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

IT 326 Data Mining Project

Diabets Disease

Final Report

Group#1

Lab Day-Time: Monday-10

Group#	1	
Section	71681	
Mambers	Name	ID
	Lujain Alhusan	444200785
	Walaa Saif Aleslam	444200088
	Lama Alkhathlan	444200844
	Noura Alzomai	444200503

1. Problem

Recently, the prevalence of diabetes and related health conditions has been increasing, becoming more common among individuals worldwide. This trend leads to numerous serious health complications, potentially resulting in chronic illnesses and reduced quality of life.

In our project, we aim to study and analyze patient data, which includes attributes such as (Pregnancies, Glucose levels, Blood Pressure, Skin Thickness, Insulin levels, BMI, Diabetes Pedigree Function, and Age**)**. By clustering individuals based on these attributes, we can identify possible patterns and risk factors associated with diabetes.

Through this analysis, we can provide valuable insights that help predict the likelihood of developing diabetes, enabling individuals to take preventive measures to safeguard their health and reduce the burden on healthcare systems. This study is critical for leveraging data to improve health outcomes and promote proactive management of diabetes risks.

2.*Data Mining Task*

In our project, we will employ two data mining tasks to help predict the likelihood of diabetes: classification and clustering. For classification, we will train our model to determine whether a patient has diabetes or not, based on a set of medical features such as glucose levels, BMI, age, blood pressure, and other related attributes. Classification will be based on the "diabetes" class.

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

3.*Data*

Dataset: Dataset link(click)

-Number of attributes: 10

- Number of objects:2769

-Class label: Outcome

Types of Attributes:

1) Id: Nominal

2) Pregnancies: Numeric ratio integer

3) Glucose: Numeric interval integer

4) BloodPressure: Numeric interval integer

5) SkinThickness: Numeric ratio integer

6) Insulin: Numeric ratio integer

7) BMI: Numeric ratio float

8) DiabetesPedigreeFunction: Numeric interval float

9) Age: Numeric ratio integer

10) Outcome: Nominal (Binary)

Columns Description:

Id: Unique identifier for each data entry.

Pregnancies: Number of times pregnant.

Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.

BloodPressure: Diastolic blood pressure (mm Hg).

SkinThickness: Triceps skinfold thickness (mm).

Insulin: 2-Hour serum insulin (mu U/ml).

BMI: Body mass index (weight in kg / height in m^2).

DiabetesPedigreeFunction: Diabetes pedigree function, a genetic score of diabetes.

Age: Age in years.

Outcome: Binary classification indicating the presence (1) or absence (0) of diabetes.

Missing Values:

1) Id: 0

2) Pregnancies: 0

3) Glucose: 0

4) BloodPressure: 0

5) SkinThickness: 0

6) Insulin: 0

7) BMI: 0

8) DiabetesPedigreeFunction: 0

9) Age: 0

10) Outcome: 0

Statical Measures for each numeric column:

-Five Number Summary:

- Id: The ID values range from 1 to 2768, indicating a unique identifier for each individual in the dataset.

- Pregnancies: The number of pregnancies varies from 0 to 17, with a mean of approximately 3.74. This suggests a diverse reproductive history among the individuals.

- Glucose: Glucose levels range from 0 to 199, with an average of 121.10. The standard deviation of 32.04 indicates considerable variability, which may suggest the presence of outliers.

- Blood Pressure: Blood pressure measurements range from 0 to 122, with a mean of 69.13. The variability in blood pressure values, as indicated by the standard deviation of 19.23, reflects different health statuses among individuals.

- Skin Thickness: Skin thickness shows a range from 0 to 99, with a mean of 20.82. This wide range and the standard deviation of 16.06 indicate significant variation in skin thickness measurements.

- Insulin: Insulin levels vary from 0 to 110, with a mean of 80.13. The high standard deviation of 112.30 suggests the presence of extreme values or outliers.

- BMI: Body Mass Index (BMI) ranges from 0.0 to 80.60, with a mean of 32.14. The standard deviation of 8.08 indicates variability in body weight relative to height.

- Diabetes Pedigree Function: This metric ranges from 0.08 to 2.42, with a mean of 0.47. The standard deviation of 0.33 indicates varying genetic predispositions to diabetes among individuals.

- Age: Ages in the dataset range from 21 to 81 years, with an average age of 33.13. The variability in age (standard deviation of 11.78) points to a diverse age distribution.

- Outcome: The outcome variable is binary, with values of 0 (non-diabetic) and 1 (diabetic). The mean of 0.34 indicates that approximately 34% of the individuals in the dataset are classified as diabetic.

These summary statistics provide a comprehensive overview of the dataset, highlighting the variability and distribution of key attributes related to diabetes and other health indicators.

Variance:

A low variance indicates that the data points tend to be close to the mean, suggesting consistency , like BMI and DiabetesPedigreeFunction.

1) Pregnancies: 11.047653

2) Glucose: 1026.337861

3) BloodPressure: 369.848213

4) SkinThickness: 257.910614

5) Insulin: 12611.724151

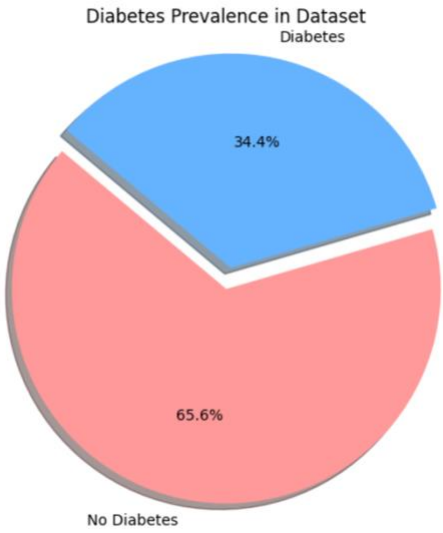
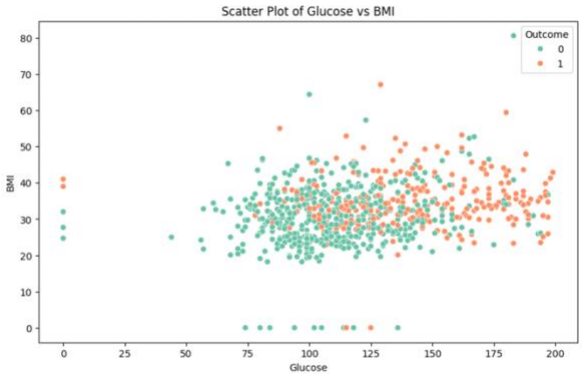
6) BMI: 65.223831

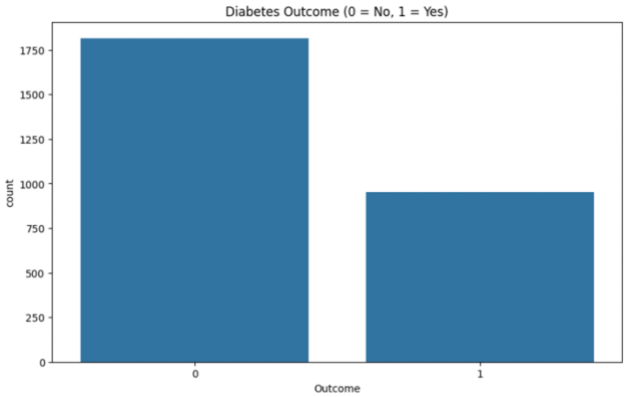
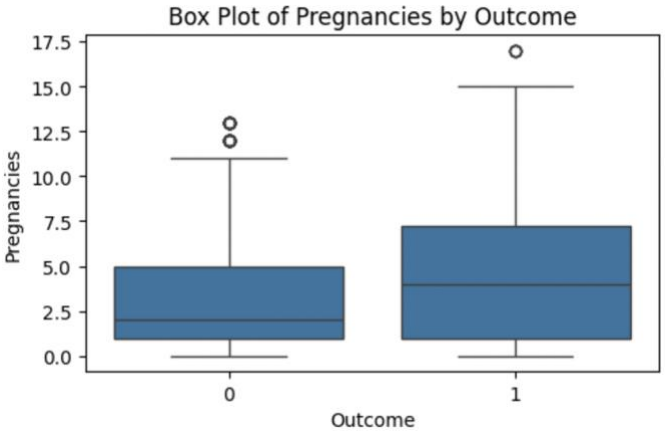
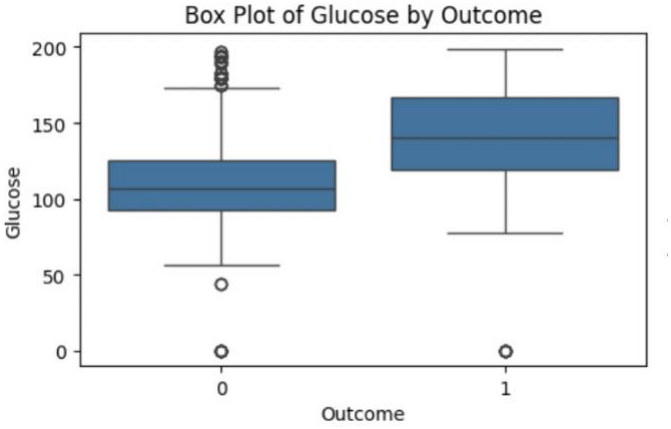
7) DiabetesPedigreeFunction: 0.106060

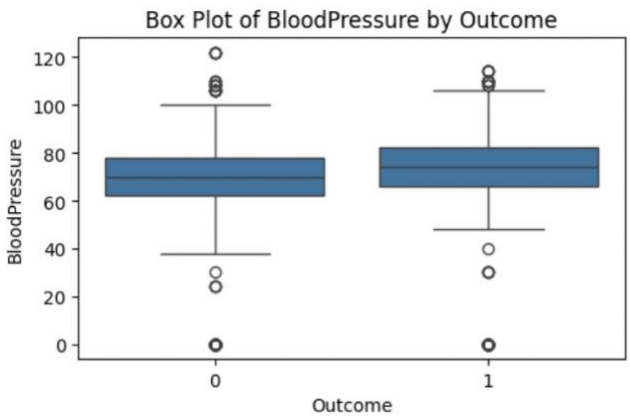
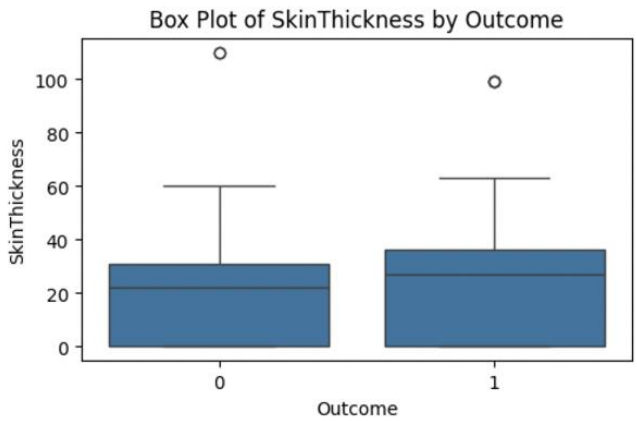
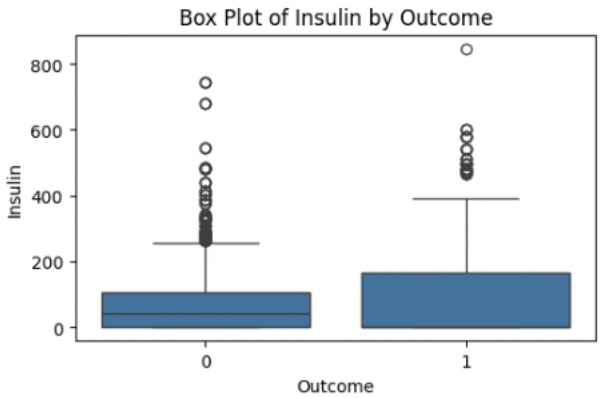
8) Age: 138.703146

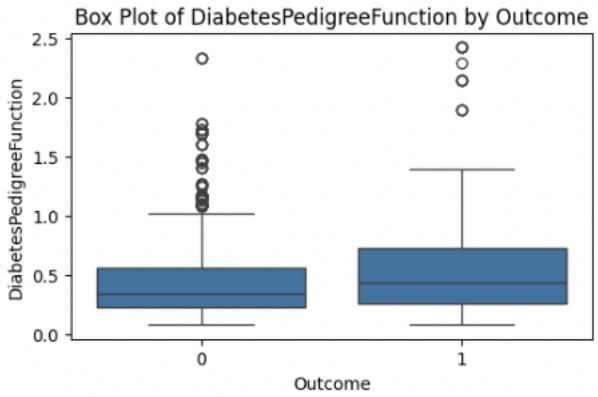
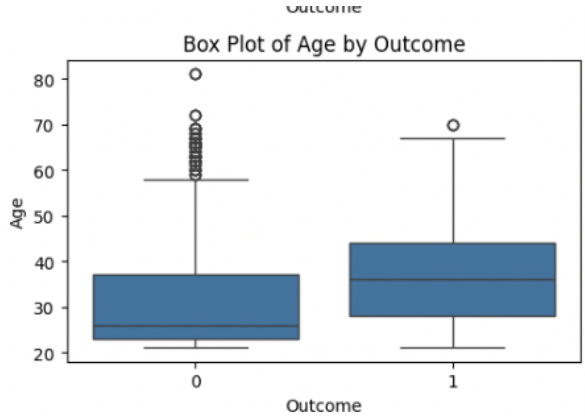
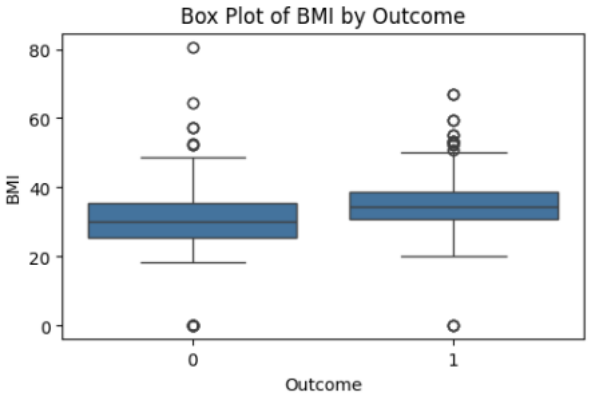
A high variance indicates that the data points are spread out over a wider range of values, suggesting more variability ,like Glucose and Insulin.

Graph's description:

Name of grap	**Picture of grap**	**Descriptio**
Pie chart		the pie chart illustrates the percentage of diabetes sufferers, which stands at 34.4%. The percentage of non-diabetes sufferers represents 65.6%, making it the larger proportion
Scatter plot		The scatter plot uses colors to distinguish between diabetic (orange) and non-diabetic (green) individuals. It shows a positive correlation between glucose levels (x-axis) and BMI (y-axis), with most non-diabetics clustering in a healthy glucose range. Outliers indicate unusual glucose or BMI levels.

Bar chart		Bar chart illustrating the classification of individuals with and without diabetes. It shows that a large number of people are not diabetic, with their count exceeding 1,750, while those diagnosed with diabetes number fewer than 1,000.
Box plot of pregnancy by outcom		-Median: Non-diabetic individuals have a higher median number of pregnancies than diabetics. - Variability: Non-diabetics show greater variability (wider IQR) and more outliers. - Range: The range of pregnancies is broader for non-diabetics. Non-diabetic individuals tend to have more pregnancies, indicating a possible link to diabetes status.
Box plot of Glucose by outcom		-Median: Diabetic individuals have a higher median glucose level than non-diabetics. -Variability: Diabetics show greater variability (wider IQR) and more outliers. -Range: Non-diabetics have a lower maximum glucose level. Diabetic individuals tend to have higher and more variable glucose levels, suggesting a significant distinction between the groups.

Box plot of BloodPressure by outcom		- Median: Diabetic individuals (Outcome 1) have a higher median blood pressure than non-diabetics (Outcome 0). - Variability: The interquartile range (IQR) is similar for both groups, indicating comparable variability in blood pressure. - Outliers: Both groups have outliers, with more extreme values noted in the diabetic group**.** Diabetic individuals tend to have higher blood pressure, highlighting a notable difference between the two groups.
Box plot of SkinThickness by outcom		- Median: The median skin thickness is similar for both diabetic (Outcome 1) and non-diabetic (Outcome 0) individuals. -Variability: The interquartile range (IQR) is also comparable, indicating similar variability in skin thickness. -Outliers: Both groups have outliers, but they are not numerous. There is little difference in skin thickness between diabetic and non-diabetic individuals, suggesting skin thickness may not be a strong differentiator for diabetes status.
Box plot of insulin by outcom		- Median: Diabetic individuals (Outcome 1) have a significantly higher median insulin level compared to non-diabetic individuals (Outcome 0). -Variability: The interquartile range (IQR) is larger for diabetics, indicating greater variability in insulin levels within this group. -Outliers: There are numerous outliers in both groups, particularly in the diabetic group, suggesting some individuals have exceptionally high insulin levels. The box plot reveals that insulin levels are markedly higher in diabetic individuals, indicating that insulin levels may be an important factor in distinguishing between diabetic and non-diabetic statuses.

Box plot of DiabetesPedigreeFunction by outcom		-Median: The median Diabetes Pedigree Function is higher for diabetic individuals (Outcome 1) compared to non-diabetic individuals (Outcome 0), suggesting a potential link between the pedigree function and diabetes risk. -Variability: The interquartile range (IQR) indicates that the diabetic group has a wider range of values, suggesting greater variability in this measure among diabetics. -Outliers: There are several outliers present in both groups, particularly in the diabetic group, which may indicate some individuals with significantly higher pedigree function scores.
Box plot of Age by outcom		-Median: The median age for non-diabetic individuals (Outcome 0) is significantly higher than for diabetic individuals (Outcome 1). This suggests that non-diabetic individuals tend to be older on average. -Variability: The interquartile range (IQR) for both groups shows that there is less variability in age among non-diabetic individuals compared to diabetic individuals, who exhibit a wider range of ages. -Outliers: The box plot indicates that both groups have outliers, particularly in the non-diabetic group, indicating some individuals are much older than the typical range. The greater variability in the diabetic group highlights a diverse age range among individuals with diabetes.
Box plot of BMI by outcom		-Median: The median BMI is similar for both diabetic (Outcome 1) and non-diabetic (Outcome 0) individuals, indicating comparable central values. -Variability: The interquartile range (IQR) shows some overlap between the two groups, suggesting similar variability in BMI. -Outliers: Both groups have several outliers, particularly in the diabetic group, which may indicate a few individuals with exceptionally high BMI. The box plot suggests that BMI does not significantly differentiate between diabetic and non-diabetic individuals, as the central tendency and variability are quite similar in both groups.

4.Data preprocessing:

As we all know, preprocessing is a crucial step in data analysis and machine learning because raw datasets often contain imperfections and inconsistencies that can affect model performance, and our chosen dataset wasn't perfect, so we decide to apply preprocessing on it.

First, we searched for any duplicate rows to eliminate it, but we didn't find any duplicated rows. Then we start to fill out missing values, but we find no null values in our data set, but we believe that the zeros in some columns such as SkinThickness, Insulin and BMI are actually missing values and have been filled with zeros; due to the fact that these values can not be 0 for any human. So we decided to treat these zeros in the mentioned columns as missing values and then replace it with the mean of the column.

Dataset before and after filling nulls

	A	B	C	D	E	F	G	H	I	J
1	Id	Pregnanci	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
2	1	6	148	72	35	0	33.6	0.627	50	1
3	2	1	85	66	29	0	26.6	0.351	31	0
4	3	8	183	64	0	0	23.3	0.672	32	1
5	4	1	89	66	23	94	28.1	0.167	21	0
6	5	0	137	40	35	168	43.1	2.288	33	1
7	6	5	116	74	0	0	25.6	0.201	30	0
8	7	3	78	50	32	88	31	0.248	26	1
9	8	10	115	0	0	0	35.3	0.134	29	0
10	9	2	197	70	45	543	30.5	0.158	53	1
11	10	8	125	96	0	0	0	0.232	54	1
12	11	4	110	92	0	0	37.6	0.191	30	0
13	12	10	168	74	0	0	38	0.537	34	1
14	13	10	139	80	0	0	27.1	1.441	57	0
15	14	1	189	60	23	846	30.1	0.398	59	1
16	15	5	166	72	19	175	25.8	0.587	51	1
17	16	7	100	0	0	0	30	0.484	32	1
18	17	0	118	84	47	230	45.8	0.551	31	1
19	18	7	107	74	0	0	29.6	0.254	31	1
20	19	1	103	30	38	83	43.3	0.183	33	0
21	20	1	115	70	30	96	34.6	0.529	32	1
22	21	3	126	88	41	235	39.3	0.704	27	0
23	22	8	99	84	0	0	35.4	0.388	50	0
24	23	7	196	90	0	0	39.8	0.451	41	1
25	24	9	119	80	35	0	29	0.263	29	1

	A	B	C	D	E	F	G	H	I	J
1	Id	Pregnanci	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
2	1	6	0.679739	0.5	0.740741	0.623984	0.490446	0.493261	50	1
3	2	1	0.267974	0.411765	0.518519	0.623984	0.267516	0.245283	31	0
4	3	8	0.908497	0.382353	0.529246	0.623984	0.16242	0.533693	32	1
5	4	1	0.294118	0.411765	0.296296	0.177778	0.315287	0.079964	21	0
6	5	5	0.470588	0.529412	0.529246	0.623984	0.235669	0.110512	30	0
7	6	3	0.222222	0.176471	0.62963	0.133333	0.407643	0.15274	26	1
8	10	8	0.529412	0.852941	0.529246	0.623984	0.458493	0.138365	54	1
9	11	4	0.431373	0.794118	0.529246	0.623984	0.617634	0.101527	30	0
10	12	10	0.810458	0.529412	0.529246	0.623984	0.630573	0.412399	34	1
11	15	5	0.797386	0.5	0.148148	0.777778	0.242038	0.457323	51	1
12	18	7	0.411765	0.529412	0.529246	0.623984	0.363057	0.158131	31	1
13	20	1	0.464052	0.470588	0.555556	0.192593	0.522293	0.405211	32	1
14	22	8	0.359477	0.676471	0.529246	0.623984	0.547771	0.278527	50	0
15	23	7	0.993464	0.764706	0.529246	0.623984	0.687898	0.33513	41	1
16	24	9	0.490196	0.617647	0.740741	0.623984	0.343949	0.166217	29	1
17	25	11	0.647059	0.823529	0.666667	0.562963	0.585987	0.158131	51	1
18	26	10	0.529412	0.470588	0.407407	0.333333	0.410828	0.114106	41	1
19	27	7	0.673203	0.556824	0.529246	0.623984	0.675159	0.160827	43	1
20	28	1	0.346405	0.411765	0	0.518519	0.159236	0.367475	22	0
21	29	13	0.660131	0.647059	0.148148	0.296296	0.127389	0.150045	57	0
22	30	5	0.477124	0.794118	0.529246	0.623984	0.506369	0.232704	38	0
23	31	5	0.424837	0.544118	0.407407	0.623984	0.566879	0.420485	60	0
24	34	6	0.313725	0.794118	0.529246	0.623984	0.05414	0.098832	28	0
25	35	10	0.509804	0.588235	0.592593	0.623984	0.299363	0.389937	45	0

As third step in data cleaning we then begin to detect and remove outliers and because our dataset is not normally distributed we found it more appropriate to use IQR (interquartile range) method which identifies outliers by calculating the range between the first quartile (Q1) and the third quartile (Q3), then we removed detected outliers.

Dataset before and after removing outliers

```
        Id  Pregnancies  Glucose  BloodPressure  SkinThickness    Insulin
0        1            6      148             72      35.000000  154.23783
1        2            1       85             66      29.000000  154.23783
2        3            8      183             64      29.289634  154.23783
3        4            1       89             66      23.000000   94.00000
4        5            0      137             40      35.000000  168.00000
...    ...          ...      ...            ...            ...        ...
2763  2764            2       75             64      24.000000   55.00000
2764  2765            8      179             72      42.000000  130.00000
2765  2766            6       85             78      29.289634  154.23783
2766  2767            0      129            110      46.000000  130.00000
2767  2768            2       81             72      15.000000   76.00000

       BMI  DiabetesPedigreeFunction  Age  Outcome
0     33.6                    0.627   50        1
1     26.6                    0.351   31        0
2     23.3                    0.672   32        1
3     28.1                    0.167   21        0
4     43.1                    2.288   33        1
...    ...                      ...  ...      ...
2763  29.7                    0.370   33        0
2764  32.7                    0.719   36        1
2765  31.2                    0.382   42        0
2766  67.1                    0.319   26        1
2767  30.1                    0.547   25        0
```

```
Number of outliers: 1062
Cleaned Dataset:
        Id  Pregnancies  Glucose  BloodPressure  SkinThickness    Insulin
0        1            6      148             72      35.000000  154.23783
1        2            1       85             66      29.000000  154.23783
2        3            8      183             64      29.289634  154.23783
3        4            1       89             66      23.000000   94.00000
5        6            5      116             74      29.289634  154.23783
...    ...          ...      ...            ...            ...        ...
2759  2760            6      102             82      29.289634  154.23783
2760  2761            6      134             70      23.000000  130.00000
2764  2765            8      179             72      42.000000  130.00000
2765  2766            6       85             78      29.289634  154.23783
2767  2768            2       81             72      15.000000   76.00000

       BMI  DiabetesPedigreeFunction  Age  Outcome
0     33.6                    0.627   50        1
1     26.6                    0.351   31        0
2     23.3                    0.672   32        1
3     28.1                    0.167   21        0
5     25.6                    0.201   30        0
...    ...                      ...  ...      ...
2759  30.8                    0.180   36        1
2760  35.4                    0.542   29        1
2764  32.7                    0.719   36        1
2765  31.2                    0.382   42        0
2767  30.1                    0.547   25        0
```

We also applied data transformation to convert our data into a suitable format or structure for analysis and model training. Normalization was performed to ensure consistent data scale. The normalization technique applied is max-min normalization. This technique scales specific attribute values to a specified range from 0 to 1. The following attributes were selected for normalization: {'BMI','Glucose','BloodPressure','SkinThickness','Insulin','DiabetesPedigreeFunction'}.

Dataset before and after applying normalization

	Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
0	1	6	148	72	35.000000	154.23783
1	2	1	85	66	29.000000	154.23783
2	3	8	183	64	29.289634	154.23783
3	4	1	89	66	23.000000	94.00000
4	5	0	137	40	35.000000	168.00000
...	...	...	...	...	...	...
2763	2764	2	75	64	24.000000	55.00000
2764	2765	8	179	72	42.000000	130.00000
2765	2766	6	85	78	29.289634	154.23783
2766	2767	0	129	110	46.000000	130.00000
2767	2768	2	81	72	15.000000	76.00000

	BMI	DiabetesPedigreeFunction	Age	Outcome
0	33.6	0.627	50	1
1	26.6	0.351	31	0
2	23.3	0.672	32	1
3	28.1	0.167	21	0
4	43.1	2.288	33	1
...	...	...	...	...
2763	29.7	0.370	33	0
2764	32.7	0.719	36	1
2765	31.2	0.382	42	0
2766	67.1	0.319	26	1
2767	30.1	0.547	25	0

min-max scaled data:

	Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
0	1	6	0.679739	0.500000	0.740741	0.623984
1	2	1	0.267974	0.411765	0.518519	0.623984
2	3	8	0.908497	0.382353	0.529246	0.623984
3	4	1	0.294118	0.411765	0.296296	0.177778
5	6	5	0.470588	0.529412	0.529246	0.623984
...	...	...	...	...	...	...
2759	2760	6	0.379085	0.647059	0.529246	0.623984
2760	2761	6	0.588235	0.470588	0.296296	0.444444
2764	2765	8	0.882353	0.500000	1.000000	0.444444
2765	2766	6	0.267974	0.588235	0.529246	0.623984
2767	2768	2	0.241830	0.500000	0.000000	0.044444

	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.490446	0.493261	50	1
1	0.267516	0.245283	31	0
2	0.162420	0.533693	32	1
3	0.315287	0.079964	21	0
5	0.235669	0.110512	30	0
...	...	...	...	...
2759	0.401274	0.091644	36	1
2760	0.547771	0.416891	29	1
2764	0.461783	0.575921	36	1
2765	0.414013	0.273136	42	0
2767	0.378981	0.421384	25	0

Then we applied discretization to the "age" column to simplify the continuous age data by grouping it into meaningful age ranges, the ranges are [0-30 , 31-50 , 51-100] This helps reduce noise, improve interpretability, and may enhance model performance by capturing patterns more effectively in categories rather than treating age as a continuous variable.

Dataset before and after applying discretization

original DataFrame:

	Age	AgeGroup
0	50	31-50
1	31	31-50
2	32	31-50
3	21	0-30
4	33	31-50
...	...	...
2763	33	31-50
2764	36	31-50
2765	42	31-50
2766	26	0-30
2767	25	0-30

And for Encoding: All the data is already in a numerical format, so there's no need to perform any encoding. And with this we were done with data preprocessing, and our cleaned dataset now is ready to apply mining techniques.

5.Data Mining Techniques:

We applied both supervised and unsupervised learning techniques to our dataset, employing classification and clustering methods. For our classification task, we chose to use a decision tree algorithm. This recursive method constructs a tree-like structure in which each leaf node represents a conclusive decision. The objective of our model is to predict whether an individual has diabetes, categorizing the outcomes as ‘1’ for those who are diabetic and ‘0’ for non- diabetic individuals. The model’s predictions are based on several features, including age, the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, and diabetes pedigree function.

As discussed, classification is a form of supervised learning, which necessitates a training dataset to develop the model. To facilitate this, we divided our dataset into two parts: a training set and a testing set. We experimented with three different training set sizes: 70%, 80%, and 90%. Additionally, we applied two methods for attribute selection: information gain (using entropy) and the Gini index.

To evaluate the effectiveness of our model and identify the best partitioning strategy, we assessed its accuracy and utilized a confusion matrix. This matrix provides a comprehensive overview of key performance metrics, including sensitivity, specificity, precision, and error rate.

Some Python packages and methods we use: Python Packages:

1. Pandas

• Purpose: Data manipulation and analysis.

• Common Methods:

- pd.read_csv(): Load data from a CSV file.
- df.drop(): Remove columns or rows from a DataFrame.

2. Scikit-Learn (sklearn)

• Purpose: A comprehensive library for machine learning.

• Common Methods:

- train_test_split(): Split the dataset into training and testing subsets.

3. Matplotlib

• Purpose: Data visualization.

• Common Methods:

- plt.plot(): Create a line plot.
- plt.xlabel(), plt.ylabel(), plt.title(): Label axes and add titles to plots. o plt.show(): Display the plot.

4. Seaborn (optional for enhanced visualization)

• Purpose: Statistical data visualization based on Matplotlib.

• Common Methods:

- *sns.scatterplot(): Create scatter plots with additional styling options.*
- *sns.heatmap(): Visualize data in a matrix format.*

In the clustering process, which is a type of unsupervised learning, we excluded the "Outcome" attribute as it serves as a class label and is not used in this analysis. Instead, we utilized all other attributes, including:

Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age,

all of which are numeric and required no conversion prior to clustering.

Clustering Steps:

We employed the K-means algorithm, which generates k clusters, each represented by a centroid.

The algorithm assigns each object to the nearest cluster and then iteratively recalculates the centroids and reassigns objects until the centroids stabilize, indicating that the clusters are correctly assigned.

Cluster Validation:

We calculated the average Silhouette Score for each cluster using the Average Silhouette Score method and visualized the results.

Additionally, we applied the Elbow method to compare three different cluster sizes to determine the optimal number by assessing the separation and compactness of the clusters.

In addition of Python packages and methods that we have mentioned before we also use:

1. NumPy

• Purpose: provides support for numerical operations and random number generation to ensure reproducibility in experiments.

• Common Methods:

- np.random.seed(#): Sets the random seed to ensure that the results of any random process (e.g., random initialization in K-means clustering) are reproducible.

2. Yellowbrick

• Purpose: A visualization library for machine learning that extends scikit-learn.

• Common Methods:

- SilhouetteVisualizer: Visualizes silhouette scores for clustering.
- elbow_method(): Helps to determine the optimal number of clusters.

3.Pipeline (from sklearn.pipeline)

• Purpose: Helps chain preprocessing steps and models into a single workflow.

• Common Methods:

- make_pipeline(): Quickly build a pipeline with pre-defined steps.

---

6. Evaluation and Comparison:

- Classification

Information Gain:

70% training – 30% testing

Figure (1) confusion matrix

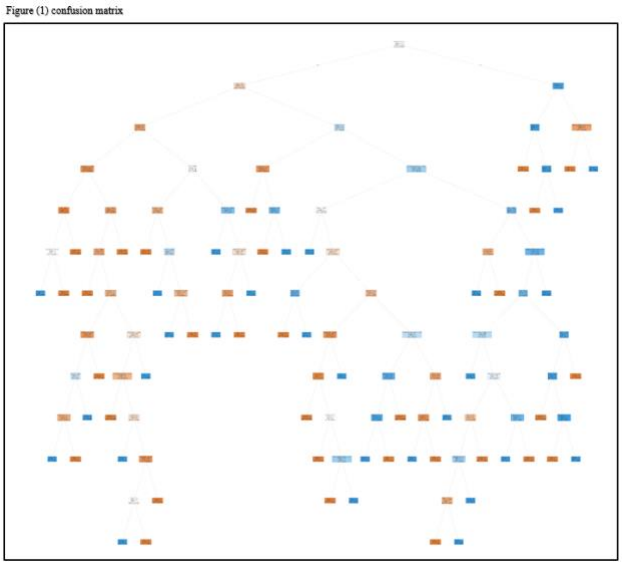
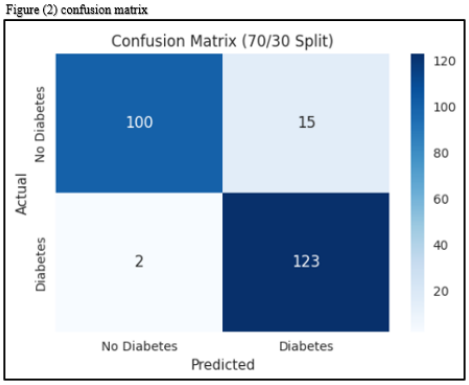


Figure (2) confusion matrix



80% training – 20% testing

Figure (1) confusion matrix

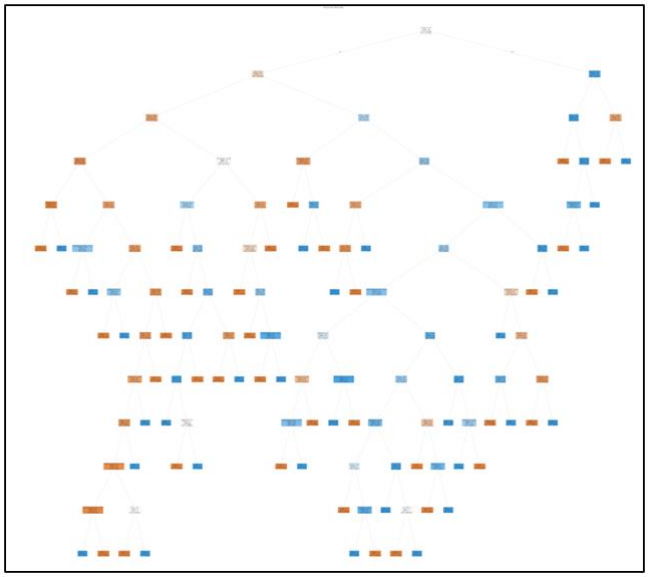
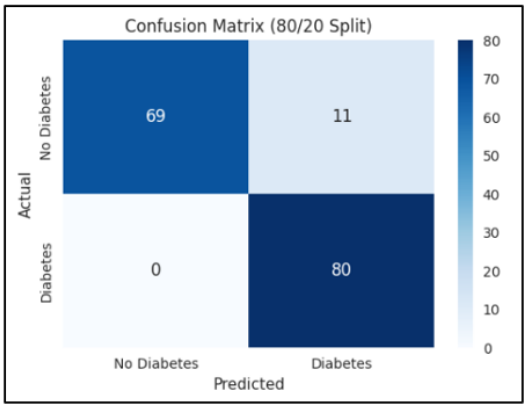


Figure (2) confusion matrix



90% training – 10% testing

Figure (1) confusion matrix

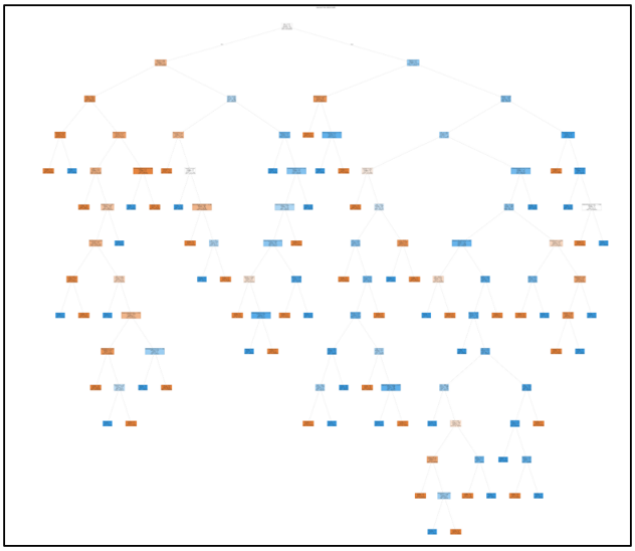
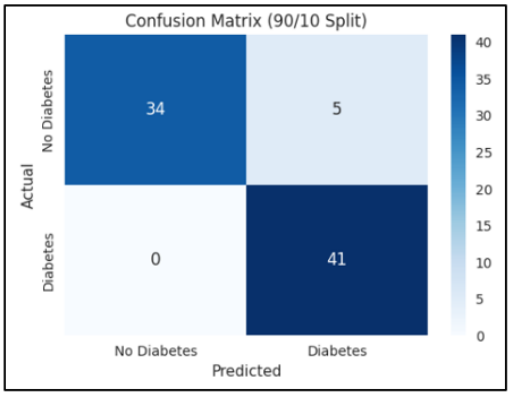


Figure (2) confusion matrix



Mining Task	Comparison Criteria
Classification for IG	We tried 3 different sizes for dataset splitting to create the decision tree

	70% training – 30% testing
Accuracy	93%
Error Rate	7%
Sensitivity	98%
Specificity	87%
Precision	89%

	80% training – 20% testing
Accuracy	93%
Error Rate	7%
Sensitivity	100%
Specificity	86%
Precision	88%

	90% training – 10% testing
Accuracy	94%
Error Rate	6%
Sensitivity	100%
Specificity	87%
Precision	89%

The 90% training, 10% testing split in gini index performs the best overall, showing the highest accuracy, lowest error rate, perfect sensitivity, and improved specificity and precision. This suggests that with more training data, the model achieves better classification performance.

Gini index:
70% training – 30% testing

Figure (1) confusion matrix

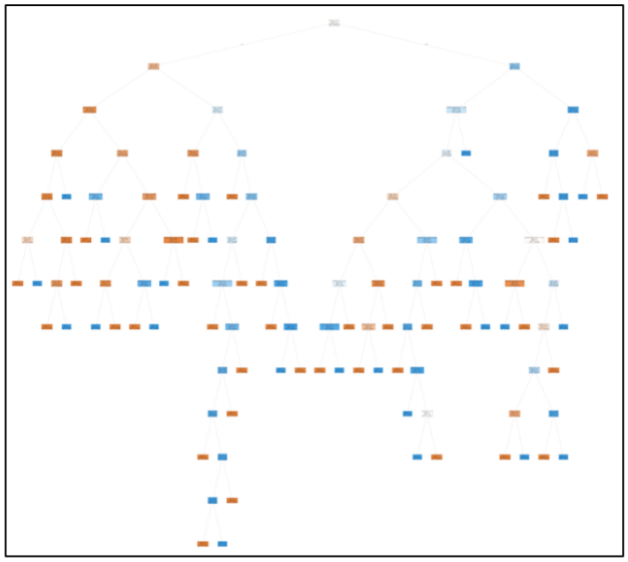
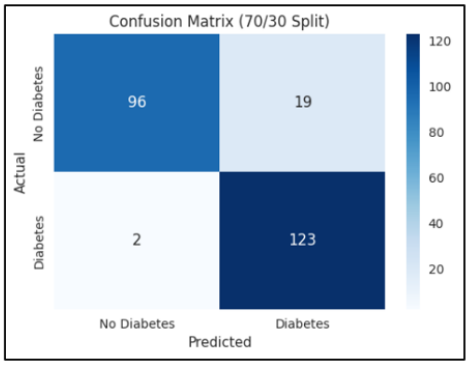


Figure (2) confusion matrix



80% training – 20% testing

Figure (1) confusion matrix

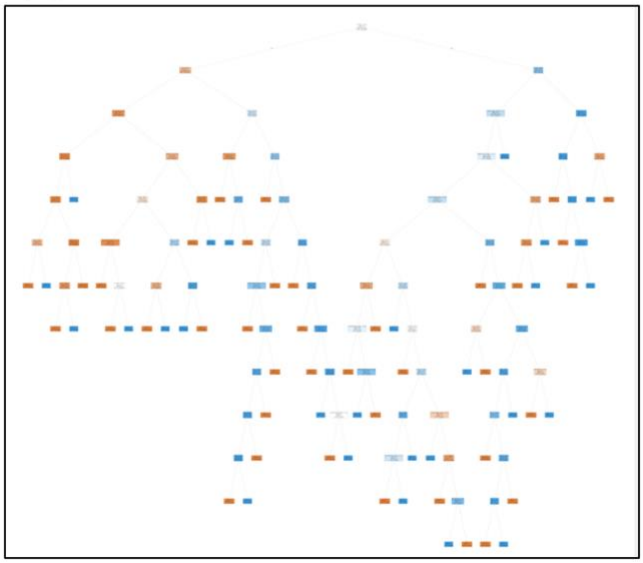
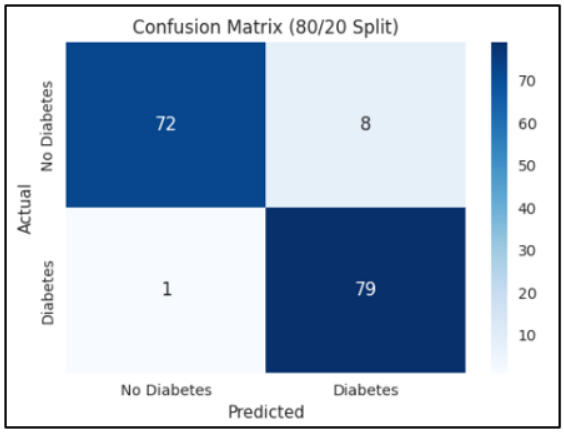


Figure (2) confusion matrix



90% training – 10% testing

Figure (1) confusion matrix

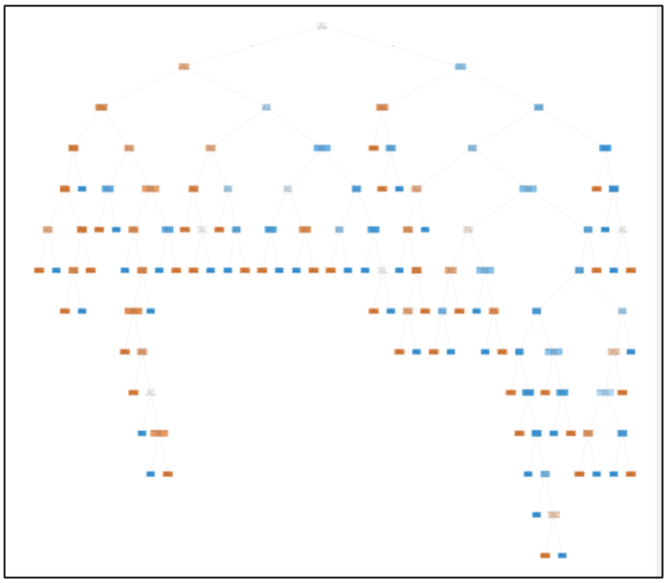
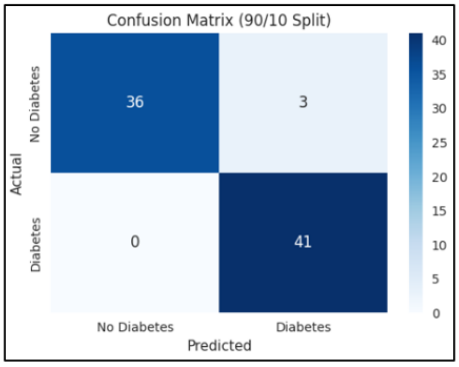


Figure (2) confusion matrix



Mining Task	Comparison Criteria
Classification for Gini index	We tried 3 different sizes for dataset splitting to create the decision tree

	70% training – 30% testing
Accuracy	91%
Error Rate	9%
Sensitivity	98%
Specificity	83%
Precision	87%

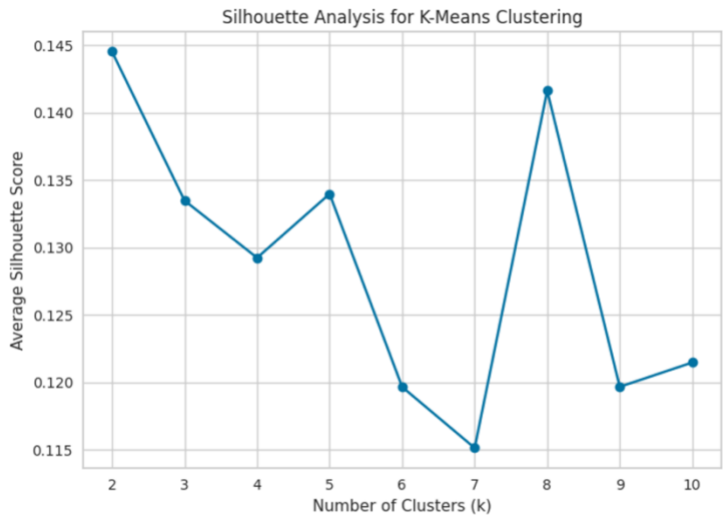
	80% training – 20% testing
Accuracy	94%
Error Rate	6%
Sensitivity	99%
Specificity	90%
Precision	91%

	90% training – 10% testing
Accuracy	96%
Error Rate	4%
Sensitivity	100%
Specificity	92%
Precision	93%

- 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:
Silhouette method is a technique used to measure the clustering quality and determine the optimal number of clusters.



Elbow method:

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

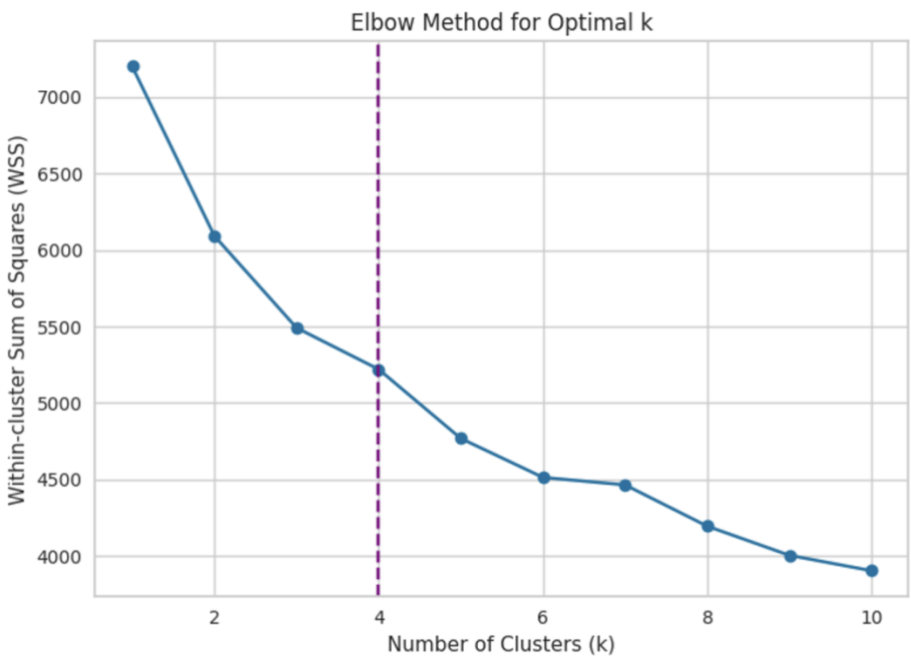


Figure (1) Silhouette scores [K=2]:

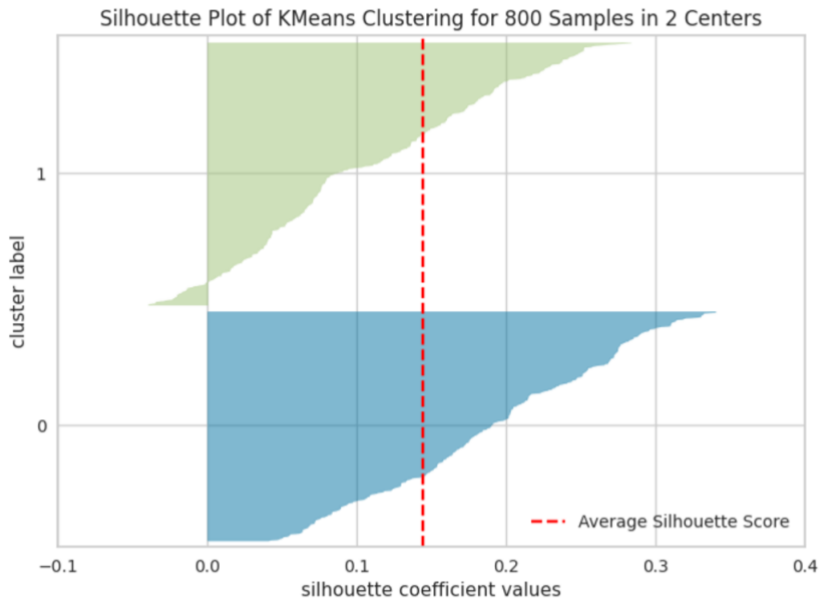


Figure (2) Silhouette scores [K=3]:

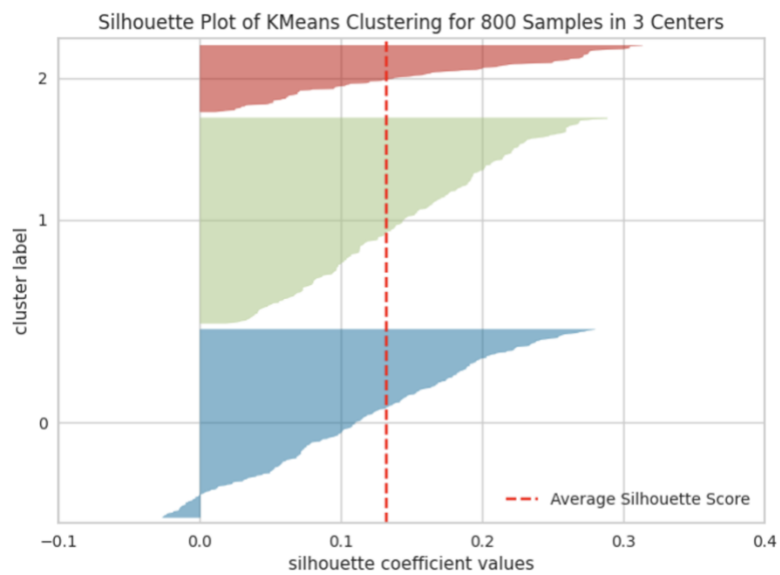
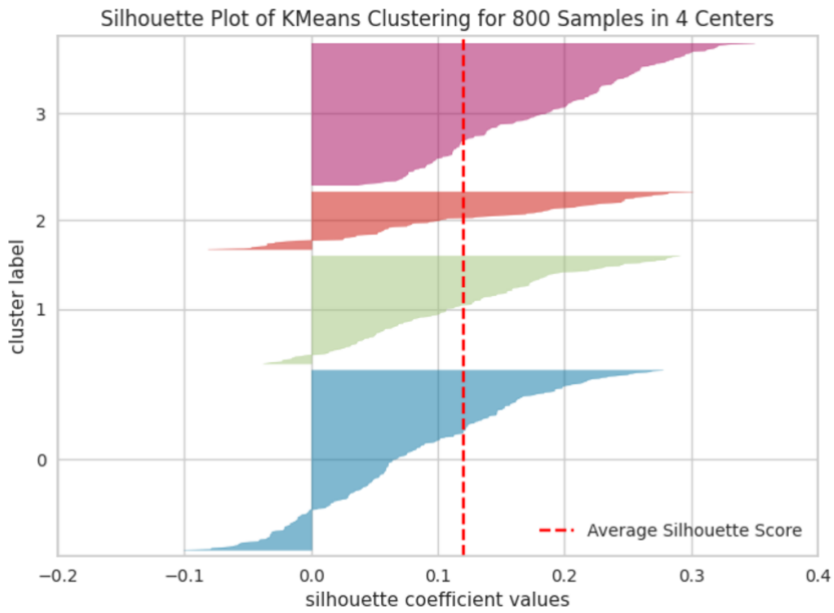


Figure (3) Silhouette scores [K=4]:



<table>
<tr><th>Mining task</th><th>Comparison Criteria</th></tr>
<tr><td>Clustering</td><td>
<table>
<tr><th></th><th>K= 2 (BEST)</th><th>K= 3</th><th>K= 4</th></tr>
<tr><td>Average Silhouette width</td><td>0.14453416465113475</td><td>0.132920274761205</td><td>0.1337386997358189</td></tr>
<tr><td>total within-cluster sum of square</td><td>6091.955083741861</td><td>5620.641135564803</td><td>5166.817672974286</td></tr>
</table>
</td></tr>
</table>

7.*Findings*:

Initially, we chose a dataset that includes patients diagnosed with diabetes, aiming to comprehend the underlying causes of this widespread condition and develop effective preventive strategies.
To enhance the effectiveness, accuracy, and precision of our results, we employed several data processing techniques aimed at improving data efficiency. We utilized a range of visualization methods, including box plots, scatter plots, and line graphs, to clarify the data and aid understanding. This facilitated the application of appropriate data processing techniques.
Based on these visualizations and additional analyses, we eliminated all empty, missing, and outlier values that could adversely affect the results. We also implemented data transformations such as normalization and feature partitioning, along with a balanced data process to ensure equal weight for certain features, thus streamlining data processing during mining tasks.
As a result, we performed data mining tasks that included classification and partitioning. For the classification, we used the Gini index and information gain metrics. By experimenting with three different sizes of training and testing datasets, we were able to achieve optimal results in both model construction and evaluation. Here are our findings:

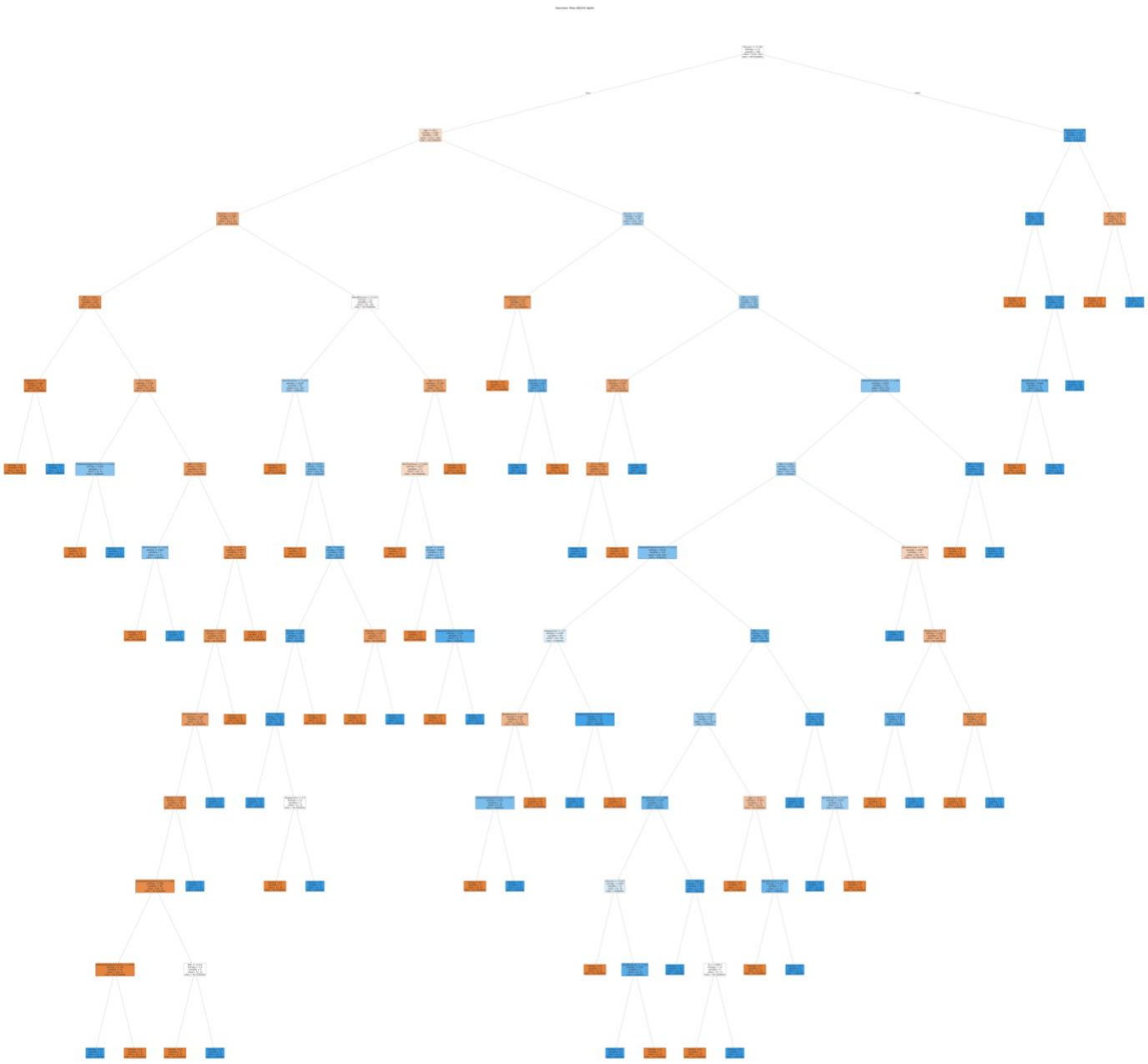
-Information Gain:

	70% training, 30% testing	80% training, 20% testing	90% training, 10% testing
Accuracy	0.93	0.93	0.94
Error Rate	0.07	0.07	0.06
Sensitivity	0.98	1.00	1.00
Specificity	0.87	0.86	0.87
Precision	0.89	0.88	0.89

Based on the presented results for the models trained using the Information Gain criterion, the following observations can be made:
1. Accuracy:
- Measures the proportion of correctly predicted instances out of the total instances.
- Values range from 0.93 to 0.94, indicating the model performs consistently well across all splits.
2. Error Rate:
- Represents the proportion of incorrect predictions.
- Lower values are better; the error rate decreases slightly from 0.07 to 0.06 as the training set size increases.
3. Sensitivity (Recall):
- Indicates the model's ability to correctly identify positive instances.

- A perfect score of 1.00 in the 80/20 and 90/10 splits suggests the model is excellent at detecting positive cases.
4. Specificity:
- Measures the model's ability to correctly identify negative instances.
- Values are around 0.86 to 0.87, suggesting moderate performance in identifying negatives.
5. Precision:
- Reflects the accuracy of positive predictions.
- Values are high (0.88 to 0.89), indicating that when the model predicts a positive case, it is likely to be correct.
The partition with the highest entropy is: 80 train - 20 test with entropy of 1.0000
The difference between the highest accuracy and the highest entropy results indicates that the two metrics are measuring different properties of the partitions.A higher accuracy means the model is more correct overall. A higher entropy means the dataset is more evenly distributed (less skewed towards one outcome or class), which may not always correlate with better performance. so: The choice depends on your objective:
If the goal is higher performance (e.g., accuracy): Choose the 90-10 partition since it maximizes accuracy. If the goal is better generalization and robustness: Consider the 80-20 partition (highest entropy) as it may better represent the underlying uncertainty or balance in the dataset.

This was the decision tree associated with this division:



The root node of this decision tree is the Glucose feature, with a split criterion of Glucose <= 0.748. This suggests that the glucose level is the most important factor for distinguishing between the "No Diabetes" and "Diabetes" classes in the dataset.

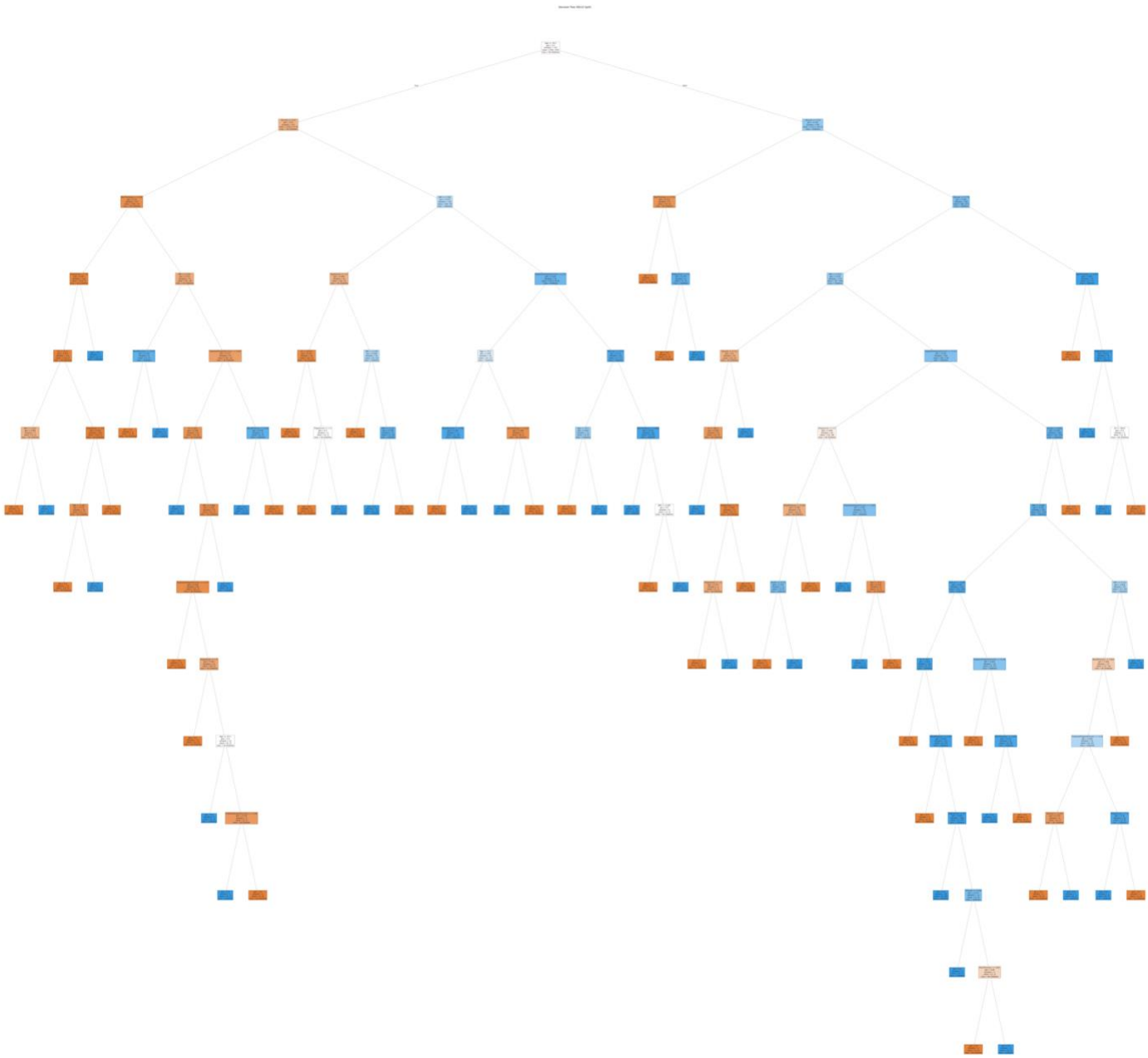
The entropy value at the root node is 1.0, which indicates a relatively high level of impurity or uncertainty in the data. As the tree grows deeper, additional splits are made based on other features to reduce this entropy and improve the classification accuracy.

Overall, this decision tree appears to be using the glucose level as the primary factor to make the initial split, and then likely incorporating other relevant features as it grows deeper to better distinguish between the two classes of "No Diabetes" and "Diabetes".

-Gini index:

	70% training, 30% testing	80% training, 20% testing	90% training, 10% testing
Accuracy	0.91	0.94	0.96
Error Rate	0.09	0.06	0.04
Sensitivity	0.98	0.99	1.00
Specificity	0.83	0.90	0.92
Precision	0.87	0.91	0.93

1. Accuracy:
- Proportion of correctly predicted instances.
- Increases from 0.91 to 0.96 as the training set size increases, indicating improved model performance with more training data.
2. Error Rate:
- Proportion of incorrect predictions.
- Decreases from 0.09 to 0.04, showing that larger training sets lead to fewer errors.
3.Sensitivity (Recall):
- Ability of the model to correctly identify positive instances.
- Remains high, with scores of 0.98 to 1.00, indicating excellent detection of positive cases.
4.Specificity:
- Ability to correctly identify negative instances.
- Shows improvement from 0.83 to 0.92, suggesting better performance in identifying negatives with more training data.
5.Precision:
- Accuracy of positive predictions.
- Increases from 0.87 to 0.93, indicating that the model becomes more reliable in its positive predictions as the training set grows. Summary:
- The model demonstrates strong performance metrics across all splits, improving with larger training datasets.
- The **90% Training, 10% Testing** partition stands out as the best, with the highest accuracy (0.96), lowest error rate (0.04), perfect sensitivity (1.00), and high precision (0.93).
- Specificity also improves significantly, reaching 0.92 in the best partition.
This was the decision tree associated with this division:



Root Feature: Age ≤ 30.5. Gini Index: 0.5. Class Distribution: [361, 359]. Conclusion: In this larger partition, "Age" is selected as the root node feature, suggesting that as the sample size increases, age becomes more influential in predicting "No Diabetes." The distribution of classes is again nearly balanced, reflecting a well- represented dataset.

Overall Conclusion: Across different partition sizes, the root node feature alternates between "Glucose" and "Age," suggesting that both are critical predictors for the target variable. The Gini index consistently being 0.5 reflects a balanced dataset at the root, and the stability of the class distribution highlights the robustness of the dataset in representing the two classes equally. Larger sample sizes may shift feature importance, as seen with the transition from "Glucose" to "Age."

-Comparison between Information Gain and Gini Index:

	Information gain	Gini index
Accuracy	0.93	0.94
Error Rate	0.07	0.06
Sensitivity	0.92	0.99
Specificity	0.92	0.88
Precision	0.89	0.90

- Accuracy: The Gini Index performs slightly better with an accuracy of 0.94 compared to Information Gain's 0.93.
- Error Rate: Gini Index has a lower error rate (0.06) compared to Information Gain (0.07),

  indicating it makes fewer incorrect predictions.
- Sensitivity: The Gini Index significantly outperforms Information Gain in sensitivity (0.99 vs.

  0.92), which means it is much better at detecting positive cases.
- Specificity: Information Gain has a higher specificity (0.92) compared to the Gini Index (0.88),

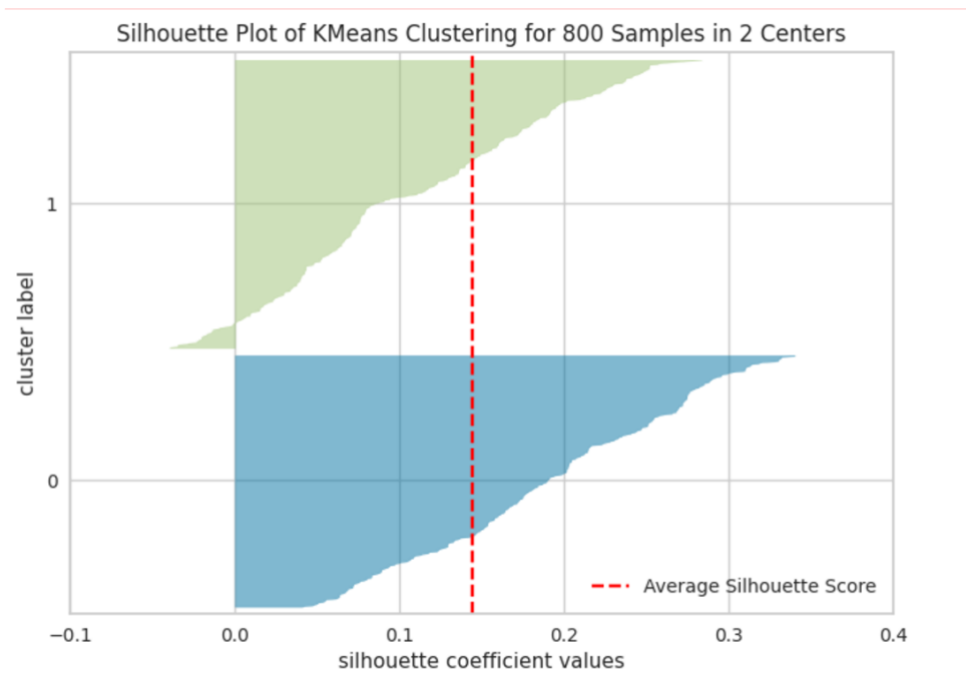
  meaning it is slightly better at identifying negative cases.
- Precision: The Gini Index is slightly better than Information Gain (0.90 vs. 0.89), meaning it has a marginally higher ability to correctly predict positive cases.

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 (BEST)	K= 3	K= 4
Average Silhouette width	0.14453416465113475	0.132920274761205	0.1337386997358189
total within-cluster sum of square	6091.955083741861	5620.641135564803	5166.817672974286

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=4 Also, having a silhouette plot of kmeans clustring of 800 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 800 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.