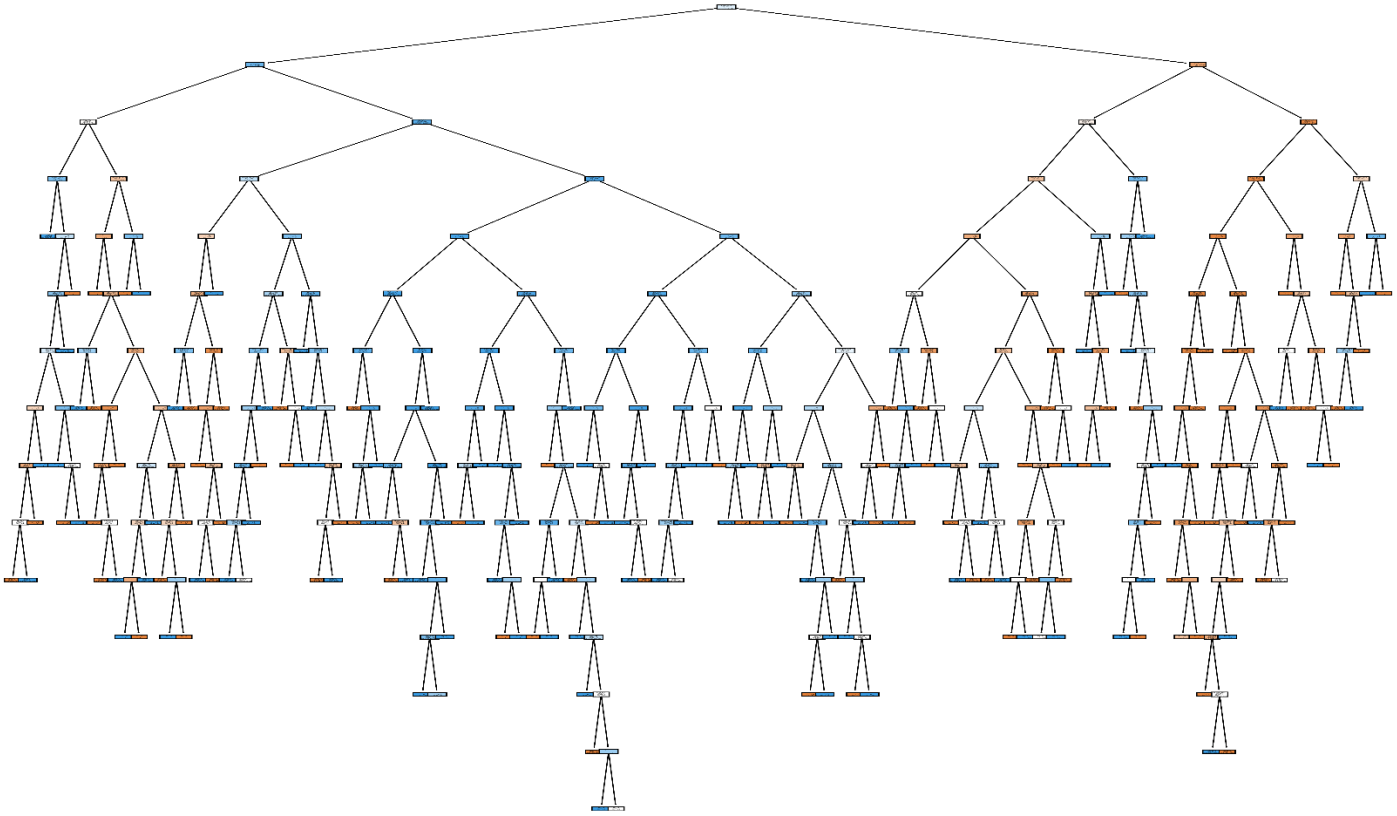
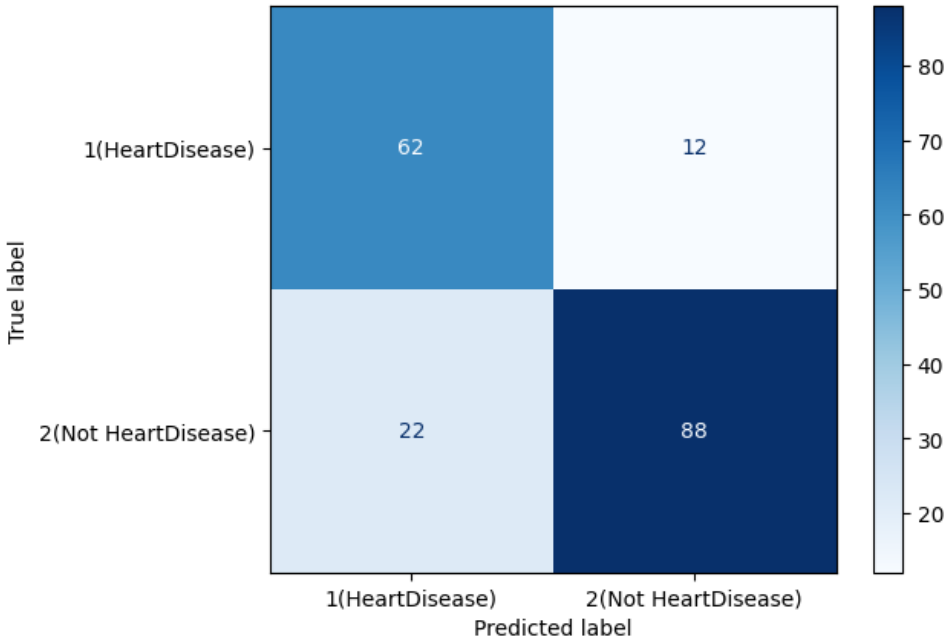


Figure (2) (confusion matrix):



Mining task	Comparison Criteria

Classification for Gini index

- 70% Training data, 30% Test data.

| Accuracy |	**76%**
| Error Rate |	**24%**
| Sensitivity |	**77%**
| Specificity |	**73%**
| Precision |	**82%**

- 60% Training data, 40% Test data.

| Accuracy |	**77%**
| Error Rate |	**23%**
| Sensitivity |	**74%**
| Specificity |	**81%**
| Precision |	**85%**

- 80% Training data, 20% Test data.

| Accuracy |	**82%**
| Error Rate |	**18%**
| Sensitivity |	**80%**
| Specificity |	**84%**
| Precision |	**88%**

Analysis of Results

1. Accuracy:

Both models achieve the same highest accuracy of 81.52% with the 80%-20% split.

2. Error Rate:

Both models show the same lowest error rate of 18.48% with the 80%-20% split.

3. Sensitivity:

Information Gain has a slightly higher sensitivity at 80.91%, compared to 80.00% for the Gini Index in the 80%-20% split. This suggests the Information Gain model is better at identifying true positives.

4. Specificity:

The Gini Index model shows higher specificity (83.78%) compared to Information Gain (82.43%) at the same split. This indicates Gini Index is better at identifying true negatives.

5. Precision:

The Gini Index model edges out with a higher precision of 88.00% compared to 87.25% for Information Gain, suggesting it has fewer false positives**.**

Conclusion:

Both models perform similarly well at the 80%-20% split, but they excel in different areas:

- Information Gain has better sensitivity, meaning it is better at identifying positive cases.
- Gini Index demonstrates better specificity and precision, indicating it is more effective at identifying
- negative cases and has fewer false positives.

Final analysis:

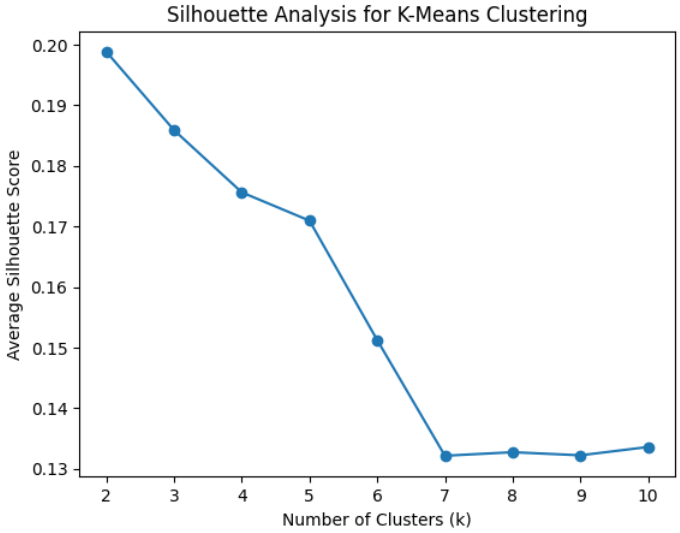
If the priority is to identify as many true positives as possible, the Information Gain model may be preferable. However, if minimizing false positives and maximizing overall precision is the goal, the Gini Index model is the better choice.

- Clustering

We selected three cluster sizes (2, 3, and 4) based on the outcomes of the validation methods we plan to apply. These sizes will then be used to execute the K-means clustering algorithm.

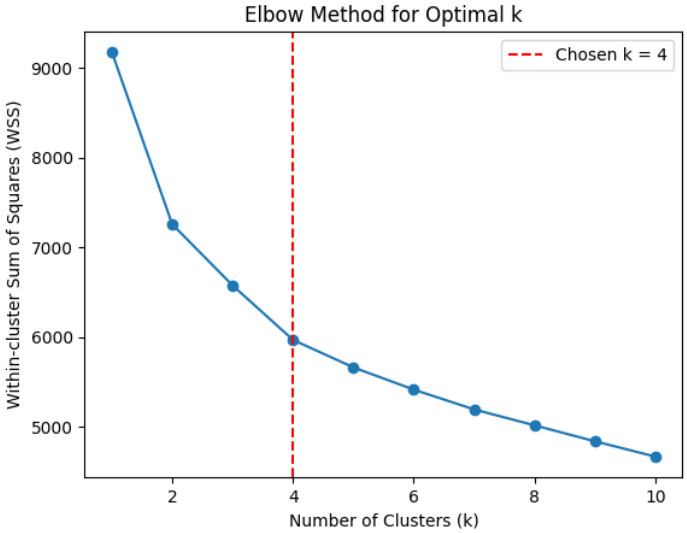
Silhouette method:

The Silhouette method evaluates clustering quality by measuring how well data points fit their cluster compared to others, with higher scores indicating better separation.



Elbow method:

The Elbow method identifies the optimal number of clusters for K-means by plotting the Within-Cluster Sum of Squares (WSS) and selecting the point where the decrease slows, forming an "elbow."



Figures:

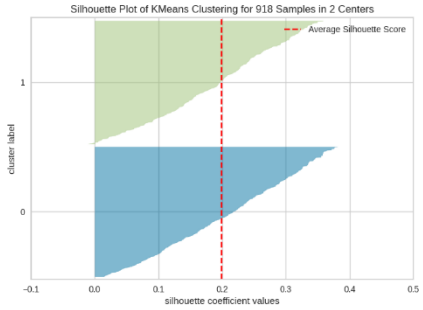


Figure (1) K=2

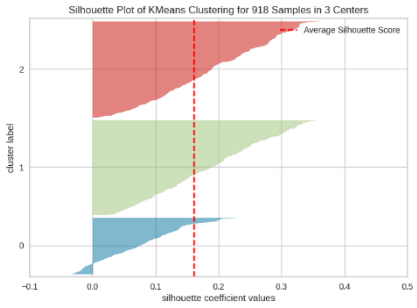


Figure (2) K=3

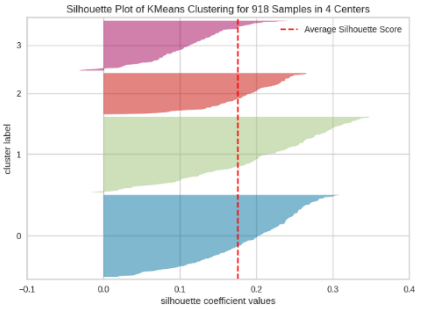


Figure (3) K=4

	K=2	K=3	K=4
Average Silhouette width	0.199	0.186	0.176
Total within-cluster sum of square	7258	6576	5968

Final decision:

Based on the metrics we've analyzed, including WSS (Within-Cluster Sum of Squares) and the average Silhouette score, we have determined that K = 2 is the most suitable and optimal choice for our clustering model. This decision is supported by the fact that K = 2 provides the highest Silhouette score among the tested values, indicating better-defined clusters with higher cohesion and separation when compared to K = 3 and K = 4. Although K = 2 has the largest WSS value, which typically decreases as K increases, the Silhouette score suggests that K = 2 is the most optimal for balancing separation and compactness of the clusters.

7-Findings

We began by selecting a dataset of patients diagnosed with heart disease to better understand the factors contributing to this serious condition and to develop effective preventive strategies.

To ensure accuracy and precision in our results, we applied several data processing techniques to enhance the dataset's efficiency. We used various visualization methods, including pie charts, box plots, scatter plots, and bar charts to clarify the data and make it easier to understand. This facilitated the application of appropriate data processing techniques. Based on these visualizations and additional analyses, we removed any empty, missing, or outlier values that could negatively affect our results.

In addition, we performed data transformations, such as normalization and feature partitioning, as well as balanced data processing to give equal weight to certain features and streamline the data mining tasks.

As a result, we conducted data mining tasks that included classification and partitioning. For classification, we utilized the Gini index and information gain metrics. By experimenting with three different sizes of training and testing data, we achieved optimal results for both model construction and evaluation. Here are our findings:

- Information Gain:

Split	Accuracy	Error Rate	Sensitivity	Specificity	Precision
70%-30%	0.7536	0.2464	0.7485	0.7615	0.8278
60%-40%	0.7690	0.2310	0.7581	0.7843	0.8316
80%-20%	0.8152	0.1848	0.8091	0.8243	0.8725

Based on the results presented for the models trained using different data splits, the following observations can be made:

- **Accuracy**: The model trained with an 80% training set and 20% testing set achieved the highest accuracy (81.52%). This indicates that a larger training set improves overall model performance.
- **Error Rate**: The error rate is lowest for the model trained with a 80%-20% split (18.48%). This suggests that this model minimizes incorrect predictions more effectively than the others.
- **Sensitivity**: The sensitivity improves significantly with a larger training set, reaching 80.91% for the 80%-20% split. This means that the model is more effective at correctly identifying positive cases as the amount of training data increases.
- **Specificity**: The model trained with an 80%-20% split also shows the highest specificity (82.43%). This indicates that it is better at accurately identifying negative cases compared to the other splits.
- **Precision**: Precision increases to 87.25% for the model trained with 80% of the data. This means that, with more training data, the model is more accurate in predicting positive instances, reducing the number of false positives.

In summary, the analysis demonstrates that increasing the size of the training data leads to improved model performance across all metrics. The model trained with 80% training set and 20% testing shows the best overall performance, indicating that a larger dataset helps the model generalize better and make more accurate predictions.

- Gini index:

Split	Accuracy	Error Rate	Sensitivity	Specificity	Precision
70%-30%	0.7572	0.2428	0.7725	0.76339	0.8165
60%-40%	0.7690	0.2310	0.7395	0.8105	0.8457
80%-20%	0.8152	0.1848	0.8000	0.8378	0.8800

Based on the results presented, the model with an **80%-20% split** is considered the best. Here are some reasons why this model outperforms the others:

- **Highest Accuracy**: The model trained using an 80% training set and 20% testing set achieved the highest accuracy rate (81.52%) among the three models. This indicates it can predict class labels more accurately than the others.
- **Highest Specificity**: This model also recorded the best specificity (83.78%) for correctly identifying negative cases. High specificity is crucial as it shows the model's ability to avoid misclassifying negative instances, making it more reliable in predicting the absence of the condition.
- **Balance Between Sensitivity and Precision**: While the sensitivity for the 80%-20% model (80.00%) is slightly lower than some other models, it remains at an acceptable level, indicating that it effectively identifies positive instances. Additionally, it achieves high precision (88.00%), meaning it makes fewer false positive predictions.
- **Lowest Error Rate**: The 80%-20% model has the lowest error rate (18.48%) among the three models, demonstrating its ability to minimize classification errors overall.

In summary, the 80%-20% model strikes a good balance between classification accuracy, specificity, and sensitivity, which is why it is considered the best based on the provided results. As the training size increases, the model's ability to distinguish between classes improves, likely leading to better performance in practical applications.

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

After selecting the best model split from Information Gain and the best split from Gini Index, which were both 80% training, 20% testing, we reviewed the values of each for comparison between Information Gain and Gini Index, and we reached the following conclusion:

Metric	Information Gain	Gini Index
Accuracy	0.7793	0.7805
Error Rate	0.2207	0.2196
Sensitivity	0.7719	0.7707
Specificity	0.7900	0.7941
Precision	0.8430	0.8141

Accuracy and Error Rate

The Gini Index achieves a comparable accuracy of **78.05%** (or **0.7805**), slightly surpassing Information Gain's accuracy of **77.93%** (or **0.7793**). This suggests that both models perform similarly in classification tasks. Furthermore, the Gini Index results in a marginally lower error rate of **21.96%** (or **0.2196**) compared to Information Gain's **22.07%** (or **0.2207**). This lower error rate indicates that the Gini Index may be slightly more reliable, minimizing misclassifications that could lead to incorrect patient management decisions.

Sensitivity and Specificity

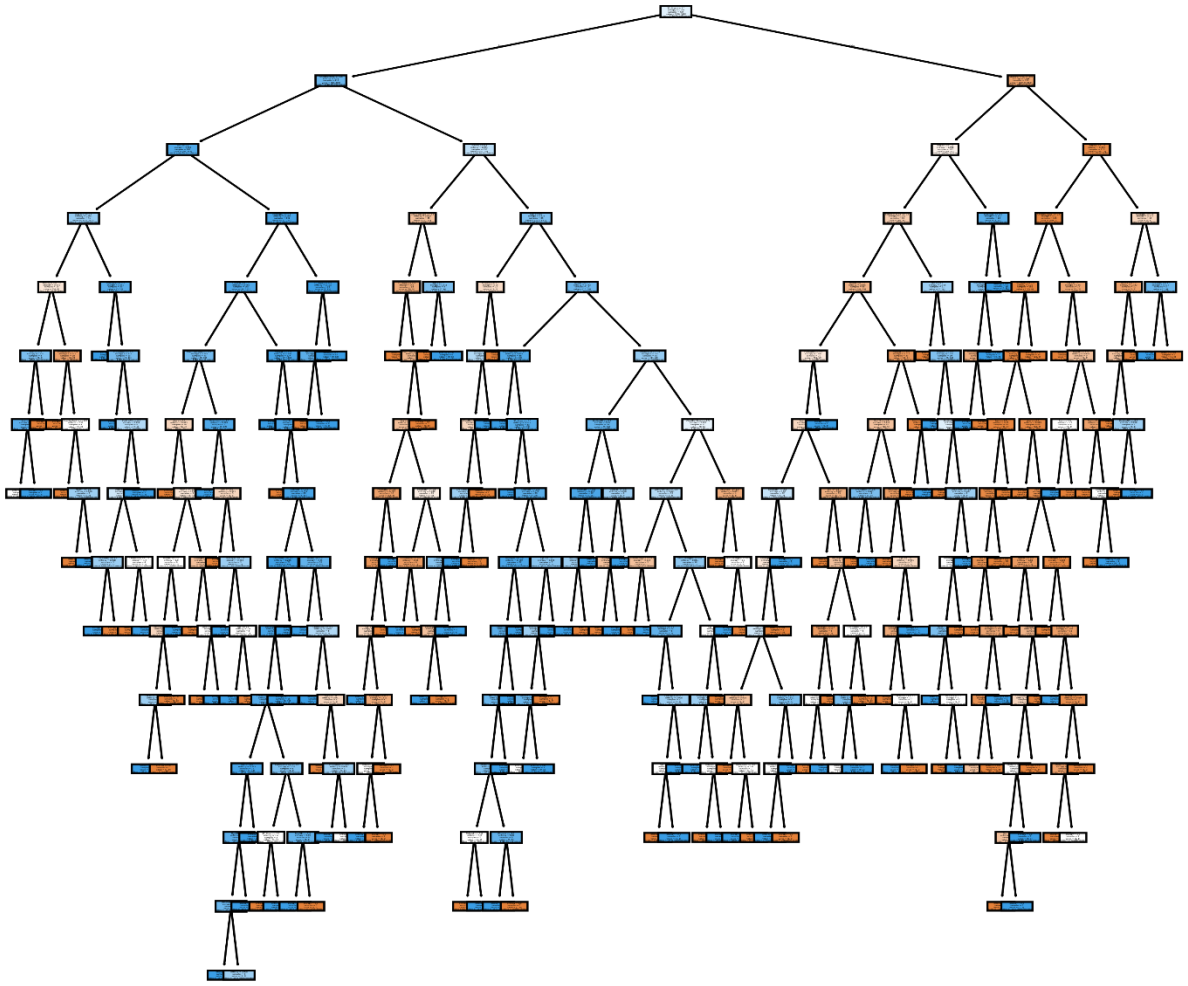
In terms of sensitivity, Information Gain slightly outperforms the Gini Index, with a sensitivity of **77.19%** (or **0.7719**) versus **77.07%** (or **0.7707**) for the Gini Index. This slight edge suggests that Information Gain is somewhat better at identifying true positive cases, which is crucial for timely interventions in patient care. However, the Gini Index demonstrates higher specificity at **79.41%** (or **0.7941**) compared to Information Gain's **79.00%** (or **0.7900**). This indicates that the Gini Index is more effective in accurately identifying negative cases, thus reducing the risk of false alarms and unnecessary treatments.

Precision

The Gini Index model achieves a precision of **81.41%** (or **0.8141**), which is slightly lower than Information Gain's precision of **84.30%** (or **0.8430**). This means that when the Gini Index predicts a positive case, it is correct **81.41%** of the time, compared to **84.30%** for Information Gain. Despite this difference, both models maintain high precision levels, ensuring that the majority of positive predictions are indeed accurate. Overall, while Information Gain excels in sensitivity and precision, the Gini Index offers advantages in accuracy and specificity, making both models valuable for different aspects of heart disease prediction.

Conclusion

Both the Gini Index and Information Gain models exhibit strong classification performance, each with unique advantages. The Gini Index achieves slightly higher accuracy (78.05%) and a lower error rate (21.96%), indicating it is marginally more reliable overall. On the other hand, Information Gain demonstrates superior sensitivity (77.19%), making it better at identifying true positives, while the Gini Index excels in specificity (79.41%), showcasing its effectiveness in accurately recognizing true negatives. In terms of precision, Information Gain also leads with 84.30%, though the Gini Index maintains a respectable 81.41%. Ultimately, the choice between these models should be guided by the specific objectives of the analysis: for maximizing true positive identification, Information Gain may be preferred, whereas for minimizing false positives and accurately identifying negatives, the Gini Index would be the better option. Conversely, Information Gain shows better sensitivity and precision in identifying positive cases. Ultimately, the choice between the two models may depend on the specific goals of classification, whether prioritizing true positives or minimizing false negatives. This was the decision tree associated with this division:



The decision tree for predicting heart disease is built on the importance of features such as Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, MaxHR, ExerciseAngina, Oldpeak, and ST_Slope. It starts with the most significant feature and progressively splits on others based on their importance.

The tree's terminal nodes classify cases into 0 (Heart Disease) or 1 (No Heart Disease) based on combinations of these attributes. This hierarchical structure reflects the complex interplay of factors affecting heart disease prediction. Understanding the tree provides valuable insights into how the model distinguishes between individuals with and without heart disease.

Findings:

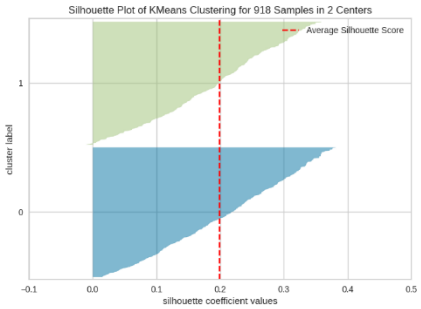
Clustering Analysis

In our analysis, we calculated the average Silhouette width for each cluster size (K) and reached the following conclusions:

	K=2	K=3	K=4
Average Silhouette width	0.199	0.186	0.176
Total within-cluster sum of square	7258	6576	5968

We found that K=2 is the best choice for our clustering model based on the metrics we analyzed, including the Within-Cluster Sum of Squares (WSS), Average Silhouette Score, and visualizations of K-means clustering. K=2 not only gave the highest Silhouette score but also showed a strong WSS value compared to K=3 and K=4.

The silhouette plot for K-means clustering with 918 samples grouped into 2 centers was an important factor in our decision for K=2. The plot showed that the clusters were clear and well-defined.



From the K-means clustering graph, we noticed that most silhouette scores were positive. This indicates that the samples fit well within their clusters and are far from other clusters. This means the clustering successfully separated the data points into distinct groups. However, it's important to mention that while having mostly positive silhouette scores is good, it doesn't mean the clustering is perfect. There might still be some overlap or confusion between clusters, especially for samples that are close to the edges, which can lead to silhouette scores near 0 or even negative values.

Conclusion

In conclusion, our clustering analysis provided helpful insights into the characteristics of patients. It showed the value of both classification and clustering models in predicting heart disease risk. However, since our dataset includes a class that indicates whether someone has heart disease, supervised learning models (like classification) are more accurate and suitable for this task. These classification models can use this known information to make better predictions, helping us understand the factors that contribute to heart disease and guiding preventive healthcare strategies.

8-References

1- F. Soriano, "Heart Failure Prediction Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data.

2- B. Siontis, V. M. Noseworthy, P. Attia, and Y. V. Murad, "Artificial Intelligence in Heart Disease Diagnosis," *PubMed*, vol. 72, no. 10, pp. 1101–1111, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31592122/.