74557King Saud University
College of Computer and Information Sciences
Information Technology department

IT 326: Data Mining
Course Project

Liver Disease

Project final Report

Group#5
LAB Day-Time:
Wednesday-8
Group members:

Group#:	5	
Section#:	74557	
Group Members	**Name**	**ID**
	Sereen Al-hmoud	443200463
	Raghad fares	443200793
	Luluh Al-yahya	443200609
	Aeshah almakhlifi	443200713

1. Problem

Recently, the prevalence of liver diseases has been increasing, becoming more common among people. This trend leads to numerous serious issues in individuals' lives, potentially resulting in fatal outcomes. In our project, we aim to study and analyse patient data, which will greatly assist in identifying possible factors and risks associated with liver diseases. By predicting the likelihood of developing liver disease, we can help many individuals take preventive measures to safeguard their health.

2. Data Mining Task

In our project, we will employ two data mining tasks to help predict the likelihood of liver diseases: classification and clustering. For classification, we will train our model to determine whether a patient suffers from liver disease or not, based on a set of medical examinations such as liver enzyme levels, bilirubin levels, age, gender, etc. Classification will be based on the "liver disease" class.

As for clustering, our model will create groups of patients who share similar characteristics, without considering the class (liver disease or not). These groups will be utilised to identify patterns and similarities in the data, potentially leading to a deeper understanding of the factors influencing liver disease and uncovering new insights if any exist.

3. Data

The Source: https://www.kaggle.com/datasets/fatemehmehrparvar/liver-disorders

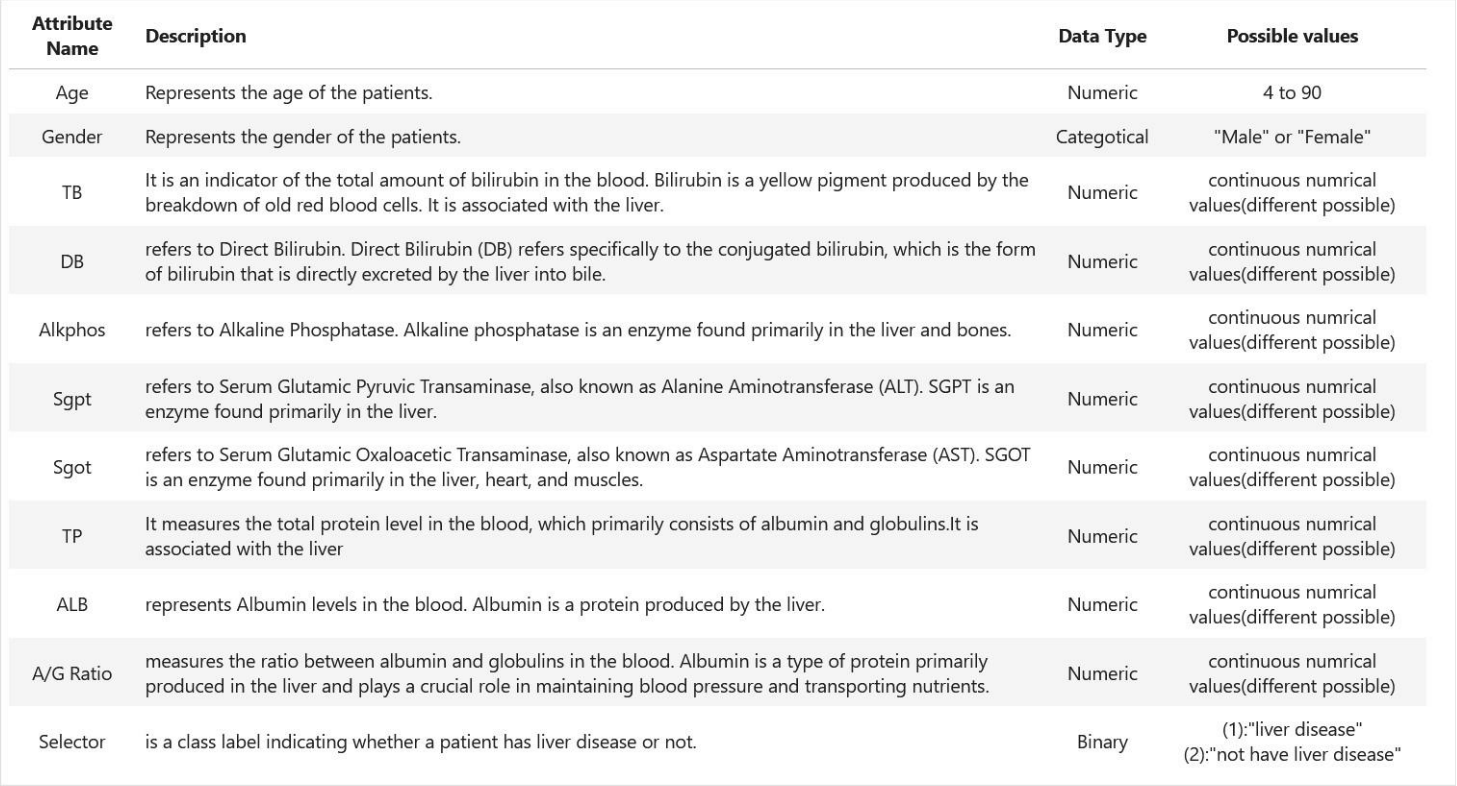
-Number of attributes: 11

-No. of objects: 583

-Class label: Selector

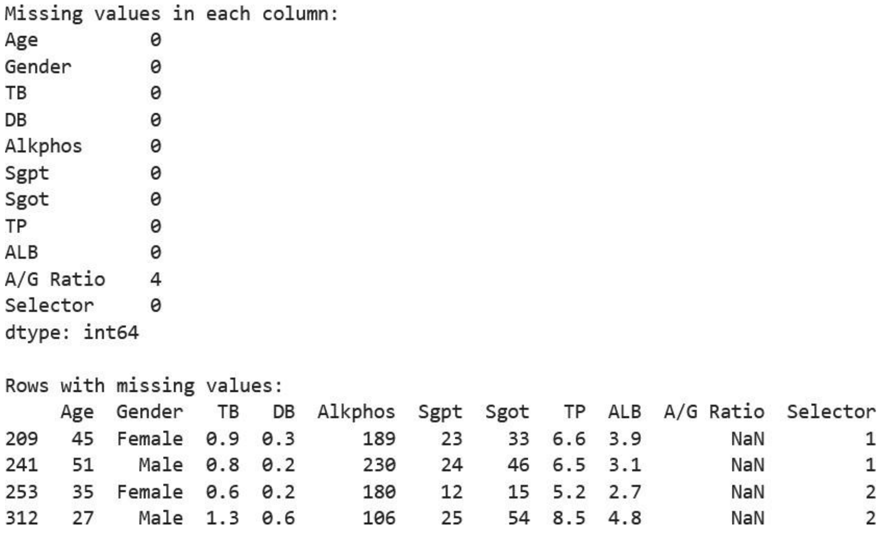
To try to understand our data, we reviewed:

- **Attributes' description**



Attribute Name	Description	Data Type	Possible values
Age	Represents the age of the patients.	Numeric	4 to 90
Gender	Represents the gender of the patients.	Categotical	"Male" or "Female"
TB	It is an indicator of the total amount of bilirubin in the blood. Bilirubin is a yellow pigment produced by the breakdown of old red blood cells. It is associated with the liver.	Numeric	continuous numrical values(different possible)
DB	refers to Direct Bilirubin. Direct Bilirubin (DB) refers specifically to the conjugated bilirubin, which is the form of bilirubin that is directly excreted by the liver into bile.	Numeric	continuous numrical values(different possible)
Alkphos	refers to Alkaline Phosphatase. Alkaline phosphatase is an enzyme found primarily in the liver and bones.	Numeric	continuous numrical values(different possible)
Sgpt	refers to Serum Glutamic Pyruvic Transaminase, also known as Alanine Aminotransferase (ALT). SGPT is an enzyme found primarily in the liver.	Numeric	continuous numrical values(different possible)
Sgot	refers to Serum Glutamic Oxaloacetic Transaminase, also known as Aspartate Aminotransferase (AST). SGOT is an enzyme found primarily in the liver, heart, and muscles.	Numeric	continuous numrical values(different possible)
TP	It measures the total protein level in the blood, which primarily consists of albumin and globulins.It is associated with the liver	Numeric	continuous numrical values(different possible)
ALB	represents Albumin levels in the blood. Albumin is a protein produced by the liver.	Numeric	continuous numrical values(different possible)
A/G Ratio	measures the ratio between albumin and globulins in the blood. Albumin is a type of protein primarily produced in the liver and plays a crucial role in maintaining blood pressure and transporting nutrients.	Numeric	continuous numrical values(different possible)
Selector	is a class label indicating whether a patient has liver disease or not.	Binary	(1):"liver disease" (2):"not have liver disease"

- **Missing values**



```
Missing values in each column:
Age          0
Gender       0
TB           0
DB           0
Alkphos      0
Sgpt         0
Sgot         0
TP           0
ALB          0
A/G Ratio    4
Selector     0
dtype: int64

Rows with missing values:
     Age  Gender   TB   DB  Alkphos  Sgpt  Sgot   TP  ALB  A/G Ratio  Selector
209   45  Female  0.9  0.3      189    23    33  6.6  3.9        NaN         1
241   51    Male  0.8  0.2      230    24    46  6.5  3.1        NaN         1
253   35  Female  0.6  0.2      180    12    15  5.2  2.7        NaN         2
312   27    Male  1.3  0.6      106    25    54  8.5  4.8        NaN         2
```

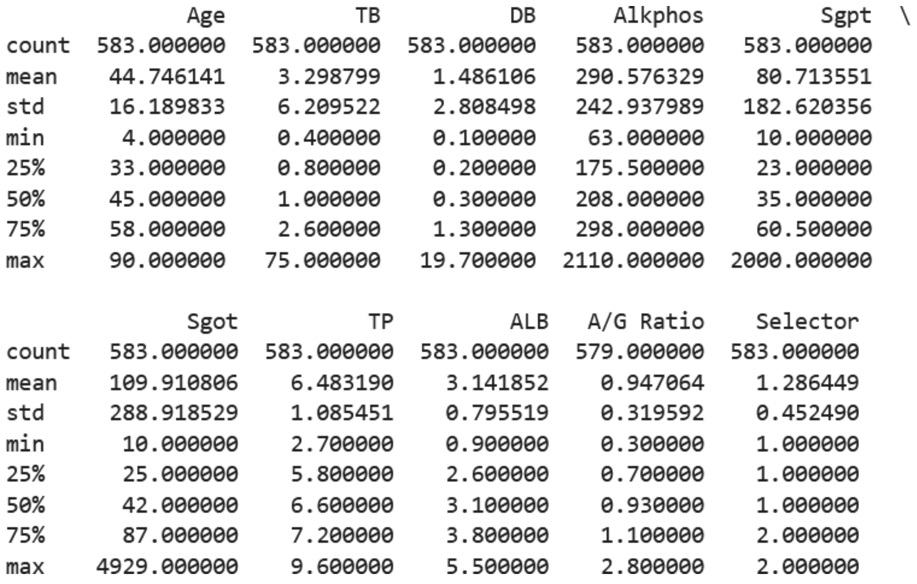
We have 4 missing values in only one attribute (A/G Ratio).

- **Statical Measures for each numeric column:**

<u>-Show Five Number Summary:</u>

using summary_stats() function. From these summary statistics, several key observations can be made:

- Age: There is significant variability in ages, ranging from 4 to 90 years, with an average of 44.74 years. This indicates that liver disease can affect individuals across a wide age range.
- Total Bilirubin (TB): The values vary significantly, with a maximum of 75 and a minimum of 0.4. The mean is 3.3, while the median is 1. This suggests the presence of extreme values or some deviation in TB levels.
- Direct Bilirubin (DB): DB values range from 0.1 to 19.7, with a mean of 1.49, indicating significant variation in direct bilirubin levels.
- Alkaline Phosphatase (Alkphos): Alkphos values range from 63 to 2110, with a mean of 290.58, indicating the presence of extreme values and significant variation in alkaline phosphatase levels.
- Serum Glutamic-Pyruvic Transaminase (Sgpt): Sgpt values range from 10 to 2000, with a mean of 80.71, indicating significant variation in Sgpt levels.
- Serum Glutamic-Oxaloacetic Transaminase (Sgot): Sgot values range from 10 to 4929, with a mean of 109.91, indicating significant variation in Sgot levels.
- Total Protein (TP): TP values range from 2.7 to 9.6, with a mean of 6.48, suggesting convergence of data and no significant variation in total protein levels.
- Albumin (ALB): ALB values range from 0.3 to 2.8, with a mean of 0.947, indicating convergence of data and no significant variation in albumin levels.
- Albumin/Globulin Ratio (A/G Ratio): The A/G Ratio ranges from 1 to 2, with a mean of 1.29, indicating convergence of data and no significant variation in the albumin/globulin ratio.
- Selector: The values are binary, limited to 1 and 2, indicating binary classification labels.



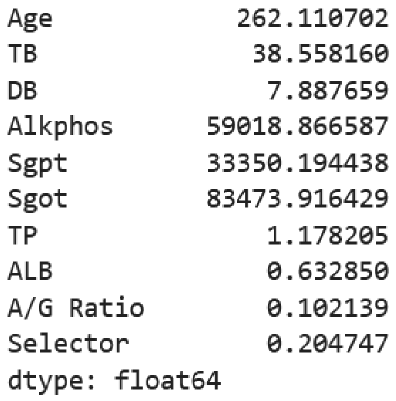
```
              Age          TB          DB      Alkphos         Sgpt  \
count  583.000000  583.000000  583.000000   583.000000   583.000000
mean    44.746141    3.298799    1.486106   290.576329    80.713551
std     16.189833    6.209522    2.808498   242.937989   182.620356
min      4.000000    0.400000    0.100000    63.000000    10.000000
25%     33.000000    0.800000    0.200000   175.500000    23.000000
50%     45.000000    1.000000    0.300000   208.000000    35.000000
75%     58.000000    2.600000    1.300000   298.000000    60.500000
max     90.000000   75.000000   19.700000  2110.000000  2000.000000

              Sgot          TP         ALB   A/G Ratio    Selector
count   583.000000  583.000000  583.000000  579.000000  583.000000
mean    109.910806    6.483190    3.141852    0.947064    1.286449
std     288.918529    1.085451    0.795519    0.319592    0.452490
min      10.000000    2.700000    0.900000    0.300000    1.000000
25%      25.000000    5.800000    2.600000    0.700000    1.000000
50%      42.000000    6.600000    3.100000    0.930000    1.000000
75%      87.000000    7.200000    3.800000    1.100000    2.000000
max    4929.000000    9.600000    5.500000    2.800000    2.000000
```

-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:

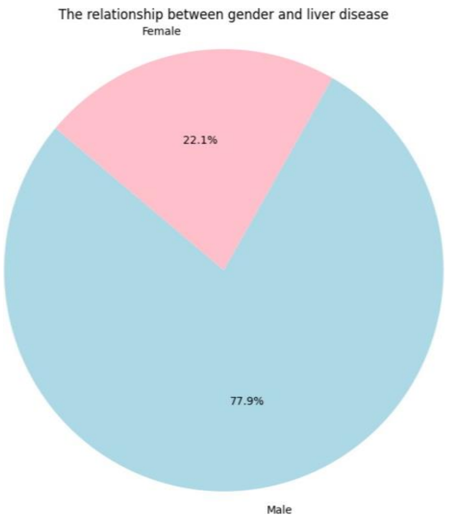
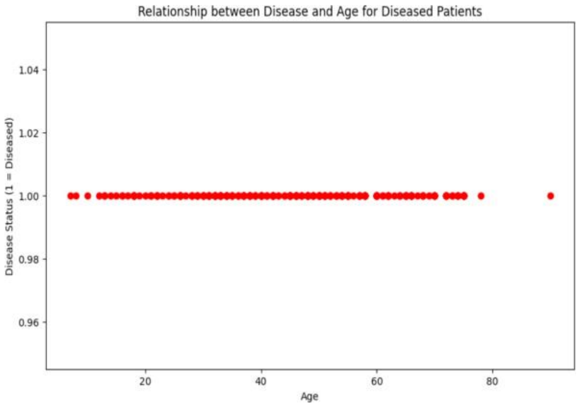
- Age: The variance is high, so the level of dispersion and spread of values is high.
- TB, DB, Alkphos, Sgpt, Sgot: The variance is very high in these columns, so the level of dispersion and spread of values is very high.
- TP, ALB, A/G Ratio, Selector: The variance is moderate to low in these columns, so the level of dispersion and spread of values is moderate to low.

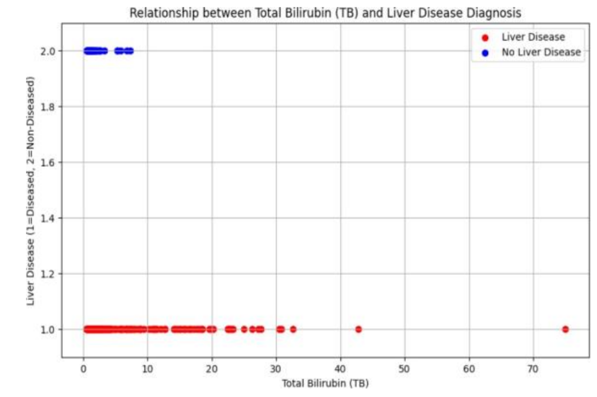
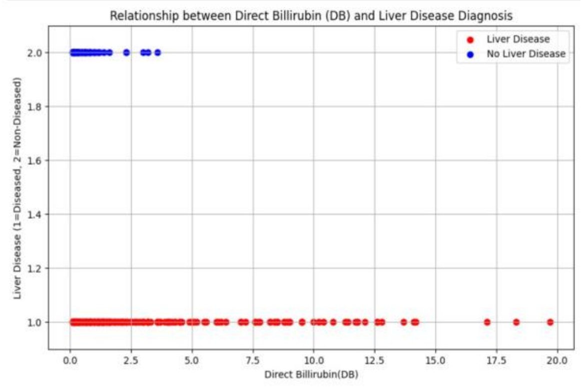
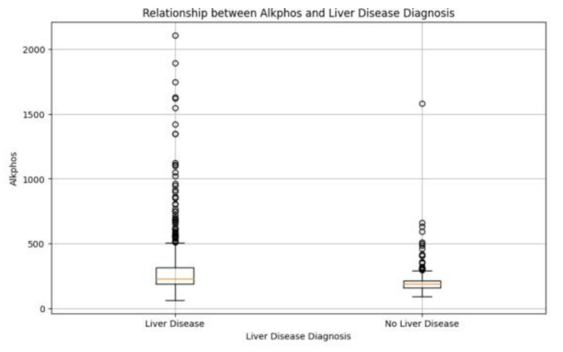
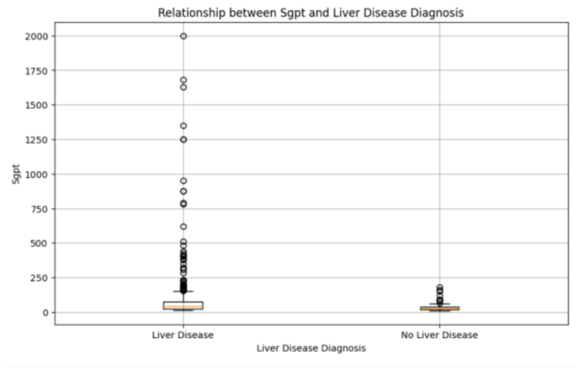


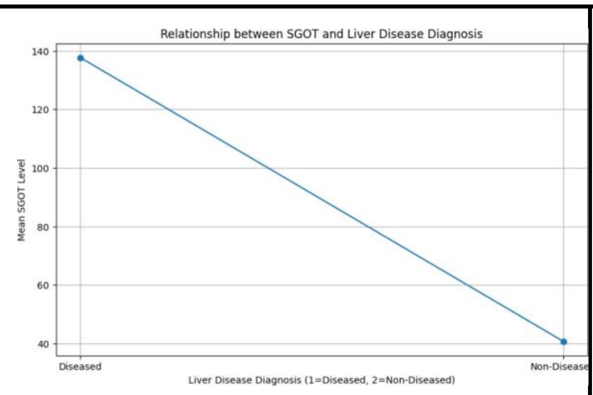
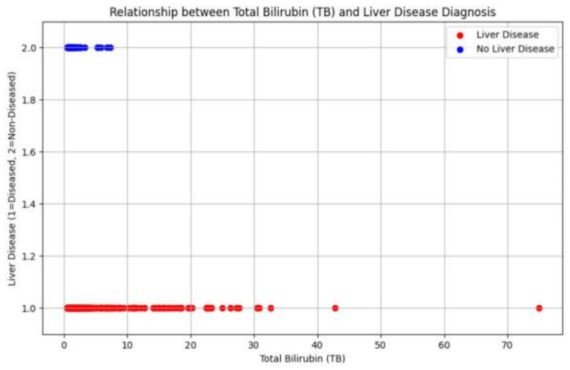
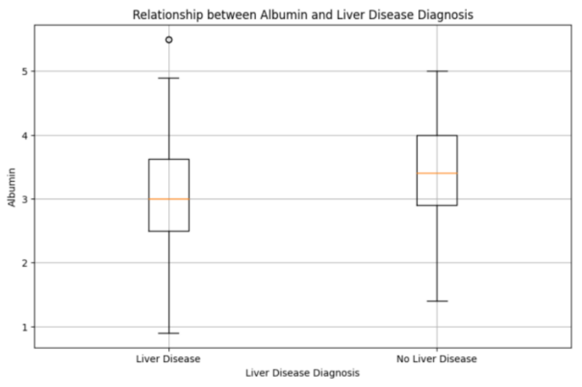
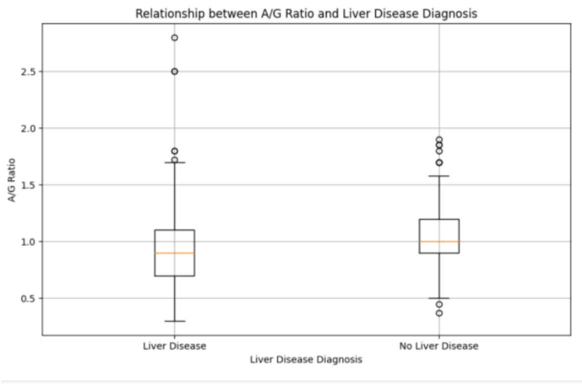
```
Age              262.110702
TB                38.558160
DB                 7.887659
Alkphos        59018.866587
Sgpt           33350.194438
Sgot           83473.916429
TP                 1.178205
ALB                0.632850
A/G Ratio          0.102139
Selector           0.204747
dtype: float64
```

- **Understanding the data through graph representations:**

To understand the relationship between liver disease and all attributes, particularly how they are associated with the likelihood of contracting liver disease, the "Selector" label class was primarily used. It indicates whether individuals are affected or unaffected and is linked to all attributes in the data. This linkage is used to extract relationships and infer whether an increase in a particular attribute suggests a likelihood of liver disease. Additionally, it helps determine if the likelihood of contracting liver disease is greater in women or men and their relationship with age (whether they are positively or negatively correlated), among other factors. This aids in understanding the factors influencing this disease and identifying indicators that could assist in early diagnosis.

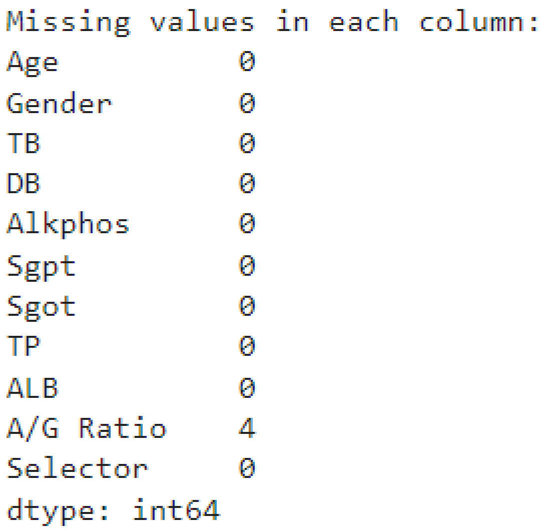
Name of Graph	Picture of Graph	Description
Pie Chart		People affected by liver disease were selected from the 'Selector' attribute for both genders of the 'Gander' attribute for comparison between the ratio of affected women and men. As a result, it has been found that men are significantly more susceptible to liver disease than women.
Scatter Chart		The chart(Scatter) above indicates that there is no clear positive or negative correlation between age and the disease, as the likelihood of contracting it varies across all age groups, indicating that the possibility of contracting it is present for all individuals of different ages.

Scatter Chart	 	The chart(Scatter) above indicates a relationship between liver disease diagnosis and elevated levels of Total Bilirubin in the blood, as patients with liver disease have higher levels of Total Bilirubin in the blood compared to non-diseased individuals.
Scatter Chart	 	The chart(Scatter) shows a relationship between DB(Direct Bilirubin) and liver disease diagnosis, as patients with liver disease have higher levels of DB (Direct Bilirubin) compared to those without liver disease.
Boxplot Chart	 	The chart(Boxplot) illustrates a direct relationship between Alkphos and liver disease, as patients with liver disease have higher levels of Alkphos compared to non-diseased individuals.
Boxplot Chart	 	The chart(Boxplot) clearly shows a relationship between liver disease and SGPT(Serum Glutamate Pyruvate Transaminase) as patients with liver disease have higher levels of this enzyme compared to non-affected individuals.

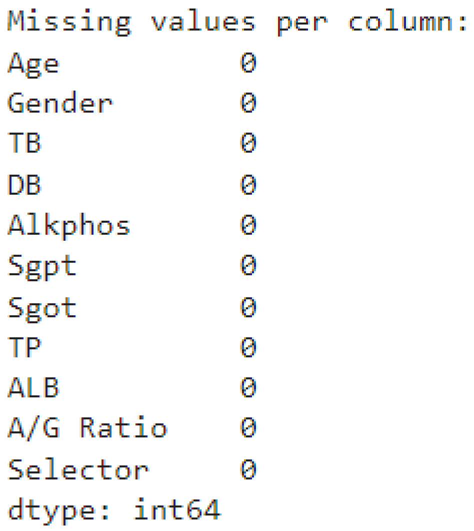
Line Graph		The chart(line Graph) clearly shows a relationship between SGOT(Serum Glutamate Oxaloacetic Transaminase) and liver disease, as patients with liver disease have higher levels of SGOT compared to non-affected individuals.
Boxplot Chart		The chart(Boxplot) reveals a relationship between TP(Total Proteins) and liver disease, despite the weak correlation. Non-affected individuals also have TP levels, but those affected have slightly higher levels.
Boxplot Chart		From the chart(BoxPlot), it is evident that there is a negative correlation between ALB(Albumin) and non-affected individuals with liver disease, as an increase in ALB indicates liver health. Individuals with liver disease have lower levels of ALB.
Boxplot Chart		The (boxplot) shows a decrease in the A/G ratio among individuals with liver disease compared to those without. This indicates that the decrease in the A/G ratio may be an indicator of changes in liver function and a marker for the occurrence of liver disease.

4- Data preprocessing:

- **Checking for missing values:**



```
Missing values in each column:
Age          0
Gender       0
TB           0
DB           0
Alkphos      0
Sgpt         0
Sgot         0
TP           0
ALB          0
A/G Ratio    4
Selector     0
dtype: int64
```

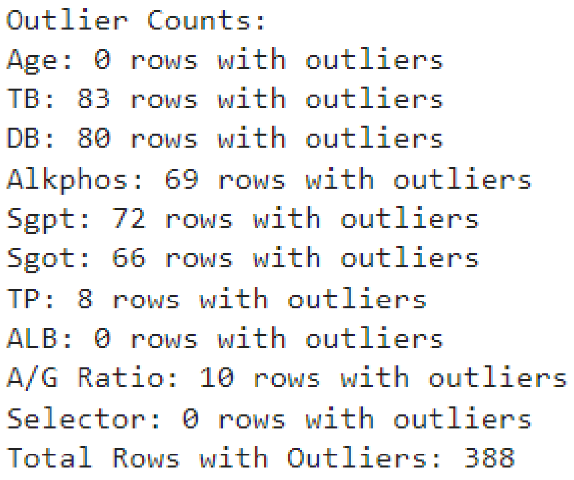


```
Missing values per column:
Age          0
Gender       0
TB           0
DB           0
Alkphos      0
Sgpt         0
Sgot         0
TP           0
ALB          0
A/G Ratio    0
Selector     0
dtype: int64
```

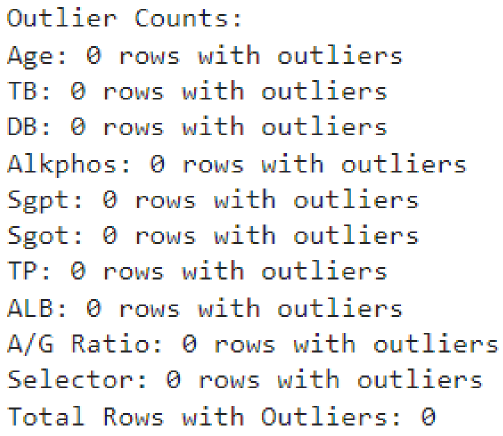
Description:

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 column which is A/G 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
TB: 83 rows with outliers
DB: 80 rows with outliers
Alkphos: 69 rows with outliers
Sgpt: 72 rows with outliers
Sgot: 66 rows with outliers
TP: 8 rows with outliers
ALB: 0 rows with outliers
A/G Ratio: 10 rows with outliers
Selector: 0 rows with outliers
Total Rows with Outliers: 388
```



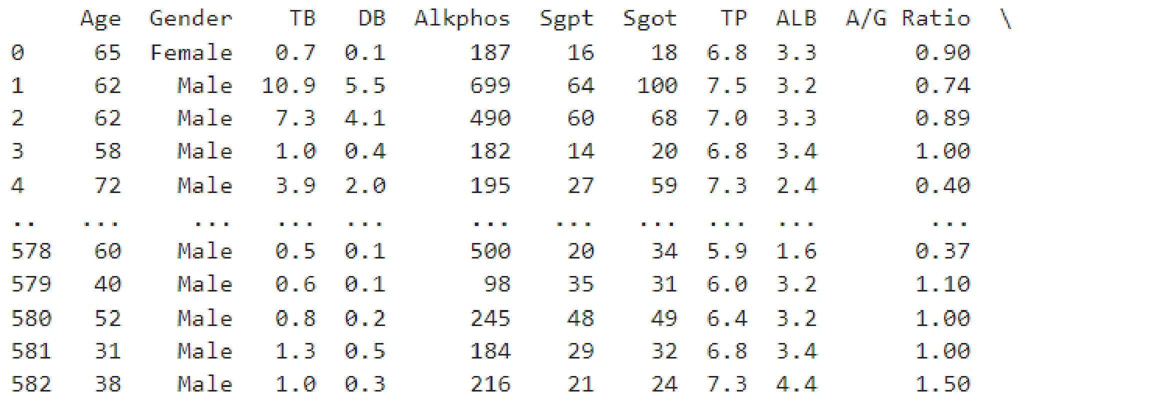
```
Outlier Counts:
Age: 0 rows with outliers
TB: 0 rows with outliers
DB: 0 rows with outliers
Alkphos: 0 rows with outliers
Sgpt: 0 rows with outliers
Sgot: 0 rows with outliers
TP: 0 rows with outliers
ALB: 0 rows with outliers
A/G Ratio: 0 rows with outliers
Selector: 0 rows with outliers
Total Rows with Outliers: 0
```

Description:

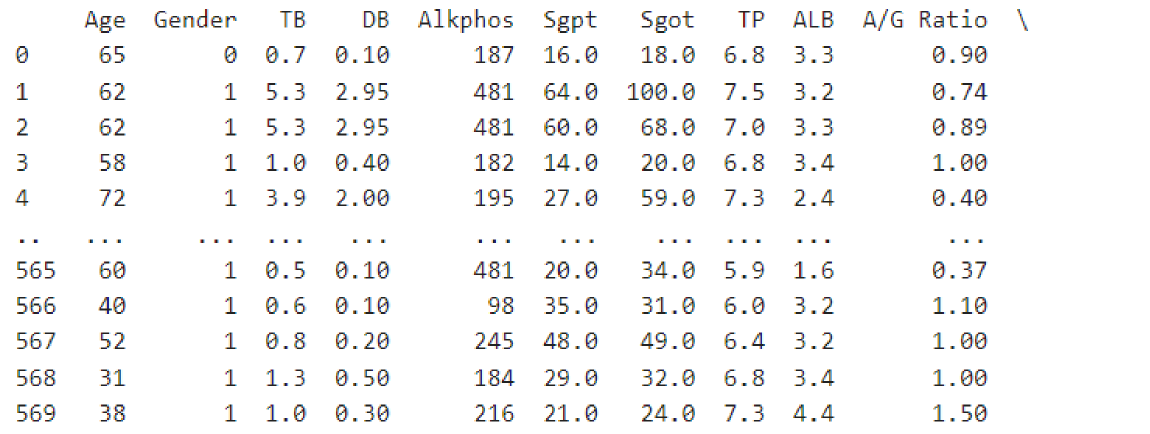
We detected a significant number of outliers in our dataset, comprising 388 rows out of 570. To handle this, we applied the interquartile range (IQR) method. Insteadof removing these outliers, we opted to cap them by replacing them with the nearest non-outlier values. This approach retains the dataset's integrity while minimizing the impact of extreme values on subsequent analyses. Through this method, we aimed to preserve valuable information while accounting for extremeobservations.

- **Data Transformation:**

1. Encoding:



```
     Age  Gender    TB   DB  Alkphos  Sgpt  Sgot   TP  ALB  A/G Ratio  \
0     65  Female   0.7  0.1      187    16    18  6.8  3.3       0.90   
1     62    Male  10.9  5.5      699    64   100  7.5  3.2       0.74   
2     62    Male   7.3  4.1      490    60    68  7.0  3.3       0.89   
3     58    Male   1.0  0.4      182    14    20  6.8  3.4       1.00   
4     72    Male   3.9  2.0      195    27    59  7.3  2.4       0.40   
..   ...     ...   ...  ...      ...   ...   ...  ...  ...        ...   
578   60    Male   0.5  0.1      500    20    34  5.9  1.6       0.37   
579   40    Male   0.6  0.1       98    35    31  6.0  3.2       1.10   
580   52    Male   0.8  0.2      245    48    49  6.4  3.2       1.00   
581   31    Male   1.3  0.5      184    29    32  6.8  3.4       1.00   
582   38    Male   1.0  0.3      216    21    24  7.3  4.4       1.50   
```

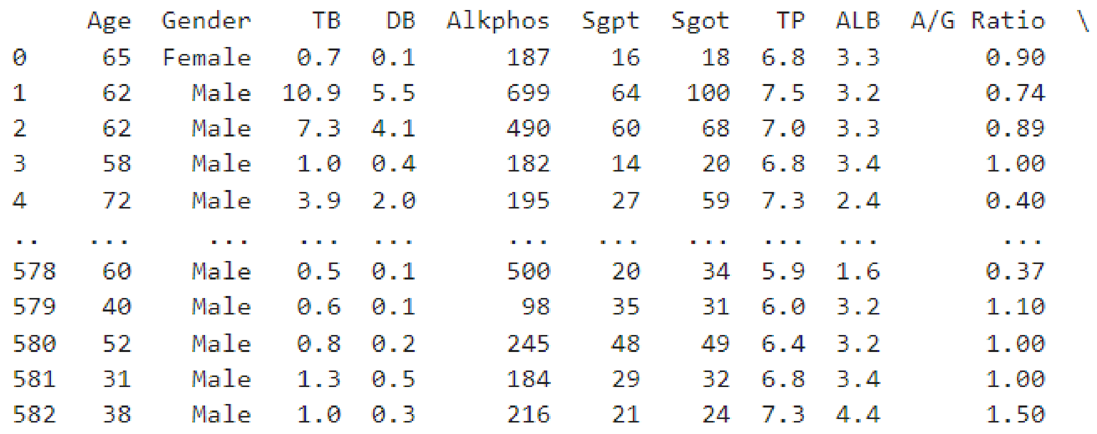


```
     Age  Gender   TB    DB  Alkphos  Sgpt   Sgot   TP  ALB  A/G Ratio  \
0     65       0  0.7  0.10      187  16.0   18.0  6.8  3.3       0.90   
1     62       1  5.3  2.95      481  64.0  100.0  7.5  3.2       0.74   
2     62       1  5.3  2.95      481  60.0   68.0  7.0  3.3       0.89   
3     58       1  1.0  0.40      182  14.0   20.0  6.8  3.4       1.00   
4     72       1  3.9  2.00      195  27.0   59.0  7.3  2.4       0.40   
..   ...     ...  ...   ...      ...   ...    ...  ...  ...        ...   
565   60       1  0.5  0.10      481  20.0   34.0  5.9  1.6       0.37   
566   40       1  0.6  0.10       98  35.0   31.0  6.0  3.2       1.10   
567   52       1  0.8  0.20      245  48.0   49.0  6.4  3.2       1.00   
568   31       1  1.3  0.50      184  29.0   32.0  6.8  3.4       1.00   
569   38       1  1.0  0.30      216  21.0   24.0  7.3  4.4       1.50   
```

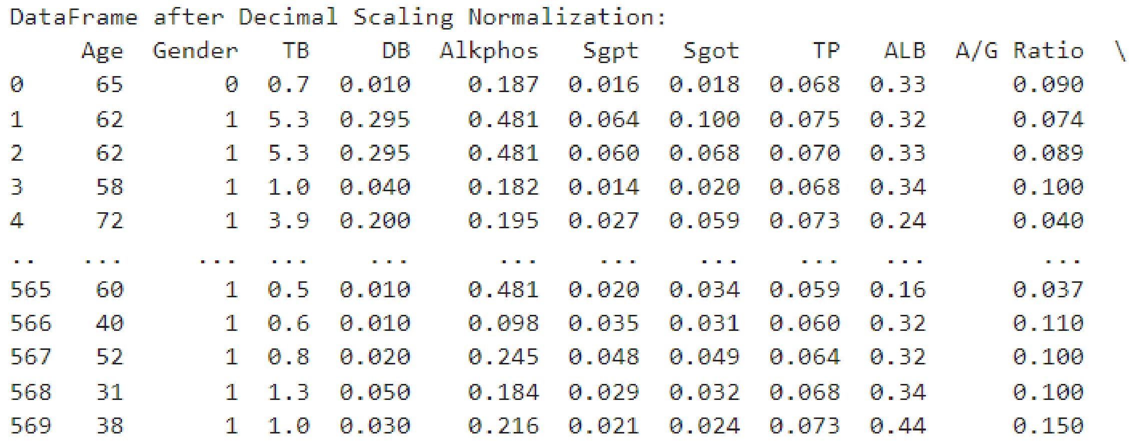
Description:

This encoding method provides a numerical representation for gender, where assigning the values 0 and 1 helps standardize the gender variable for computational purposes.where 1 corresponds to male and 0 corresponds to female. This enables easierprocessing and analysis of gender-related data in various algorithms and models.

2. Normalization:



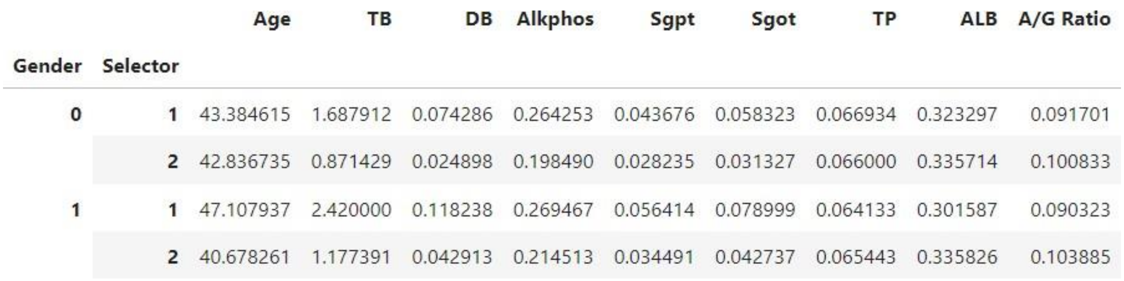
```
     Age  Gender    TB   DB  Alkphos  Sgpt  Sgot   TP  ALB  A/G Ratio  \
0     65  Female   0.7  0.1      187    16    18  6.8  3.3       0.90
1     62    Male  10.9  5.5      699    64   100  7.5  3.2       0.74
2     62    Male   7.3  4.1      490    60    68  7.0  3.3       0.89
3     58    Male   1.0  0.4      182    14    20  6.8  3.4       1.00
4     72    Male   3.9  2.0      195    27    59  7.3  2.4       0.40
..   ...     ...   ...  ...      ...   ...   ...  ...  ...        ...
578   60    Male   0.5  0.1      500    20    34  5.9  1.6       0.37
579   40    Male   0.6  0.1       98    35    31  6.0  3.2       1.10
580   52    Male   0.8  0.2      245    48    49  6.4  3.2       1.00
581   31    Male   1.3  0.5      184    29    32  6.8  3.4       1.00
582   38    Male   1.0  0.3      216    21    24  7.3  4.4       1.50
```



```
DataFrame after Decimal Scaling Normalization:
     Age  Gender   TB     DB  Alkphos   Sgpt   Sgot     TP   ALB  A/G Ratio  \
0     65       0  0.7  0.010    0.187  0.016  0.018  0.068  0.33      0.090
1     62       1  5.3  0.295    0.481  0.064  0.100  0.075  0.32      0.074
2     62       1  5.3  0.295    0.481  0.060  0.068  0.070  0.33      0.089
3     58       1  1.0  0.040    0.182  0.014  0.020  0.068  0.34      0.100
4     72       1  3.9  0.200    0.195  0.027  0.059  0.073  0.24      0.040
..   ...     ...  ...    ...      ...    ...    ...    ...   ...        ...
565   60       1  0.5  0.010    0.481  0.020  0.034  0.059  0.16      0.037
566   40       1  0.6  0.010    0.098  0.035  0.031  0.060  0.32      0.110
567   52       1  0.8  0.020    0.245  0.048  0.049  0.064  0.32      0.100
568   31       1  1.3  0.050    0.184  0.029  0.032  0.068  0.34      0.100
569   38       1  1.0  0.030    0.216  0.021  0.024  0.073  0.44      0.150
```

Here in the Normalization method, we normalize the attributes and unify their scalesince 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.

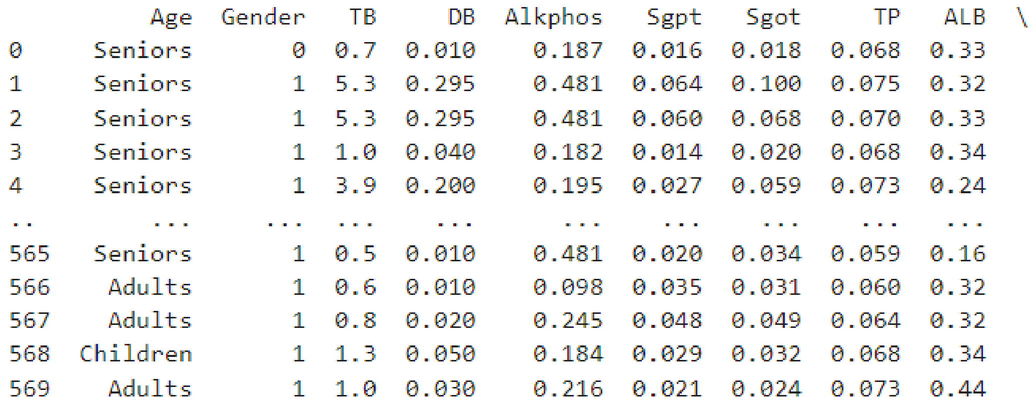
3. Aggregation:



		Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G Ratio
Gender	Selector									
0	1	43.384615	1.687912	0.074286	0.264253	0.043676	0.058323	0.066934	0.323297	0.091701
	2	42.836735	0.871429	0.024898	0.198490	0.028235	0.031327	0.066000	0.335714	0.100833
1	1	47.107937	2.420000	0.118238	0.269467	0.056414	0.078999	0.064133	0.301587	0.090323
	2	40.678261	1.177391	0.042913	0.214513	0.034491	0.042737	0.065443	0.335826	0.103885

In the aggregation method, we grouped the "Gender" and "Selector" 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 selected or not selected. By aggregating the data in this way, we can identify any patterns or differences in attribute means based on gender and selection 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:

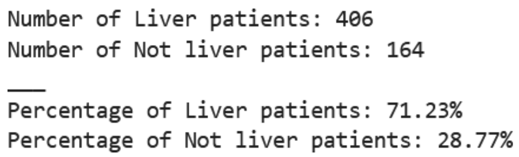


```
          Age  Gender   TB     DB  Alkphos   Sgpt   Sgot     TP   ALB  \
0     Seniors       0  0.7  0.010    0.187  0.016  0.018  0.068  0.33
1     Seniors       1  5.3  0.295    0.481  0.064  0.100  0.075  0.32
2     Seniors       1  5.3  0.295    0.481  0.060  0.068  0.070  0.33
3     Seniors       1  1.0  0.040    0.182  0.014  0.020  0.068  0.34
4     Seniors       1  3.9  0.200    0.195  0.027  0.059  0.073  0.24
..        ...     ...  ...    ...      ...    ...    ...    ...   ...
565   Seniors       1  0.5  0.010    0.481  0.020  0.034  0.059  0.16
566    Adults       1  0.6  0.010    0.098  0.035  0.031  0.060  0.32
567    Adults       1  0.8  0.020    0.245  0.048  0.049  0.064  0.32
568  Children       1  1.3  0.050    0.184  0.029  0.032  0.068  0.34
569    Adults       1  1.0  0.030    0.216  0.021  0.024  0.073  0.44
```

In the discretization method, we categorize numerical age values into three groups:Children (0-17 years), Adults (18-64 years), and Seniors (65+ years). This simplifiesdata 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:**

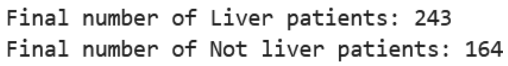
Before starting the Data Mining Technique, we investigated whether the data was balanced or not:



```
Number of Liver patients: 406
Number of Not liver patients: 164
___
Percentage of Liver patients: 71.23%
Percentage of Not liver patients: 28.77%
```

In the beginning, we reviewed the percentage for each of the two classes in the Liver Class (Liver patients, non-liver patients), and we noticed that the percentage is imbalanced (not ranging between 40% to 60%).

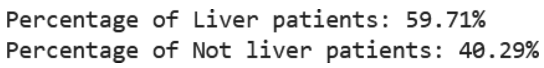
- Process of correcting data balancing



```
Final number of Liver patients: 243
Final number of Not liver patients: 164
```

- Data after the balancing process:

By using the "resample" function, we reduced the number of samples in the majority class (patients) to achieve balance between the two classes. This helps prevent the model from being biased towards the majority class and improves its ability to generalize to new data.



```
Percentage of Liver patients: 59.71%
Percentage of Not liver patients: 40.29%
```

We finally calculated the percentage for each class to ensure that the data has become balanced. The two classes represent liver patients and non-liver patients, and it is indeed balanced as the percentage of each class ranges from 40% to 60%.

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 has liver disease (selector), categorizing the results into(‘1’ that means have liver disease) or (‘0’ that means not have liver disease) . It makes predictions based on several attributes:(age, gender, TB,DB,TP,Alkphos,Sgpt,Sgot,A/G Ratio,ALB).

As we touched on before, classification is a type of supervised learning, so we need training data to train the model, so we split our dataset into two subsets which are training data and testing data. we tried 3 different sizes of training subsets which are 70%, 60%, and 80% and use two attribute selection measures (IG (entropy) Gini index) .To evaluate our model and determine the better partitioning we look at its accuracy and useing a confusion matrix that is summarizes the basic measures for performance evaluation like sensitivity, specificity, precision, and error rate.

In the clustering process, which is a type of unsupervised learning, we omitted the "selector" class label attribute since it does not use class labels. Instead, we utilized all other attributes such as:(age, gender, TB,DB,TP,Alkphos,Sgpt,Sgot,A/G Ratio,ALB).all of which are numeric and require no conversion prior to clustering. For creating the clusters, we employed the K-means algorithm. This algorithm generates K clusters, each represented by the centroid of the cluster. It assigns each object to the closest cluster, then iteratively recalculates the centroids and reassigns the objects until the centroids stabilize, indicating correct cluster assignment.
For cluster validation, we calculated the average silhouette score of each cluster using theAverage Silhouette Score method and visualized these scores. Additionally, we used WSS method to compare three different cluster sizes to determine the optimal number by assessing the separation and compactness of the clusters.

6- **Evaluation and Comparison:**

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

Figure (1) (decision tree):

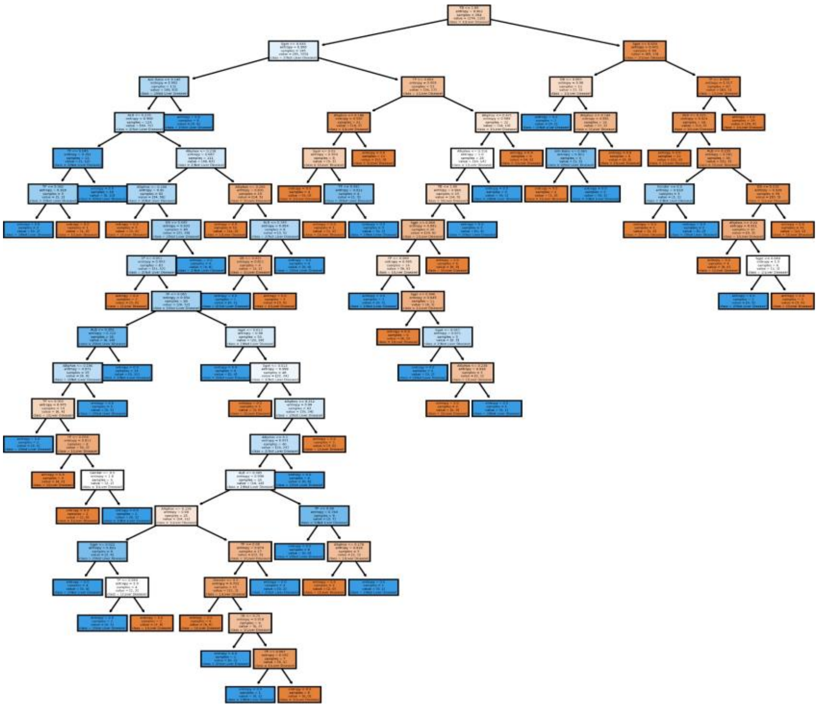
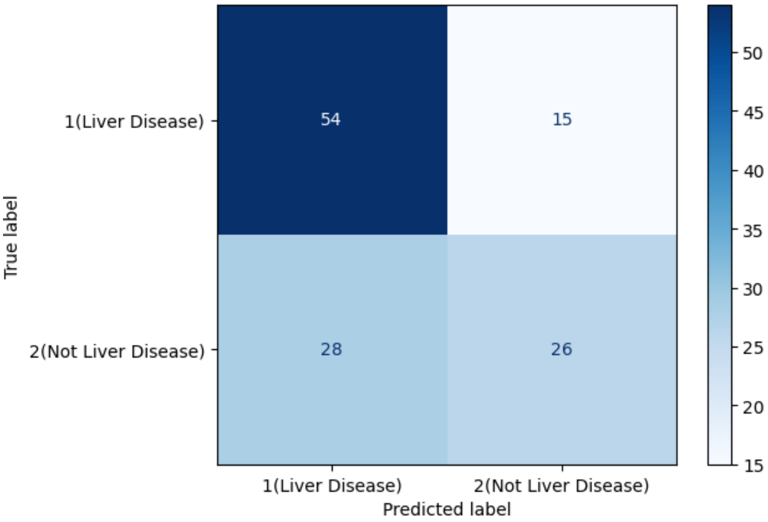


Figure (2) (confusion matrix):



- Classification [60% Training and 40% Test] Information Gain:

Figure (1) (decision tree):

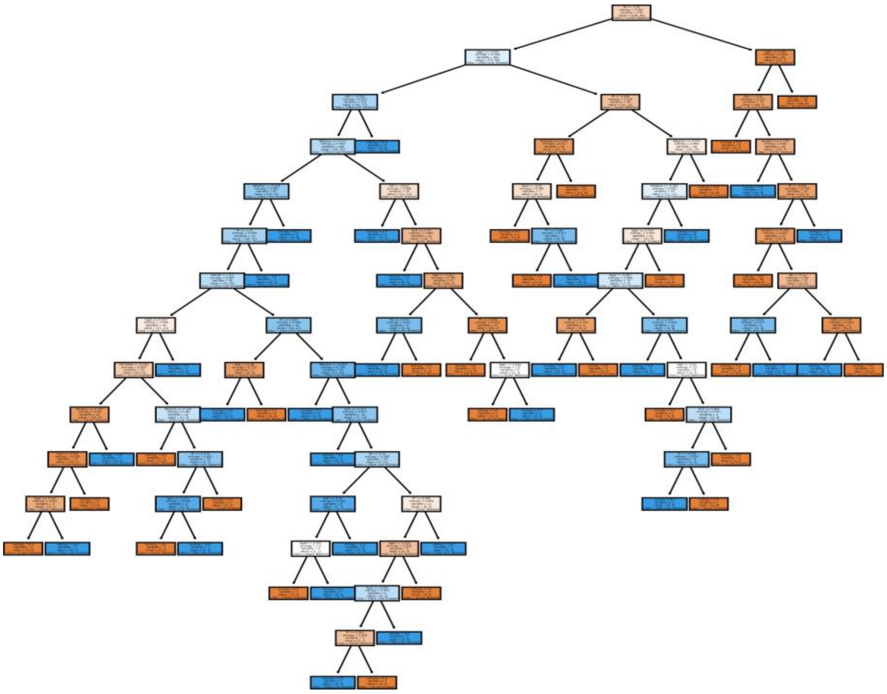
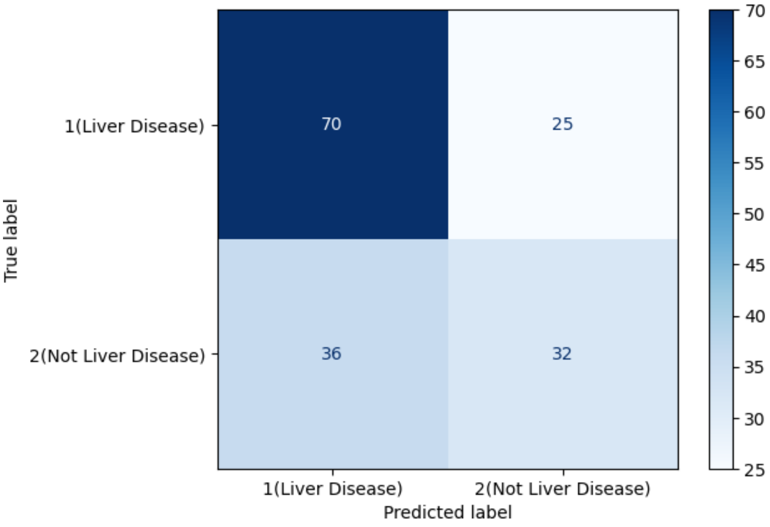


Figure (2) (confusion matrix):



- Classification [80% training and 20% test] Information Gain:

Figure (1) (decision tree):

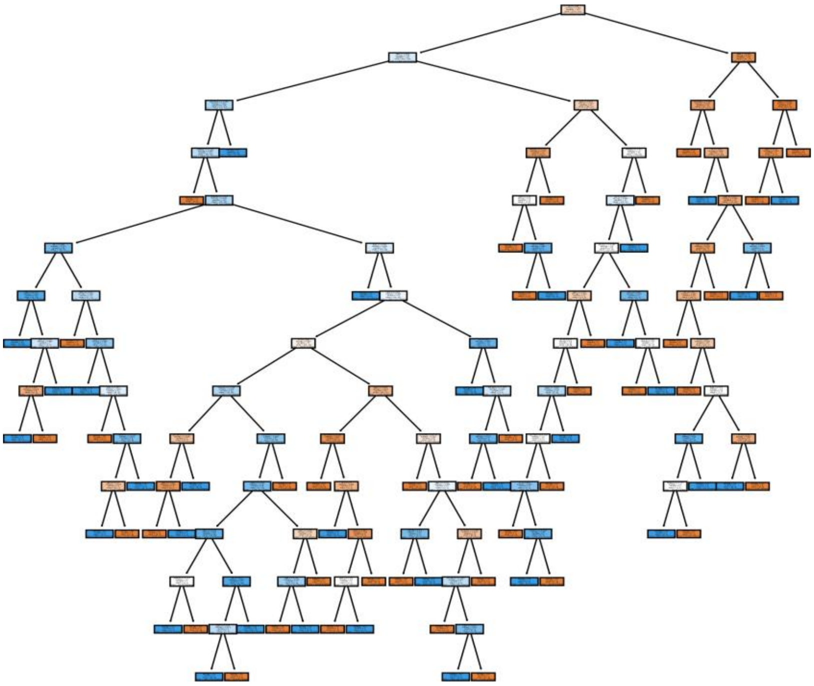
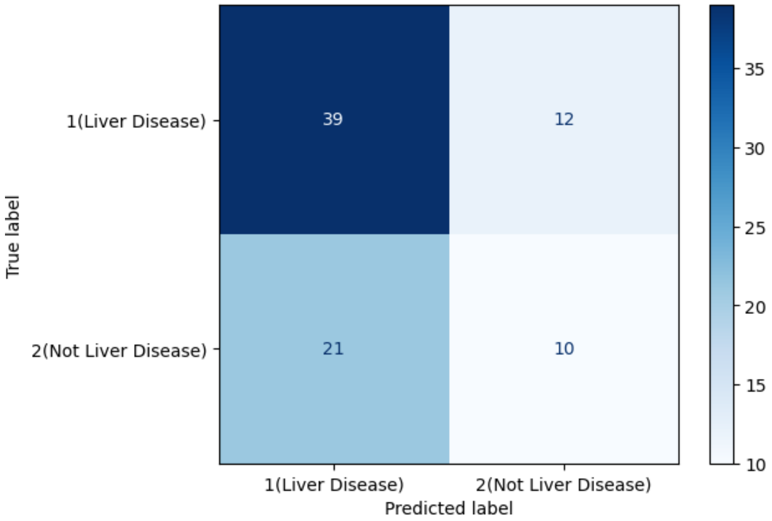


Figure (2) (confusion matrix):



<table>
<tr><th>Mining task</th><th>Comparison Criteria</th></tr>
<tr><td>Classification for
Information Gain</td><td>
We tried 3 different sizes for dataset splitting to create thedecision tree:
- 70% Training data, 30% Test data.
<table>
<tr><th></th><th></th></tr>
<tr><td>Accuracy</td><td>65%</td></tr>
<tr><td>precision</td><td>63%</td></tr>
<tr><td>sensitivity</td><td>48%</td></tr>
<tr><td>specificity</td><td>78%</td></tr>
<tr><td>Error rate</td><td>34%</td></tr>
</table>
- 60% Training data, 40% Test data.
<table>
<tr><th></th><th></th></tr>
<tr><td>Accuracy</td><td>62.5%</td></tr>
<tr><td>precision</td><td>56%</td></tr>
<tr><td>sensitivity</td><td>47%</td></tr>
<tr><td>specificity</td><td>73%</td></tr>
<tr><td>Error rate</td><td>37.4%</td></tr>
</table>
</td></tr>
</table>

-80% Training data, 20% Test data.

Accuracy	62.1%
precision	50%
sensitivity	38%
specificity	76%
Error rate	37.8%

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

Figure (1) (decision tree):

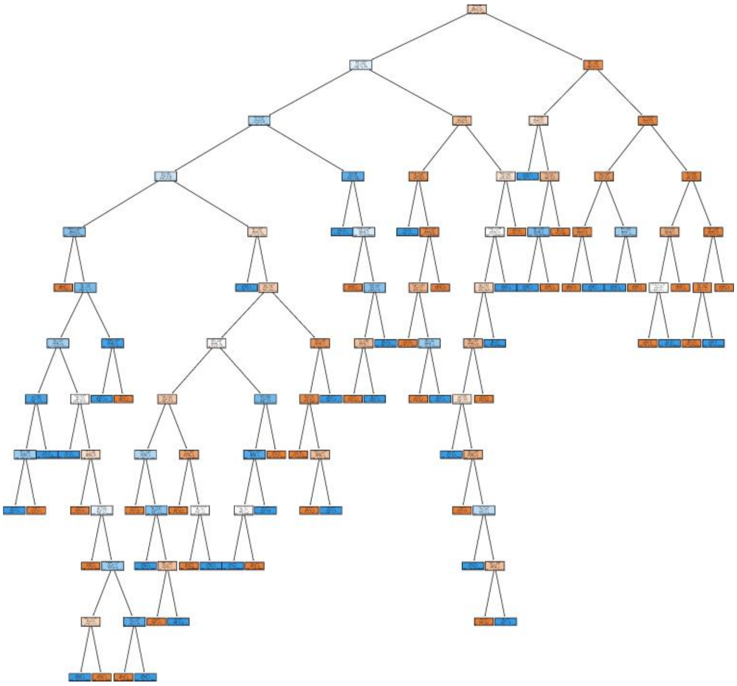
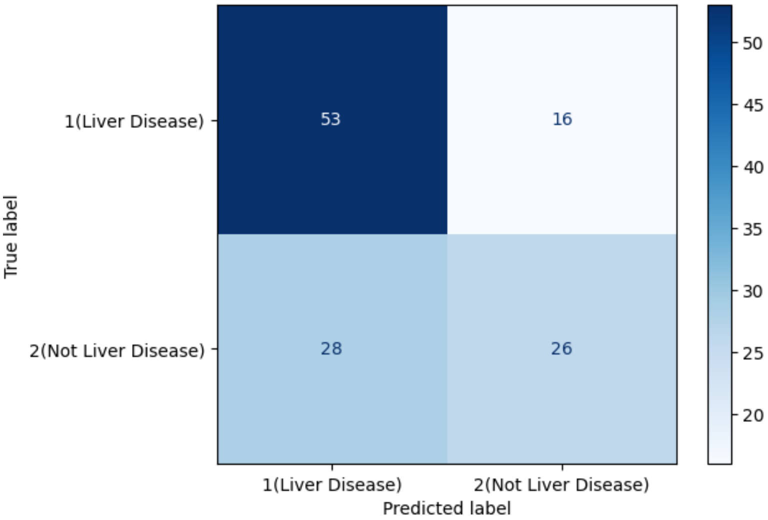


Figure (2) (confusion matrix):



- Classification [60% Training and 40% Test] Gini Index:

Figure (1) (decision tree):

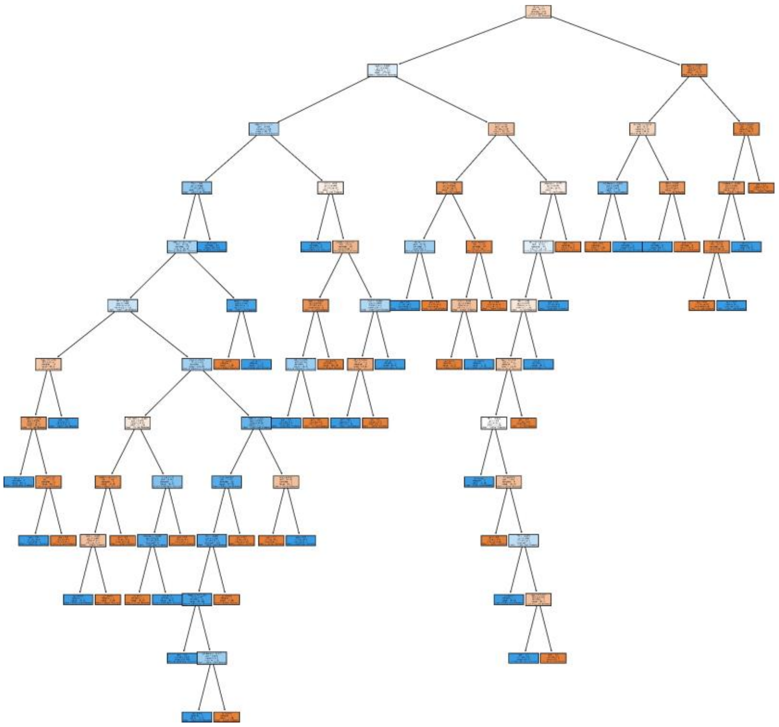
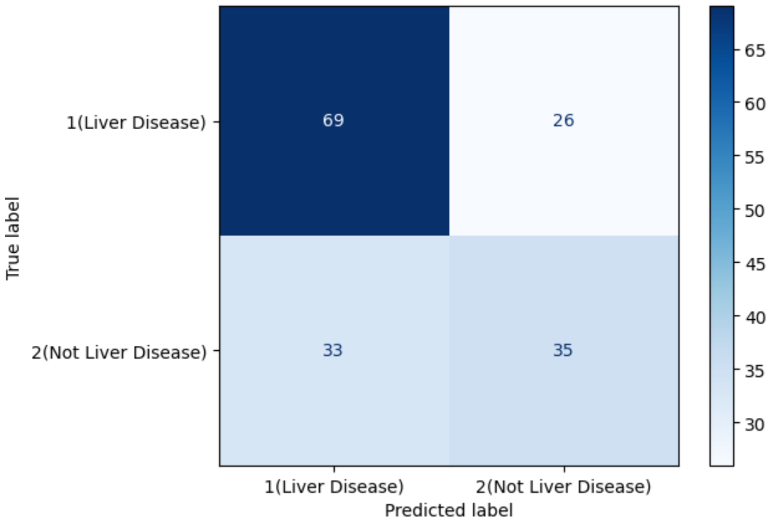


Figure (2) (confusion matrix):



- Classification [80% training and 20% test] Gini Index:

Figure (1) (decision tree):

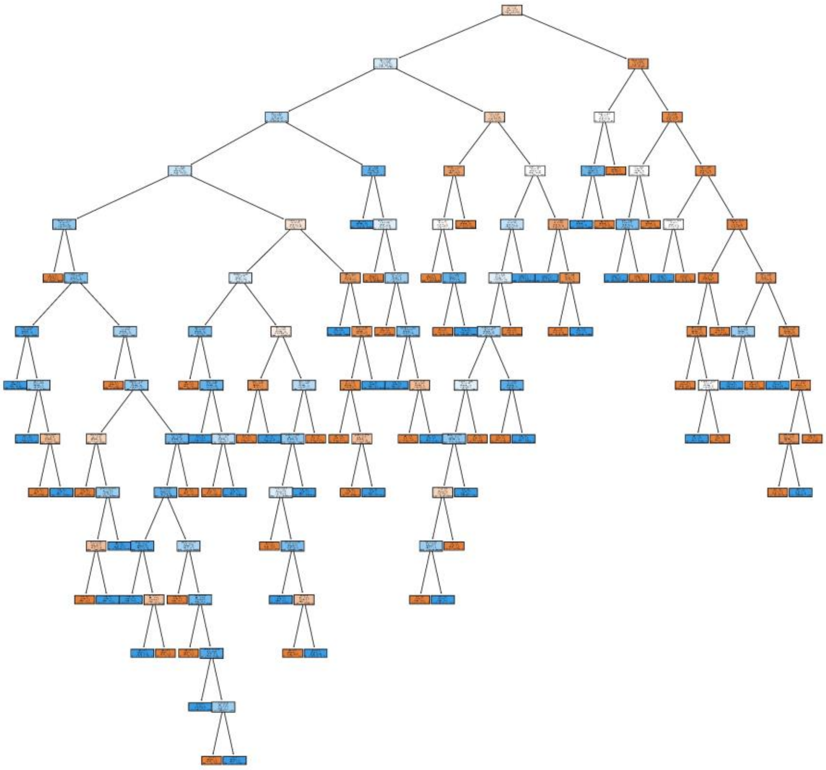
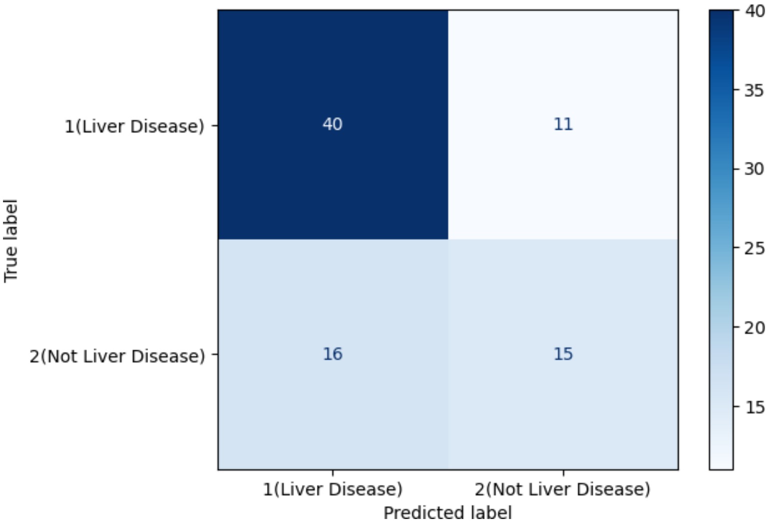


Figure (2) (confusion matrix):



<table>
<tr><th>Mining task</th><th>Comparison Criteria</th></tr>
<tr><td>Classification for
Gini Index</td><td>
<p>We tried 3 different sizes for dataset splitting to create thedecision tree:</p>
<p>- 70% Training data, 30% Test data.</p>
<table>
<tr><th></th><th></th></tr>
<tr><td>Accuracy</td><td>64%</td></tr>
<tr><td>precision</td><td>61%</td></tr>
<tr><td>sensitivity</td><td>48%</td></tr>
<tr><td>specificity</td><td>76%</td></tr>
<tr><td>Error rate</td><td>35%</td></tr>
</table>
<p>- 60% Training data, 40% Test data.</p>
<table>
<tr><th></th><th></th></tr>
<tr><td>Accuracy</td><td>63%</td></tr>
<tr><td>precision</td><td>57.3%</td></tr>
<tr><td>sensitivity</td><td>51%</td></tr>
<tr><td>specificity</td><td>72%</td></tr>
<tr><td>Error rate</td><td>36%</td></tr>
</table>
</td></tr>
</table>

-80% Training data, 20% Test data.

Accuracy	67%
precision	57.6%
sensitivity	48%
specificity	78%
Error rate	32%

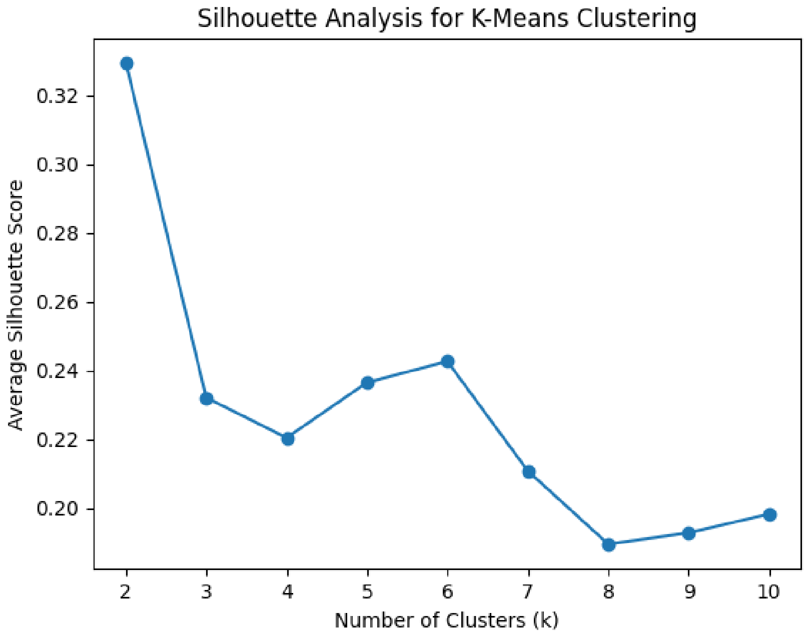
- **the better partitioning:**

In summary, the 80%-20% split using the Gini Index yields better overall performance, with high accuracy, low error rate, and high values for sensitivity, specificity, and precision. which is why it is considered the best based on the provided results.

- **Clustering**
  We choose 3 different sizes [2,3,6] based on the result of the validation methods that we will apply then we will use these sizes to perform the k-means clustering.

Silhouette method:

The Silhouette method is a technique used to evaluate the quality of clustering results. It measures how well each data point fits within its assigned cluster compared to neighboring clusters.



Elbow method:

The Elbow method is a technique used to determine the optimal number of clusters in a dataset for K-means clustering.

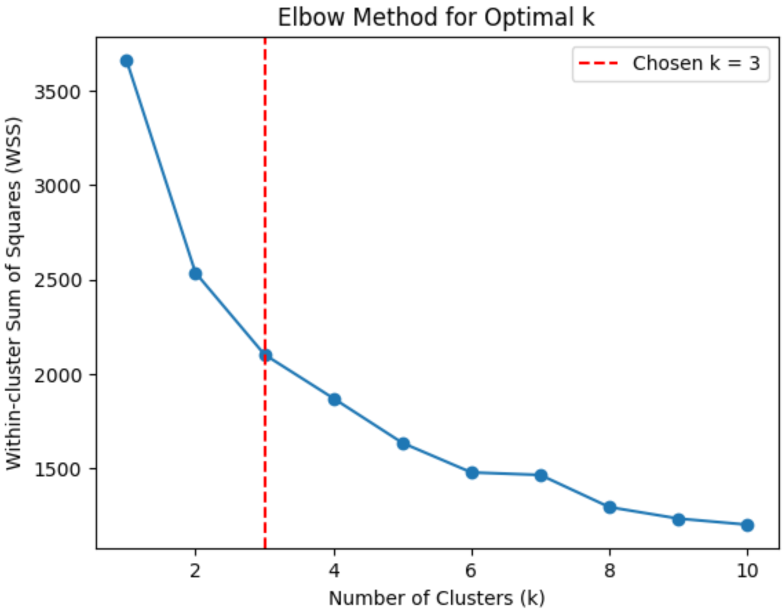


Figure (1): silhouette scores [K=2]

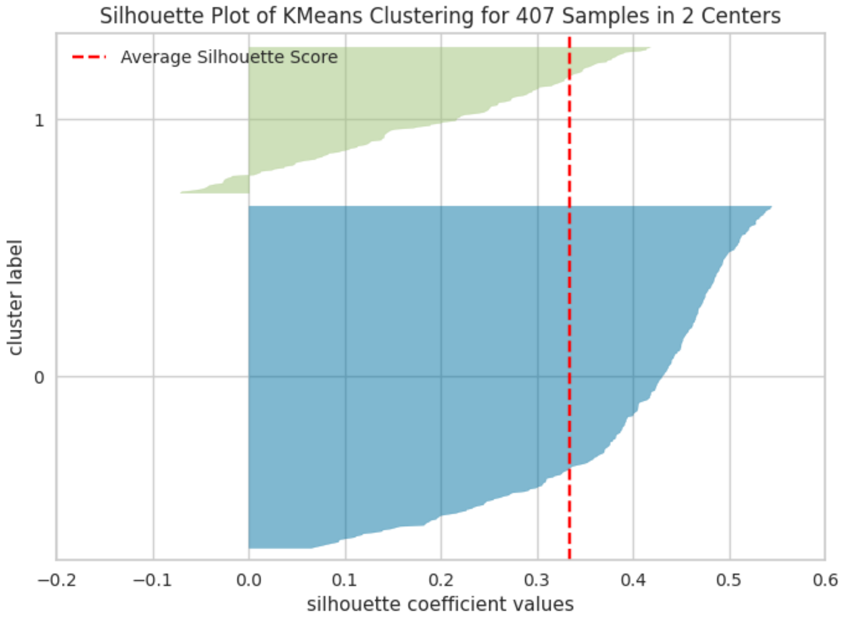


Figure (2): silhouette scores [K=3]

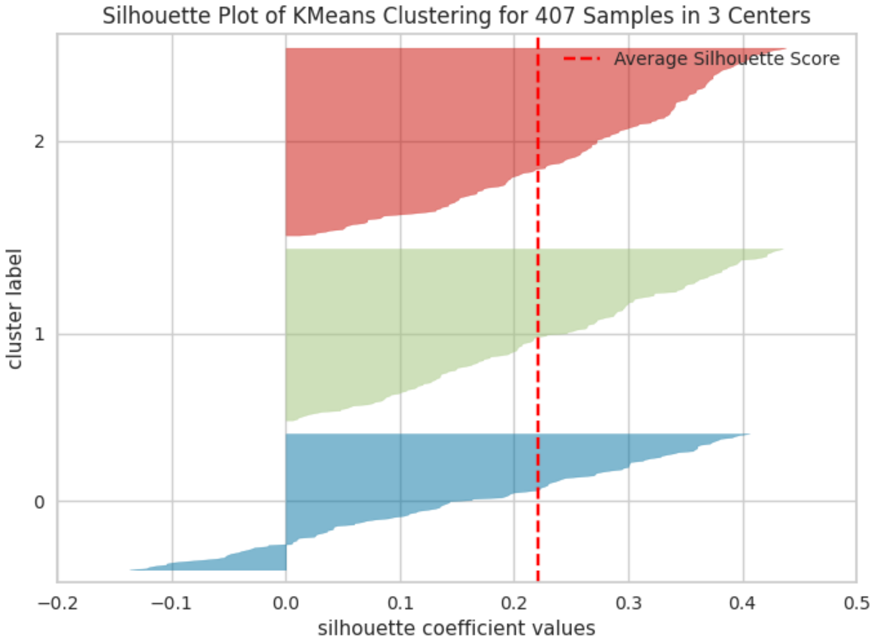
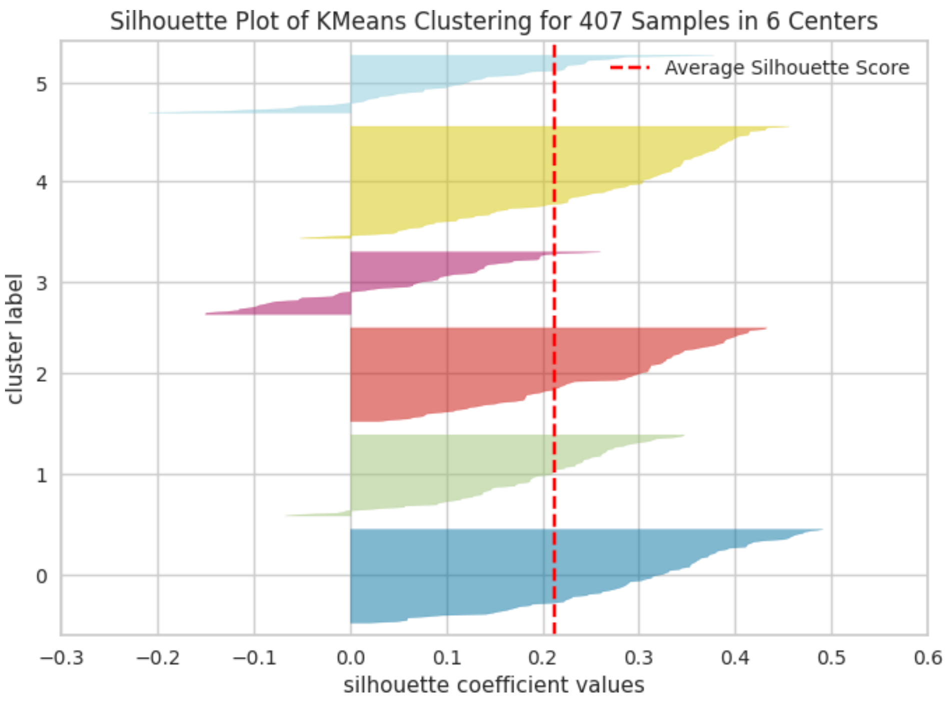


Figure (3): silhouette scores [K=6]



Mining task	Comparison Criteria			
Clustring	We tried 3 different sizes for dataset splitting to create the decision tree: K=2, K=3, K=6			
	No. of clusters	K=2 (BEST)	K=3	K=6
	Average Silhouette width	0.329	0.232	0.2427
	total within-cluster sum of square	2537.0	2125.6	1526.79

7. Findings:

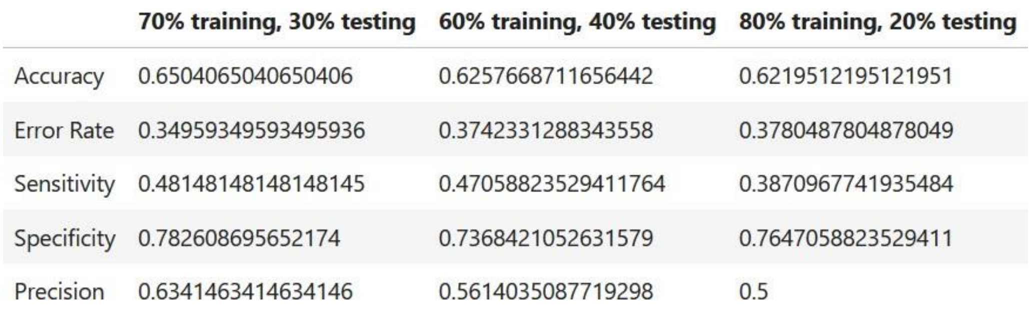
Initially, we selected a dataset representing patients diagnosed with liver disease with the aim of understanding the causes of this prevalent condition and implementing appropriate preventive measures.

To ensure effectiveness, accuracy, and maximum precision in our results, we applied several data processing techniques to enhance data efficiency. Utilizing various visualization methods such as box plots, scatter plots, and line graphs, we clarified the data and facilitated comprehension, enabling the application of suitable data processing techniques. Based on these visualizations and other analyses, we removed all empty, missing, and outlier values that could potentially impact the results negatively.

Furthermore, we implemented data transformations, including normalization، feature partitioning, and balanced data process to assign equal weight to certain features and streamline data processing during mining tasks.

Consequently, we conducted data mining tasks, encompassing classification and partitioning. For classification, we employed the Gini index and information gain metrics. Experimenting with three different sizes of training and testing data allowed us to achieve optimal results for both model construction and evaluation. Here are our findings:

- **Information Gain:**



	70% training, 30% testing	60% training, 40% testing	80% training, 20% testing
Accuracy	0.6504065040650406	0.6257668711656442	0.6219512195121951
Error Rate	0.34959349593495936	0.3742331288343558	0.3780487804878049
Sensitivity	0.48148148148148145	0.47058823529411764	0.3870967741935484
Specificity	0.782608695652174	0.7368421052631579	0.7647058823529411
Precision	0.6341463414634146	0.5614035087719298	0.5

Based on the presented results for the models trained using the Information Gain criterion, the following observations can be made:

-Accuracy: The model trained with a 70% training set and 30% testing set achieved the highest accuracy (65%). This indicates that the 70-30 split model performs slightly better in terms of overall accuracy.

-Error Rate: The model trained with an 80% training set and 20% testing set exhibited the highest error rate (37.8%). Hence, the 70-30 split model has the lowest error rate, suggesting better performance in minimizing classification errors.

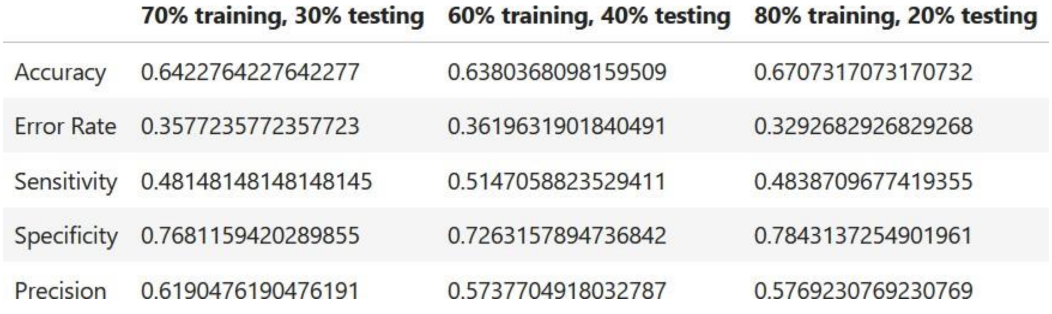
-Sensitivity: The model trained with a 70-30 split achieved the highest sensitivity (48%). This implies that the 70-30 split model is more effective in correctly identifying positive instances.

-Specificity: The model trained with a 70-30 split obtained the highest specificity (78%). Thus, the 70-30 split model exhibits better performance in correctly identifying negative instances.

-Precision: The model trained with a 70-30 split achieved the highest precision (63%). This indicates that the 70-30 split model is more accurate in predicting positive instances.

In summary, the model trained with a **<u>70% training set and 30% testing</u>** set generally performs better across various evaluation metrics compared to the other partitioning schemes.

- **Gini index:**



	70% training, 30% testing	60% training, 40% testing	80% training, 20% testing
Accuracy	0.6422764227642277	0.6380368098159509	0.6707317073170732
Error Rate	0.3577235772357723	0.3619631901840491	0.3292682926829268
Sensitivity	0.48148148148148145	0.5147058823529411	0.4838709677419355
Specificity	0.7681159420289855	0.7263157894736842	0.7843137254901961
Precision	0.6190476190476191	0.5737704918032787	0.5769230769230769

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

-Highest Accuracy: The model trained using an 80% training and 20% testing split achieved the highest accuracy rate among the three compared models. This means it can predict class labels more accurately compared to the other models.

-Highest Specificity: The model achieving the best specificity for negative classification (the negative class) is crucial because it signifies its ability to avoid errors in classifying negative instances. Therefore, the model with the highest specificity can be more reliable in predicting the absence of the condition.

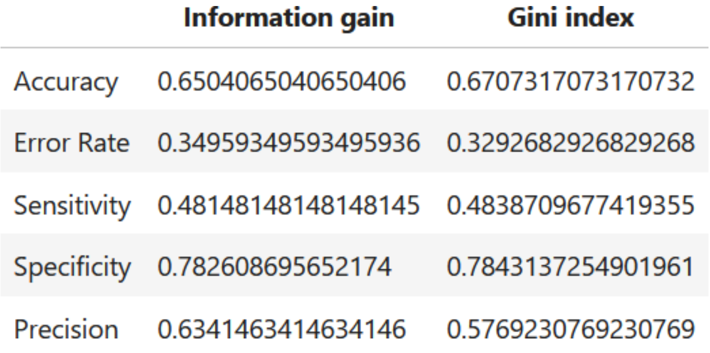
-Balance between Sensitivity and Precision: Although the sensitivity for the 80-20 model is slightly lower compared to some other models (around 48%), it still remains at an acceptable level, indicating its ability to identify positive class instances effectively. Additionally, it achieves high accuracy, meaning it can correctly predict the classification of instances.

-Lowest Error Rate: The 80-20 model has the lowest error rate among the three models, indicating 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.

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

After selecting the best model split from Information Gain, which was 70% training, 30% testing, and the best split from Gini Index, which was 80% training, 20% 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	0.6504065040650406	0.6707317073170732
Error Rate	0.34959349593495936	0.3292682926829268
Sensitivity	0.48148148148148145	0.4838709677419355
Specificity	0.782608695652174	0.7843137254901961
Precision	0.6341463414634146	0.5769230769230769

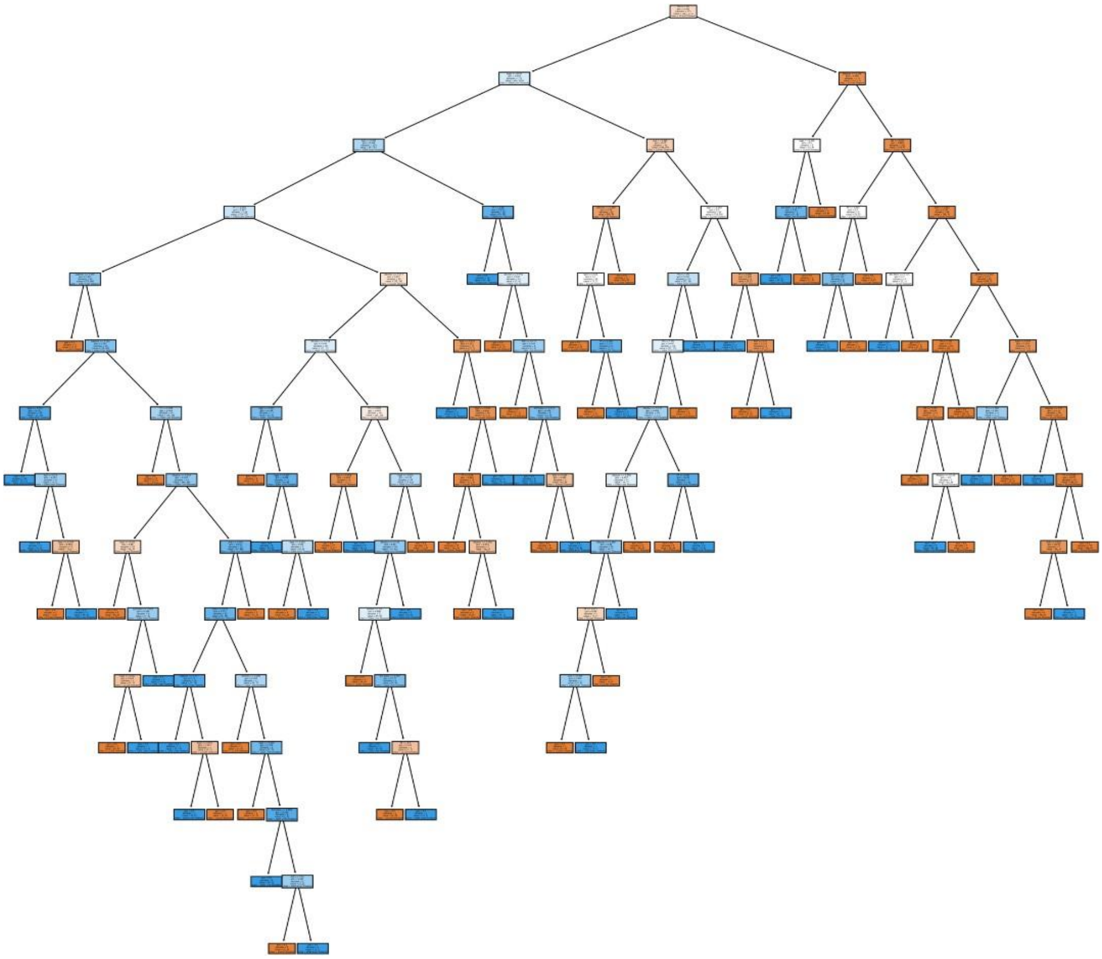
-Accuracy and Error Rate: The Gini Index split provides higher accuracy (67.07% or o.67) compared to Information Gain (65.04% or 0.65), resulting in a lower error rate of (32.93% or 0.32) for Gini Index compared to (34.96% or 0.34) for Information Gain. This indicates that the Gini Index model classifies cases more accurately, making it more reliable.

-Sensitivity and Specificity: Information Gain split slightly outperforms in sensitivity (48.15% or 0.48) compared to Gini Index (48.39% or 0.48), but there's a marginal difference. However, the Gini Index split achieves higher specificity (78.43% or 0.7843) compared to Information Gain (78.26% or 0.7826). Specificity reflects the model's ability to correctly identify negative cases, making the Gini Index model more reliable in predicting negative instances.

-Precision: The Gini Index split achieves lower precision (57.69% or 0.57) compared to Information Gain (63.41% or 0.63), meaning when the model predicts positive cases, it's correct (57.69% or 0.57) of the time compared to (63.41% or 0.63) for Information Gain. However, both values remain high and acceptable.

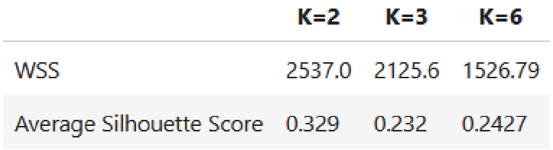
Based on these reasons, it can be concluded that the **<u>80%-20% split</u>** using the Gini Index yields better overall performance, with high accuracy, low error rate, and high values for sensitivity, specificity, and precision.

This was the decision tree associated with this division:



Show us the decision tree for predicting liver disease is built on the importance of Total Bilirubin (TB) and further splits on features such as Sgot, Alkphos, ALB, TP, DB, and Sgpt. The tree's structure reveals complex decision pathways based on combinations of these features, highlighting the multifaceted nature of liver disease prediction. Terminal nodes provide the final predicted outcome (1 for liver disease, 2 for no liver disease) based on the feature values and their importance. The model heavily relies on Total Bilirubin, Alkaline Phosphatase, Serum Glutamic Oxaloacetic Transaminase, and Serum Glutamic Pyruvic Transaminase for accurate predictions, considering specific combinations of these features. The depth and complexity of the decision tree demonstrate the diverse factors considered by the model in assessing liver disease. Understanding the decision tree provides valuable insights into the model's inner workings and its ability to predict liver disease.

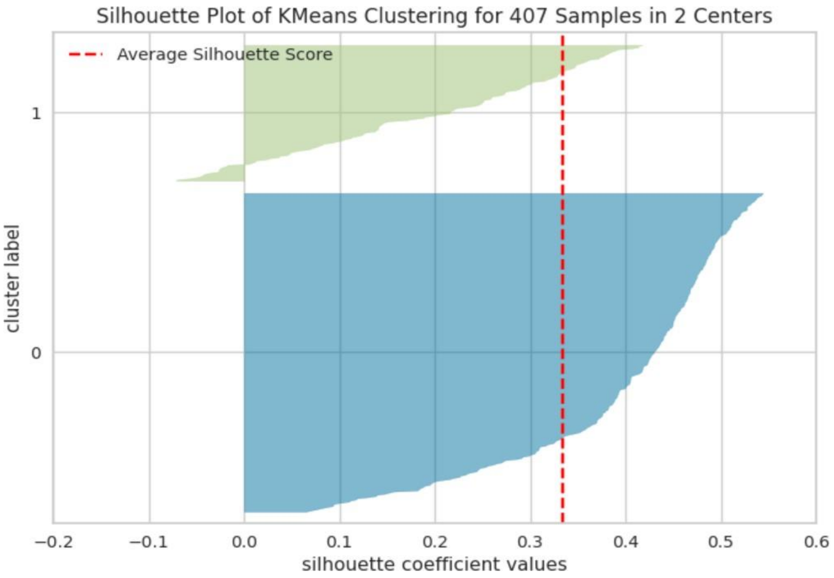
For Clustering, we used K-means algorithm with 3 different K to find the optimal number of clusters, we calculated the average silhouette width for each K, and we concluded the following results:



	K=2	K=3	K=6
WSS	2537.0	2125.6	1526.79
Average Silhouette Score	0.329	0.232	0.2427

We've decided that K=2 is the best choice for our clustering model based on the metrics we've analyzed(WSS, Average Sihouette Score, Visualization of K-mean). This choice is because K=2 gives the highest silhouette width, also k=2 have a highest value of WSS Comparison of WSS value for K=3,k=6
Also, having a silhouette plot of kmeans clustring of 407 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:



From the graph of KMeans Clustering for 407 Samples in 2 Centers, the fact that most of the silhouette scores with a positive value reinforces the notion that the samples are well-matched to their clusters and are distant from neighboring clusters. This indicates that the clustering solution has successfully separated the data points into distinct and well-defined clusters.
Note that while most silhouette scores being positive is a positive indicator, it does not necessarily imply that the clustering solution is "extremely perfect" or flawless. There might still be some degree of overlap or ambiguity between clusters, especially if there are samples as above in the first center with silhouette scores close to 0 or negative values.

Finally, both models are useful in predicting whether a person may develop liver disease or not, helping us achieve our goal of understanding the underlying causes of the disease - such as elevated enzyme levels and others. However, since our data includes a "Selector" class category indicating whether a person is affected or not by liver disease, 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:

- Fatemeh Mehrparvar, "Liver Disorders Dataset", Kaggle, Available: https://www.kaggle.com/datasets/fatemehmehrparvar/liver-disorders

- "Liver disease accounts for two million annual deaths globally: the need for joint and robust policy and practices across liver diseases." PubMed, Available: https://pubmed.ncbi.nlm.nih.gov/36990226/#:~:text=Liver%20disease%20accounts%20for%20two,related%20deaths%20occur%20in%20men

- American Liver Foundation, "How Many People Have Liver Disease?" Available: https://liverfoundation.org/about-your-liver/facts-about-liver-disease/how-many-people-have-liver-disease/

- “Labs and Lecture Slides," College of Computer Science, Department of Information Technology, King Saud University.