



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

IT 326: Data Mining
2nd Semester 1446 H

Group #6
Wednesday 10-12

Section #:	52846	
Group #:	Group #6 Wednesday 10-12	
Group Members	NAME	ID
	leen aldbays	444200652
	Walah Alsaeed	444201689
	Lamiis Alsaleh	444201135
	Mariah alnfisah	444200965
	Norah Alkathiri	444200439

1. Problem

Alzheimer's Disease remains a growing global concern with no definitive cure. One of the major challenges lies in its early detection, as symptoms often develop gradually and can be confused with normal aging. This delay in diagnosis impacts on the quality of care and emotional well-being of both patients and their families. Given the complex nature of the disease affected by lifestyle, genetics, and clinical factors, there is a crucial need to identify the key predictors of Alzheimer's and develop models that can accurately assess the risk. Solving this problem can help facilitate timely intervention, improve treatment planning, and offer families the opportunity to better prepare and support their loved ones.

2. Data Mining Task

In this project, we applied two data mining tasks: classification and clustering to support the prediction of Alzheimer's Disease risk.

For classification, we train a model to predict whether an individual is at risk of developing Alzheimer's based on a range of medical and lifestyle attributes such as age, family history, education level, blood pressure, diabetes, and more. The classification is based on the "Diagnosis" class, which indicates whether a person is diagnosed with Alzheimer's or not (binary: diagnosed or not diagnosed).

As for clustering, the model groups individuals with similar characteristics into clusters without using the diagnosis label(Class label). These clusters help uncover common patterns and traits among individuals, offering deeper insights into possible risk factors. This approach may also reveal hidden relationships between features, supporting early detection efforts and tailored prevention strategies.

3. Data

- The Source: <https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset>
- Number of attributes: 34
- Number of Objects: 2149
- Class label: Diagnosis

- Attributes' description

	Attribute Name	Description	Data Type	Possible Values
0	PatientID	Unique identifier for each patient	Numeric	Range: 4701 - 6899
1	Age	Age of the patient in years	Numeric	Range: 60 - 90
2	Gender	Gender of the patient (0 = Female, 1 = Male)	Binary	0, 1
3	Ethnicity	Encoded ethnic background of the patient	Numeric	Range: 0 - 3
4	EducationLevel	Level of education encoded numerically	Numeric	Range: 0 - 3
5	BMI	Body Mass Index of the patient	Numeric	Range: 15.00885118 - 39.59276746
6	Smoking	Whether the patient smokes (0 = No, 1 = Yes)	Binary	0, 1
7	AlcoholConsumption	Level of alcohol consumed by the patient	Numeric	Range: 0.002003099 - 15.09329336
8	PhysicalActivity	Physical activity level or frequency	Numeric	Range: 0.003618017 - 9.987429413
9	DietQuality	Score representing the quality of diet	Numeric	Range: 0.00938472 - 9.998349679
10	SleepQuality	Score representing sleep quality	Numeric	Range: 4.00262866 - 9.99940317
11	FamilyHistoryAlzheimers	Family history of Alzheimer's (0 = No, 1 = Yes)	Binary	0, 1
12	CardiovascularDisease	Presence of cardiovascular disease (0 = No, 1 = ...)	Binary	0, 1
13	Diabetes	Whether the patient has diabetes (0 = No, 1 = ...)	Binary	0, 1
14	Depression	Presence of diagnosed depression (0 = No, 1 = ...)	Binary	0, 1
15	HeadInjury	History of head injury (0 = No, 1 = Yes)	Binary	0, 1
16	Hypertension	Presence of high blood pressure (0 = No, 1 = Yes)	Binary	0, 1
17	SystolicBP	Systolic blood pressure measurement	Numeric	Range: 90 - 179
18	DiastolicBP	Diastolic blood pressure measurement	Numeric	Range: 60 - 119
19	CholesterolTotal	Total cholesterol level	Numeric	Range: 150.0933156 - 299.9933525
20	CholesterolLDL	Low density lipoprotein (LDL) cholesterol level	Numeric	Range: 50.230738656 - 199.9658651
21	CholesterolHDL	High-density lipoprotein (HDL) cholesterol level	Numeric	Range: 20.00343401 - 99.98032408
22	CholesterolTriglycerides	Triglycerides cholesterol level	Numeric	Range: 50.40719362 - 399.9418616
23	MMSE	Mini-Mental State Examination score	Numeric	Range: 0.005312146 - 29.99138056
24	FunctionalAssessment	Assessment of patient's functional abilities	Numeric	Range: 0.000459594 - 9.996467073
25	MemoryComplaints	Presence of self-reported memory complaints	Binary	0, 1
26	BehavioralProblems	Behavioral changes or abnormalities	Binary	0, 1
27	ADL	Activities of Daily Living score	Numeric	Range: 0.001287928 - 9.999747122
28	Confusion	Confusion episodes reported (0 = No, 1 = Yes)	Binary	0, 1
29	Disorientation	Disorientation symptoms (0 = No, 1 = Yes)	Binary	0, 1
30	PersonalityChanges	Notable personality changes (0 = No, 1 = Yes)	Binary	0, 1
31	DifficultyCompletingTasks	Difficulty completing everyday tasks (0 = No, ...)	Binary	0, 1
32	Forgetfulness	Reported forgetfulness (0 = No, 1 = Yes)	Binary	0, 1
33	Diagnosis	Alzheimer's diagnosis class/category	Binary	0, 1
34	DoctorInCharge	Encoded identifier of responsible doctor	Nominal	XXXXConfid

- Missing values

```
Missing values in each column:
PatientID      0
Age            0
Gender         0
Ethnicity      0
EducationLevel 0
BMI            0
Smoking        0
AlcoholConsumption 0
PhysicalActivity 0
DietQuality    0
SleepQuality   0
FamilyHistoryAlzheimers 0
CardiovascularDisease 0
Diabetes       0
Depression     0
HeadInjury     0
Hypertension   0
SystolicBP     0
DiastolicBP    0
CholesterolTotal 0
CholesterolLDL 0
CholesterolHDL 0
CholesterolTriglycerides 0
MMSE          0
FunctionalAssessment 0
MemoryComplaints 0
BehavioralProblems 0
ADL           0
Confusion      0
Disorientation 0
PersonalityChanges 0
DifficultyCompletingTasks 0
Forgetfulness  0
Diagnosis      0
dtype: int64
```

Rows with missing values:

```
0      0
1      0
2      0
3      0
4      0
..
2144   0
2145   0
2146   0
2147   0
2148   0
Length: 2149, dtype: int64
```

We have no missing values. All columns are complete.

-Statistical measures

To understand our dataset more thoroughly, we computed the following statistical measures to give us a better understanding of the data:

Five number summaries for numeric columns						
	Age	BMI	SystolicBP	DiastolicBP	CholesterolTotal	CholesterolLDL
Min.	60	15.008	90	60	150.09	50.23
Q1	67	21.61	112	74	190.25	87.19
Q2	75	27.82	134	91	225.08	123.34
Q3	83	33.86	157	105	262.03	161.73
Max.	90	39.99	179	119	299.99	199.96
	CholesterolHDL	Hypertension	Gender	CardiovascularDisease	Diabetes	MMSE
Min.	20.003	0	0	0	0	0.005
Q1	39.095	0	0	0	0	7.16
Q2	59.768	0	1	0	0	14.44
Q3	78.93	0	1	0	0	22.16
Max.	99.98	1	1	1	1	29.99

-Outliers

Our data was clear of any outliers, which is further proved by the data processing step in point 4, and the Visual data representation (Box Plot) below.

-Value Variance

Variance measures how spread-out values are. A higher variance means more dispersion, while a lower variance means values are closer to the mean. Therefore, our variance results indicate:

```
Age                80.824080
BMI                52.091413
SystolicBP        673.368875
DiastolicBP       309.495923
CholesterolTotal  1809.841576
CholesterolLDL    1880.660612
CholesterolHDL    535.421368
Hypertension       0.126792
Gender             0.250077
CardiovascularDisease 0.123502
Diabetes           0.128096
MMSE              74.186375
dtype: float64
```

SystolicBP, CholesterolTotal, and CholesterolLDL:

These columns have a high variance with values (673.37, 1809.84, 1880.66) respectively. This indicates a significant level of dispersion, reflecting a widespread in blood pressure and cholesterol levels.

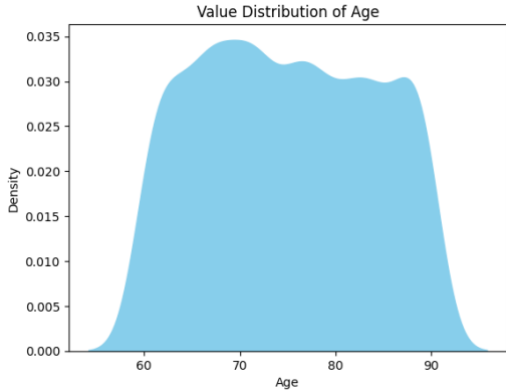
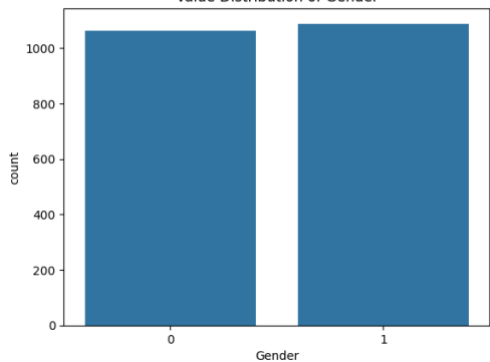
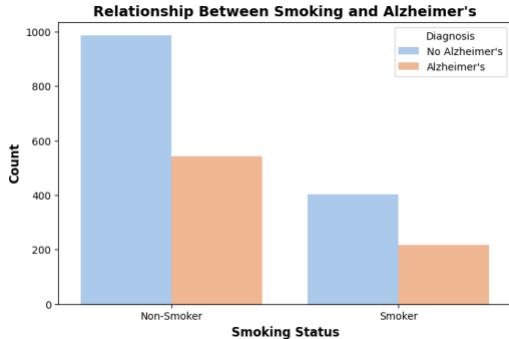
Age, BMI, DiastolicBP, CholesterolHDL, Gender, and MMSE:

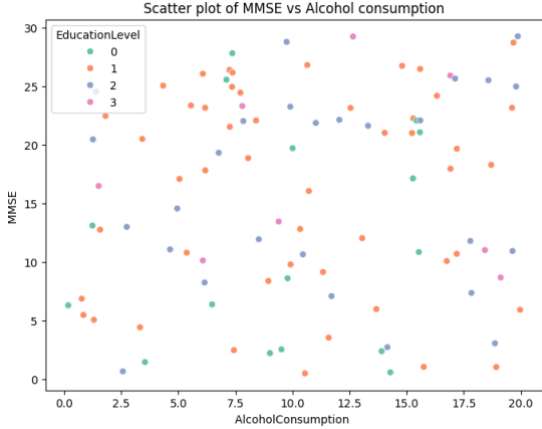
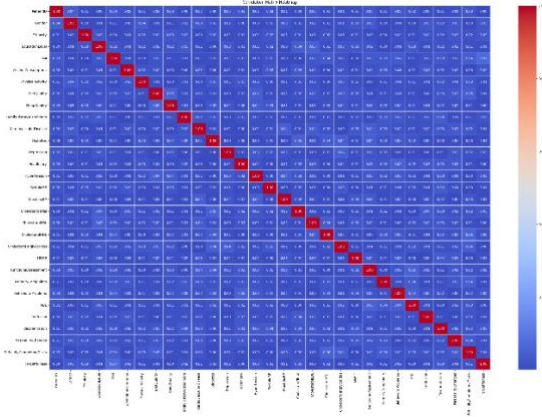
These columns exhibit a moderate level of variance with values (80.82, 52.09, 309.50, 535.42, 0.25, and 0.08) respectively. This suggests a noticeable spread in age, BMI, and blood pressure, but not as extreme as the previous group. The variance in Gender (0.25) indicates a nearly equal distribution of males and females, ensuring a fair representation in gender-based analysis

Hypertension, CardiovascularDisease, Diabetes, and Diagnosis:

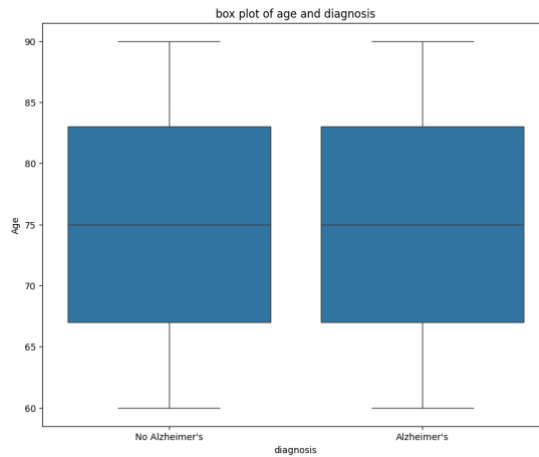
These columns have low variance with values (0.13, 0.12, 0.13, and 0.23) respectively. This indicates that most individuals have similar values in these categories, reflecting limited variability in the dataset.

-Visual Data Representation

Plot	Image	Description
Density plot (numeric data)	 <p>This density plot illustrates how ages of individuals are distributed, as we can see the age value that takes up most of the graph is between 60-70 which is the most seen in the dataset</p>	
Bar Chart (binary data)	 <p>The bar chart shows the distribution of gender in dataset, 0 = female, 1 = male.</p> <p>both genders almost appeared equally in the dataset except that there is slightly higher number of males.</p>	
Clustered bar chart (Nominal data)	 <p>We used a bar chart to show the relationship between smoking and Alzheimer's. The results indicate that non-smokers are more common in both groups, but there are still many smokers among Alzheimer's patients.</p>	

<p>Scatter plot (numeric data)</p>	 <p>A scatter plot titled "Scatter plot of MMSE vs Alcohol consumption". The x-axis is labeled "AlcoholConsumption" and ranges from 0.0 to 20.0 with increments of 2.5. The y-axis is labeled "MMSE" and ranges from 0 to 30 with increments of 5. The plot contains numerous data points colored by "EducationLevel", with a legend showing levels 0 (green), 1 (orange), 2 (blue), and 3 (pink). The points are widely scattered across the plot area, showing no clear linear trend or significant clustering based on education level.</p>	<p>The scatter plot displays the relationship between MMSE scores and alcohol consumption across different education levels.</p> <p>The points are dispersed randomly without forming a clear pattern or trend, suggesting that there is no significant or direct correlation between alcohol consumption and cognitive performance.</p> <p>Additionally, the varying colors representing education levels do not show noticeable clustering, indicating that education level does not have a straightforward impact on this relationship either. So in summary these 3 columns are not strongly correlated to each other which will be further proved in the preprocessing step.</p>
<p>Heatmap (numeric data)</p>	 <p>A heatmap titled "Columns Not Correlated" showing the correlation matrix for various features. The color scale on the right ranges from 0.0 (blue) to 1.0 (red). The diagonal elements are all 1.0 (red). Most off-diagonal elements are blue, indicating low correlation. A few elements are slightly darker blue or light red, but none reach the 0.75 threshold mentioned in the text.</p>	<p>To further detect any correlation between the columns of our dataset we generated a heatmap and it doesn't show any highly correlated columns above the threshold of 0.75, which is proven by the scatter plot above. (A clearer image can be seen in page 12).</p>

Box plot



The boxplot shows the distribution of age for individuals with and without Alzheimer's, indicating that the majority of cases are within the age range of 70 to 80 years, after examining this box plot we can also determine other values such as the minimum and maximum value of each box plot along with any outliers. as we can see both box plots have a maximum value of 90 and a minimum value of 60 and they're both clear of any outliers.

4. Data preprocessing:

-Detecting the outliers:

1-Using Z-score method:

```
The outliers for the dataset: Empty DataFrame
Columns: [PatientID, Age, Gender, Ethnicity, EducationLevel, BMI, Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality, FamilyHistoryAlzheimers, CardiovascularDisease, Diabetes, Depression, HeadInjury, Hypertension, SystolicBP, DiastolicBP, CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides, MMSE, FunctionalAssessment, MemoryComplaints, BehavioralProblems, ADL, Confusion, Disorientation, PersonalityChanges, DifficultyCompletingTasks, Forgetfulness, Diagnosis]
Index: []

[0 rows x 34 columns]

Results Description:

No detected outliers
```

2-using IQR method:

```
outliers using IQR: 0
Empty DataFrame
Columns: [PatientID, Age, Gender, Ethnicity, EducationLevel, BMI, Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality, FamilyHistoryAlzheimers, CardiovascularDisease, Diabetes, Depression, HeadInjury, Hypertension, SystolicBP, DiastolicBP, CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides, MMSE, FunctionalAssessment, MemoryComplaints, BehavioralProblems, ADL, Confusion, Disorientation, PersonalityChanges, DifficultyCompletingTasks, Forgetfulness, Diagnosis]
Index: []

[0 rows x 34 columns]

Results Description:

No detected outliers
```

How to calculate outliers:

Z-score Method:

Z-score shows how many standard deviations a value is from the mean.

- Threshold = 2
- Data points with Z-score > 2 or < -2 are considered outliers.

IQR Method:

The IQR is calculated as:

$$\text{IQR} = Q3 - Q1$$

Then:

- Upper bound = $Q3 + 1.5 \times \text{IQR}$
 - Lower bound = $Q1 - 1.5 \times \text{IQR}$
- Any value outside these bounds is considered an outlier.

The results indicate that no outliers were detected in the dataset using both Z-score and IQR methods. This confirms that the dataset is clean, well-distributed, and does not contain anomalies or extreme values that may affect the performance of data mining models.

Hence, no special handling for outliers is required.

- Show duplicates:

```
➦ Number of duplicate rows: 0  
Empty DataFrame  
Columns: [PatientID, Age, Gender, Ethnicity, EducationLevel, BMI, Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQ  
Index: []  
  
[0 rows x 34 columns]
```

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

- Data Transformation:

1. Encoding:

Encoding is the process of converting non-numeric data into numeric format. This allows the model to effectively process categorical features such as gender, education or ethnicity. However, we did not perform this process on our dataset because it was already pre-encoded (Columns: Gender – Ethnicity – Education Level) in the following manner:

Gender	Ethnicity	Education Level
0: Male 1: Female	0: Caucasian 1: African American 2: Asian 3: Other	0: None 1: High School 2: Bachelor's 3: Higher

Sample of the pre-encoded columns:

C	D	E
Gender	Ethnicity	EducationLevel
0	0	2
0	0	0
0	3	1
1	0	1
0	0	0
1	1	1
0	3	2
0	0	1

2. Normalization:

Before:

	BMI	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	\
0	22.927749	13.297218	6.327112	1.347214	9.025679	
1	26.827681	4.542524	7.619885	0.518767	7.151293	
2	17.795882	19.555085	7.844988	1.826335	9.673574	
3	33.800817	12.209266	8.428001	7.435604	8.392554	
4	20.716974	18.454356	6.310461	0.795498	5.597238	
	SystolicBP	CholesterolTotal	CholesterolLDL	CholesterolHDL		\
0	142	242.366840	56.150897	33.682564		
1	115	231.162595	193.407995	79.028477		
2	99	284.181858	153.322762	69.772292		
3	118	159.582240	65.366637	68.457491		
4	94	237.602184	92.869700	56.874305		
	CholesterolTriglycerides	MMSE	FunctionalAssessment	ADL		
0	162.189143	21.463532	6.518877	1.725883		
1	294.630909	20.613267	7.118696	2.592424		
2	83.638324	7.356249	5.895077	7.119548		
3	277.577358	13.991127	8.965106	6.481226		
4	291.198780	13.517609	6.045039	0.014691		

After:

	BMI	AlcoholConsumption	PhysicalActivity	DietQuality	\
0	0.316960	0.665183	0.633375	0.133931	
1	0.473058	0.227170	0.762862	0.050995	
2	0.111553	0.978276	0.785408	0.181896	
3	0.752163	0.610751	0.843804	0.743443	
4	0.228472	0.923204	0.631707	0.078698	
...	
2144	0.965137	0.078006	0.405291	0.655316	
2145	0.114035	0.938860	0.135925	0.289848	
2146	0.018717	0.229779	0.989841	0.811960	
2147	0.011650	0.433901	0.636096	0.125543	
2148	0.731706	0.394686	0.657802	0.794079	
...	
	SleepQuality	SystolicBP	CholesterolTotal	CholesterolLDL	\
0	0.837564	0.584270	0.615567	0.039538	
1	0.525021	0.280899	0.540822	0.956205	
2	0.945597	0.101124	0.894520	0.688497	
3	0.731994	0.314607	0.063302	0.101085	
4	0.265892	0.044944	0.583781	0.284763	
...	
2144	0.589092	0.359551	0.869803	0.298125	
2145	0.759124	0.696629	0.242102	0.301733	
2146	0.294609	0.280899	0.579928	0.708162	
2147	0.720376	0.146067	0.614435	0.015042	
2148	0.979802	0.853933	0.889283	0.280291	
...	
	CholesterolHDL	CholesterolTriglycerides	MMSE	\	
0	0.171039	0.319802	0.715606		
1	0.738026	0.698711	0.687251		
2	0.622290	0.095072	0.245145		
3	0.605851	0.649922	0.466410		
4	0.461019	0.688892	0.450619		
...		
2144	0.511894	0.526737	0.039881		
2145	0.920845	0.908578	0.215192		
2146	0.996222	0.699201	0.567120		
2147	0.766192	0.271351	0.134235		
2148	0.774181	0.477749	0.370488		
...		
	FunctionalAssessment	ADL			
0	0.652102	0.172486			
1	0.712108	0.259154			
2	0.589697	0.711936			
3	0.896823	0.640894			
4	0.604699	0.001341			
...			
2144	0.023830	0.449224			

Here in the Normalization method, we applied Min-Max scaling to several numerical attributes such as BMI, Alcohol Consumption, Physical Activity, Systolic and Diastolic Blood Pressure, Cholesterol levels, MMSE, Functional Assessment, and ADL. This transformation unified the scale of these features to a range between 0 and 1, helping to balance their influence and improve model performance. Normalization ensures that each feature contributes equally during the analysis and model training phases.

3. Discretization:

Before:

```
0    73
1    89
2    73
3    74
4    89
Name: Age, dtype: int64
```

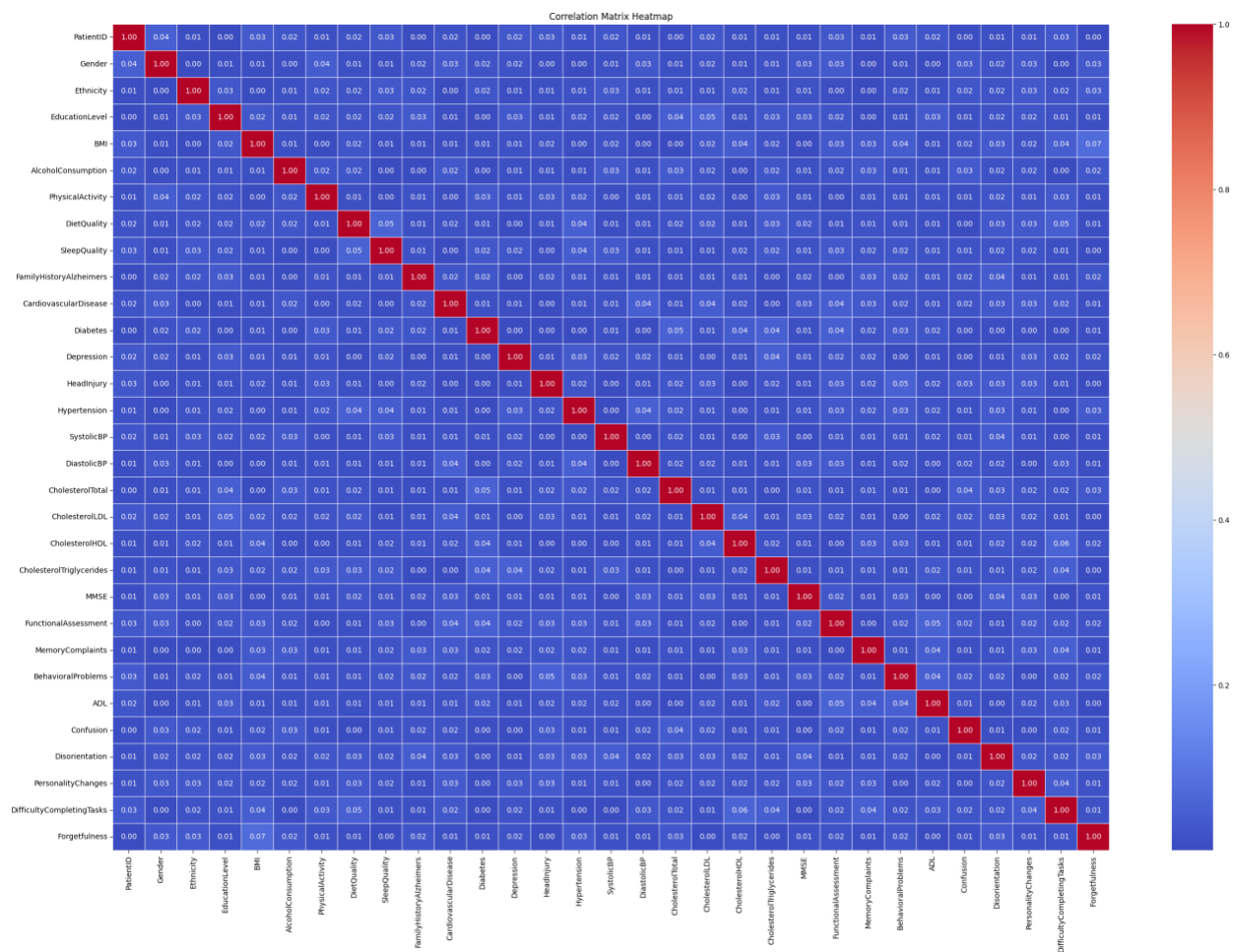
After:

```
0    73-78
1    85-90
2    73-78
3    73-78
4    85-90
...
2144  60-66
2145  73-78
2146  73-78
2147  73-78
2148  67-72
Name: Age, Length: 2149, dtype: category
Categories (5, object): ['60-66' < '67-72' < '73-78' < '79-84' < '85-90']
```

The Discretization step, we transformed the continuous Age values into categorical intervals. This technique grouped individuals into five age ranges: (60–66), (67–72), (73–78), (79–84), and (85–90). This transformation improves the interpretability of the dataset and enables better pattern recognition for age-related trends.

4. Removing highly correlated columns:

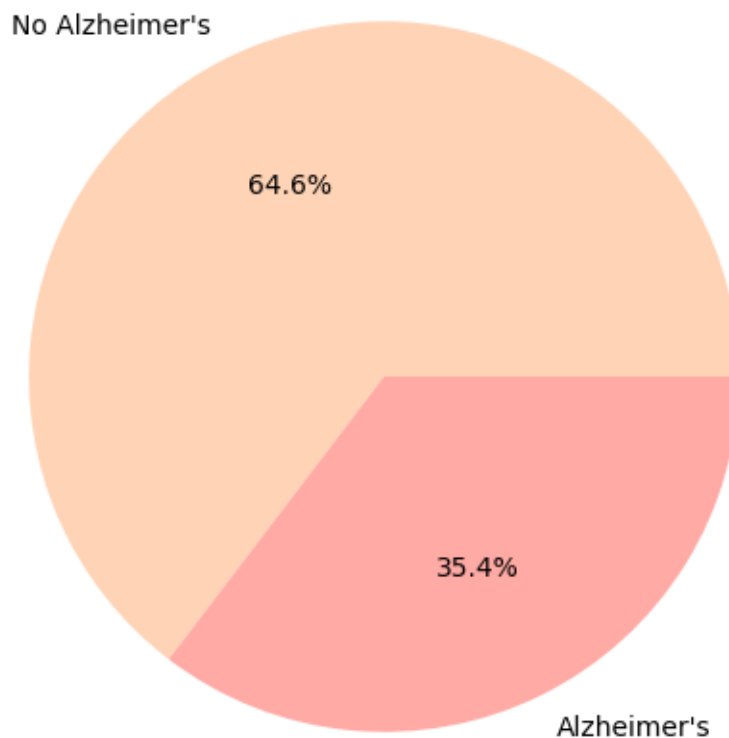
This process involves detecting and removing columns that have a strong linear relationship between them. To perform this data transformation process, we first calculated the correlation coefficient with a threshold of 0.75 for all numeric columns to detect highly correlated features. However, no strongly correlated columns were found, so no columns were removed. To further investigate, we generated a correlation matrix heatmap. The heatmap clearly shows that there is no value above the threshold of 0.75 which means there is no correlation above the threshold we've set, therefore no further handling will be done.



- Balance Data:

Data balancing is the process of adjusting the distribution of samples in a dataset so that each class has an equal number of samples. This step is crucial to prevent class imbalance from affecting the model's performance and reducing any bias towards a particular class. In our dataset we have two possible values of our class label "Diagnosis", we used a pie chart to illustrate the percentage distribution of the class label and found that it was slightly imbalanced, with 64.6% No Alzheimer's and 35.4% Alzheimer's. No further balancing was performed on the dataset as per our instructor's advice.

Class Distribution to Detect Class Imbalance



5. Data Mining Technique:

We used 2 techniques for this data mining process which are Classification (Supervised) and clustering (Unsupervised).

All packages and methods used:

	Package	Methods
Classification	<ul style="list-style-type: none">scikit-learn (sklearn) package: we used this package a lot due to the variety of methods that assist us in the process such as (decision tree, Logistic regression classifier and many more.Pickle: we used this package to save and load our model during the prediction process.	<code>train_test_split()</code> , <code>StandardScaler()</code> , <code>MinMaxScaler()</code> , <code>zscore()</code> , <code>LogisticRegression()</code> , <code>DecisionTreeClassifier()</code> , <code>plot_tree()</code> , <code>GaussianNB()</code> , <code>RFE()</code> , <code>confusion_matrix()</code> , <code>accuracy_score()</code> , <code>recall_score()</code> , <code>precision_score()</code> , <code>ConfusionMatrixDisplay()</code> , <code>dump()</code> , <code>load()</code>
Clustering	<ul style="list-style-type: none">scikit-learn (sklearn) package: we used this package in the clustering process as well to implement the Kmeans method we followed as well as computing the silhouette scores.Yellowbrick package: complemented the previous package by enabling us to visualize the silhouette scores for each cluster.Kneed package: used along with the elbow methods to identify the elbow point which tells us the optimal number of clusters.	<code>KMeans()</code> , <code>silhouette_score()</code> , <code>silhouette_samples()</code> , <code>SilhouetteVisualizer()</code> , <code>KneeLocator()</code>
Extra	<ul style="list-style-type: none">Pandas: used for loading the dataset and manipulating it to what we see fit (e.g. splitting into features and target).Numpy: used for numerical operations.matplotlib.pyplot: created visualizations for data presentation such as the decision tree.Seaborn: complements matplotlib by creating more attractive and informative statistical graphics	<code>read_csv()</code> , <code>columns.tolist()</code> , <code>remove()</code> , <code>subplots()</code> , <code>title()</code> , <code>show()</code> , <code>savefig()</code>

1. Classification technique:

In this process we used the decision tree where the model recursively splits the data at decision nodes based on the attribute selection measures. We've chosen to reach the final decision which assists in predicting the class label (Positive or Negative), we specifically used the Information Gain and Gini Index as the attribute selection measures.

Firstly, to start the classification process, we first split the data into features and target. This is a vital step to ensure that the model differentiates between the input variables (features) that will be utilized to predict the final output (target). In our case specifically the target was the class label "Diagnosis", and the rest of the dataset was used as features.

Secondly, after the initial splitting of target and features, we had to split the dataset into two sections. We split the dataset into a training section, used to train the classifier and build the decision tree, and a testing section, used to evaluate the model's predictive performance. We tried 3 separate splits which are:

- A. 80% for training and 20% for testing.
- B. 70% for training and 30% for testing.
- C. 60% for training and 40% for testing.

Thirdly, we visualized our results using the confusion matrix and decision tree to help us gain a better understanding of our model's ability to predict. After that we assessed the results by calculating additional performance measures, such as accuracy and sensitivity, this would help us in comparing the 3 splits above to determine the best one.

Lastly, we saved the model and created new data to test the model's prediction ability in all 3 splits. The model was able to accurately assign the new data to the correct class labels, demonstrating its effectiveness in making predictions and distinguishing between individuals at risk and those not at risk.

2. Clustering technique:

Clustering falls under unsupervised learning which means it does not require a class label, so we started by removing the class label column from the dataset. In addition, we ensure our dataset is clear of any non-numeric column values by preprocessing it in the previous phase.

After performing what's mentioned previously. the next step is to scale the attributes, this step ensures all attributes are on the same comparable scale meaning that no attribute will influence or outweigh the other attributes.

In our data mining process, we used K-means clustering which segments the data into a K number of clusters and assigns each data point into its closest centroid, after the initial assignment is done, we iteratively recalculate the center and reassign the objects into clusters. This is done until we minimize the total within-cluster variance (WSS) which gives an estimation of the dispersion of the observations within a cluster.

We decided to perform K-means clustering with 3 different sizes of K; to determine the three sizes we followed two methods:

- Number of K with highest average silhouette coefficient: this is a metric used to evaluate the quality of clustering, it measures how well an object fits within its assigned cluster in comparison to other clusters, after calculations we determined 4 to be the best number of clusters with the highest average silhouette coefficient, closely followed by 5 clusters.

- Elbow method: commonly used technique to identify the optimal number of clusters for K-means clustering. It works by plotting the within-cluster sum of squares (inertia) against different values of k. Inertia indicates how closely the data points are grouped within a cluster, and the Elbow point is where increasing the number of clusters no longer leads to a significant drop in inertia. After calculations we determined 7 to be the best number of clusters.

After determining the 3 values of the number of clusters, K-means clustering on all values. following that we calculated WSS, B-Cubed precision and recall which assisted us in determining the clustering performance in all 3 values. Along with our calculations we visualized the silhouette coefficient for all clusters, to give us a better understanding of our results.

6. Evaluation and Comparison:

- **Classification**

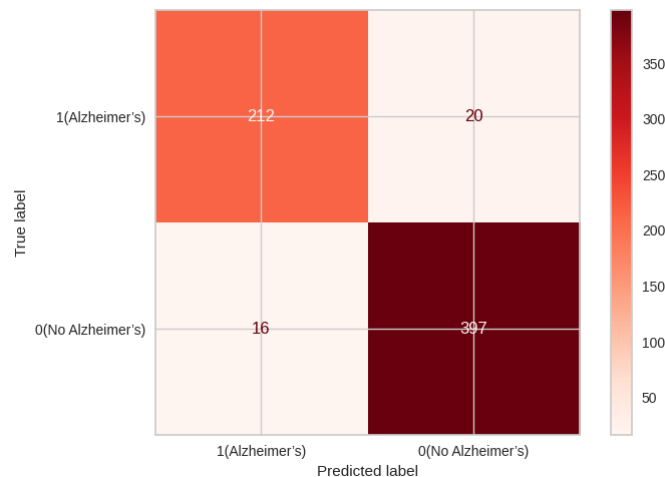
In our dataset we used two classification methods 1- Information gain 2-Gini index and with both methods we used three different splits to the dataset

- 1- 70% training 30% testing
- 2- 80% training 20% testing
- 3- 60% training 40% testing

- **Information Gain**

1- 70% training 30% testing split

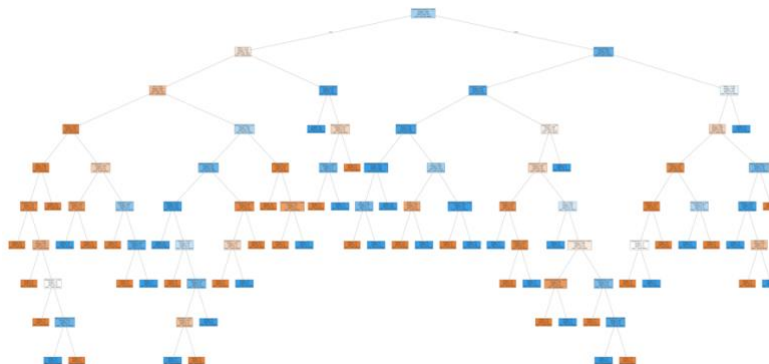
- **Confusion matrix**



- **True Positives (TP):** 212 (Correctly predicted Alzheimer's)
- **False Negatives (FN):** 20 (Alzheimer's patients misclassified as non-Alzheimer's)
- **False Positives (FP):** 16 (Non-Alzheimer's patients misclassified as having Alzheimer's)
- **True Negatives (TN):** 397 (Correctly predicted non-Alzheimer's)

The confusion matrix shows that the model performed well in distinguishing between Alzheimer's and non-Alzheimer's cases. It correctly predicted most instances in both classes, with only a few misclassifications. The number of false positives and false negatives is low, indicating that the model is reliable and balanced in its predictions.

- **Decision Tree**

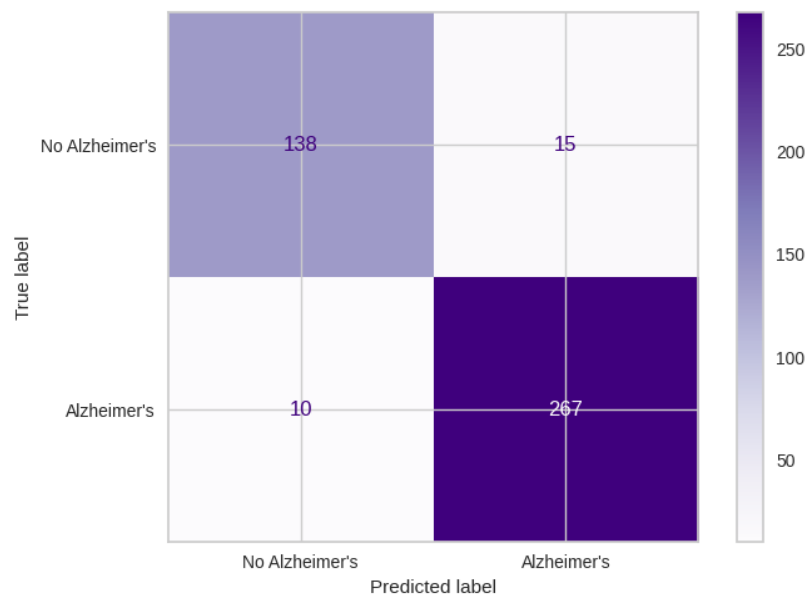


- **Accuracy**

Accuracy: **94.88%** This indicates that the model correctly classified approximately 95 out of every 100 test samples. Accuracy reflects the overall correctness of the model across both classes patients with and without Alzheimer's.

2- 80% training 20% testing split

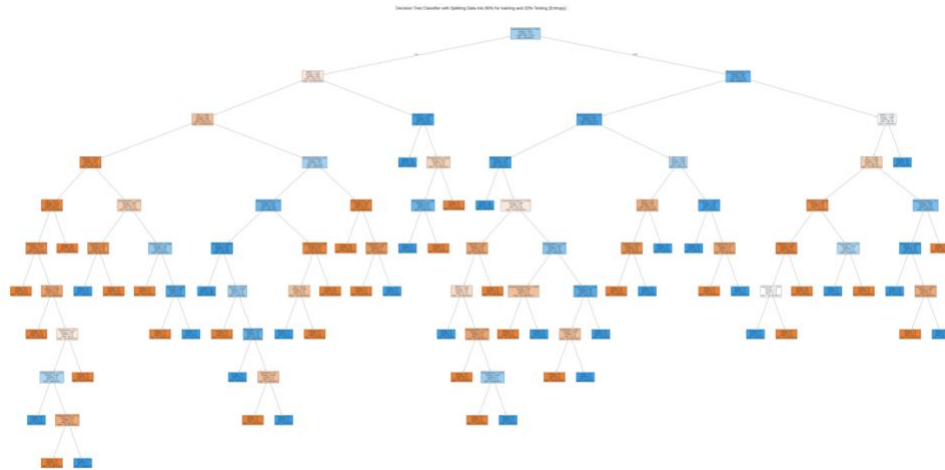
- **Confusion Matrix**



- **True Positives (TP):** 267 (Correctly predicted Alzheimer's)
- **False Negatives (FN):** 10 (Alzheimer's patients misclassified as non-Alzheimer's)
- **False Positives (FP):** 15 (Non-Alzheimer's patients misclassified as having Alzheimer's)
- **True Negatives (TN):** 138 (Correctly predicted non-Alzheimer's)

These results indicate that the model performs very well in detecting both classes, with very few misclassifications. The high number of true positives and true negatives reflects the model's reliability in classifying Alzheimer's cases using the 80/20 train-test split.

- **Decision Tree**

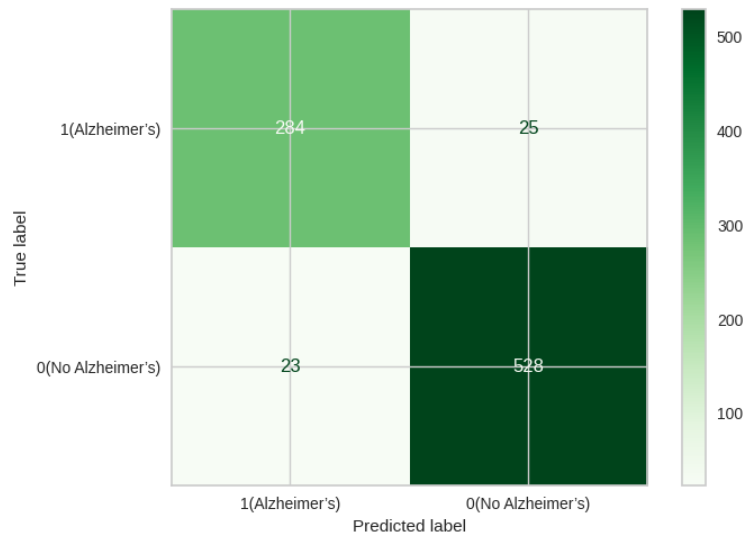


- **Accuracy**

The model achieved an accuracy of **94%** using the 80/20 train-test split with Information Gain, reflecting strong overall performance and effective generalization on unseen data.

3- 60% testing 40% training split

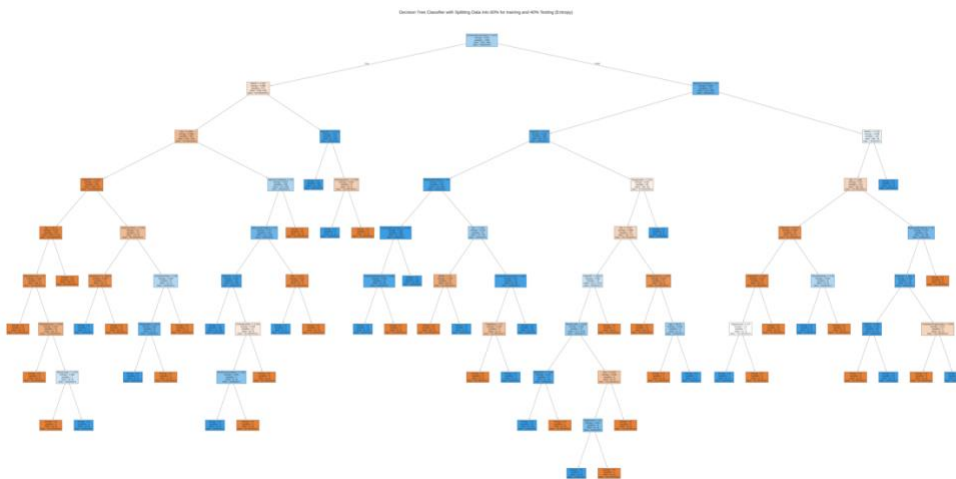
- **Confusion Matrix**



- **True Positives (TP):** 284 (Correctly predicted Alzheimer's)
- **False Negatives (FN):** 25 (Alzheimer's patients misclassified as non-Alzheimer's)
- **False Positives (FP):** 23 (Non-Alzheimer's patients misclassified as having Alzheimer's)
- **True Negatives (TN):** 528 (Correctly predicted non-Alzheimer's)

The confusion matrix shows that the model performed well, with high numbers of correct predictions in both classes and a low number of misclassifications.

- **Decision Tree**



- **Accuracy**

The model achieved an accuracy of **94.42%** using the 60/40 train-test split with Information Gain. This high accuracy demonstrates the model's strong performance even when evaluated on a larger test set, confirming its effectiveness in correctly classifying Alzheimer's and non-Alzheimer's cases.

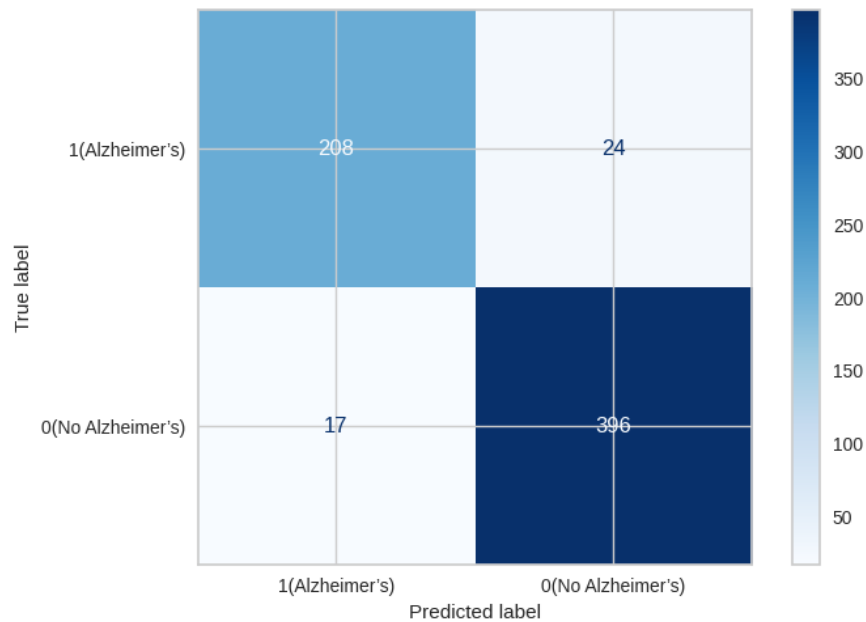
Accuracy Comparison Across Data Splits Using Information Gain

Split	70% training 30% testing	80% training 20% testing	60% training 40% testing
Accuracy	94.88%	94%	94.42%

Findings:

The table above compares the model's accuracy across three different training-testing splits using Information Gain as the attribute selection method. All three splits yielded high accuracy, with slight variations. The 70/30 split achieved the highest accuracy at **94.88%**, followed closely by the 60/40 split at **94.42%**, and the 80/20 split at **94%**. These results suggest that the model performs consistently well across different data sizes, **with the 70/30 split offering a slightly better balance between training and testing data.**

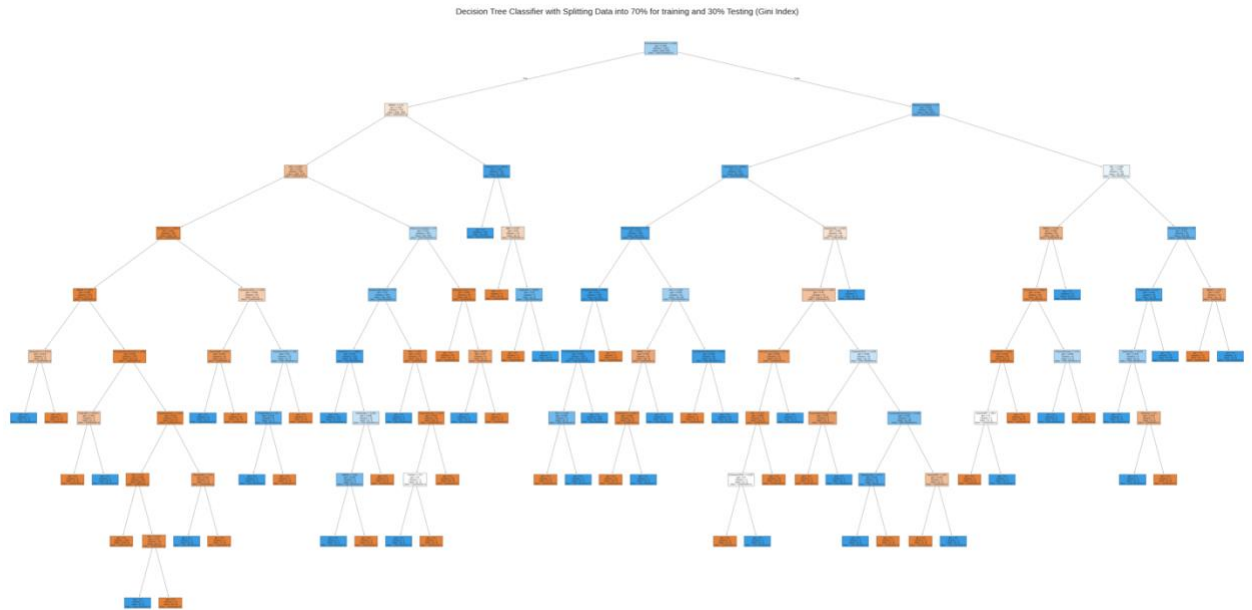
- **Gini Index**
- **70% training 30% testing split**
- **Confusion Matrix**



- **True Positives (TP):** 208 (Correctly predicted Alzheimer's)
- **False Negatives (FN):** 24 (Alzheimer's patients misclassified as non-Alzheimer's)
- **False Positives (FP):** 17 (Non-Alzheimer's patients misclassified as having Alzheimer's)
- **True Negatives (TN):** 396 (Correctly predicted non-Alzheimer's)

The confusion matrix shows that the model performed well with the Gini Index, correctly identifying most cases with a relatively small number of misclassifications.

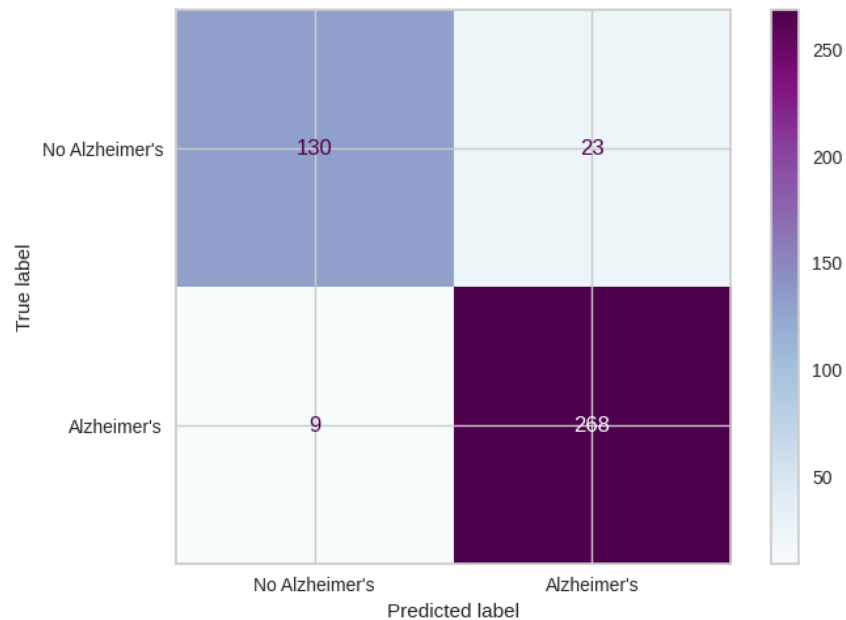
- **Decision Tree**



- **Accuracy**

The model achieved an accuracy of **93.64%** using the 70/30 train-test split with Gini Index as the attribute selection method. This indicates strong overall performance, with the model correctly classifying the majority of both Alzheimer's and non-Alzheimer's cases.

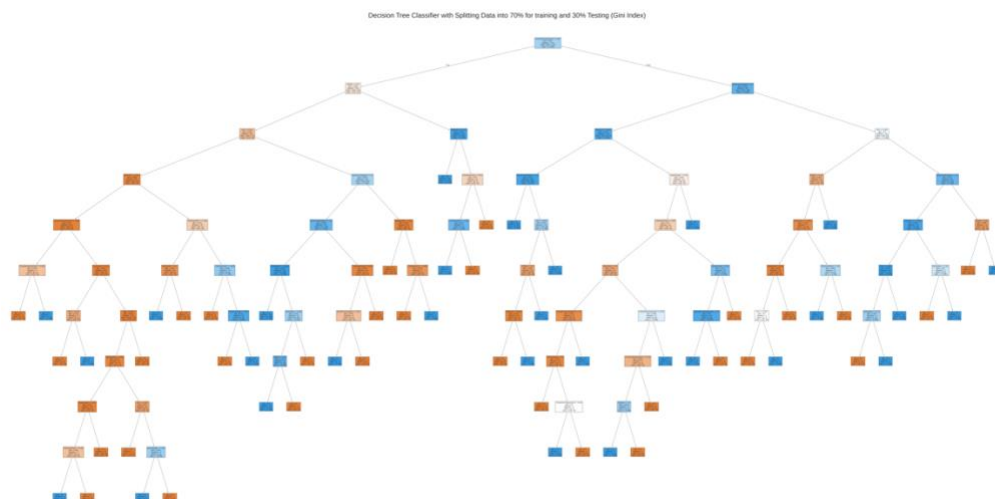
- **80% training 20% testing split**
- **Confusion Matrix**



- **True Positives (TP):** 268 (Correctly predicted Alzheimer's)
- **False Negatives (FN):** 9 (Alzheimer's patients misclassified as non-Alzheimer's)
- **False Positives (FP):** 23 (Non-Alzheimer's patients misclassified as having Alzheimer's)
- **True Negatives (TN):** 130 (Correctly predicted non-Alzheimer's)

The confusion matrix reflects strong model performance with the Gini Index, showing high accuracy in detecting Alzheimer's cases and a relatively low number of misclassifications in both classes.

- **Decision Tree**



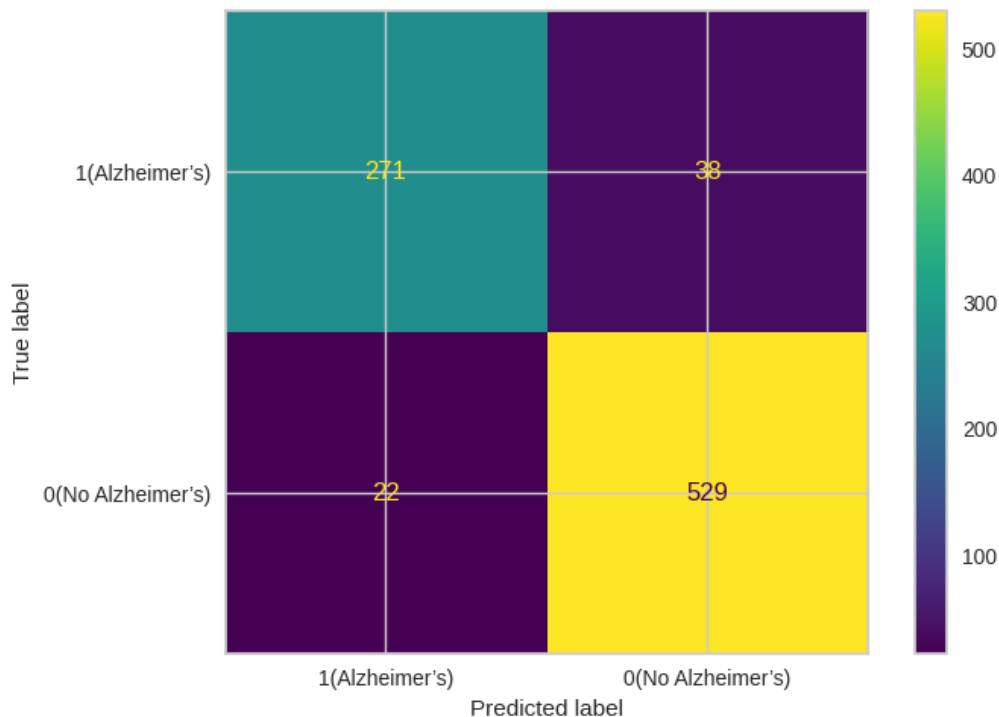
- **Accuracy**

The model achieved an accuracy of **92%** using the 80/20 train-test split with Gini Index, which is slightly lower than the accuracies obtained with the same split using Information Gain and other splits.

Nonetheless, the model still performed well, correctly classifying most Alzheimer's and non-Alzheimer's cases.

3-60% testing 40% training split

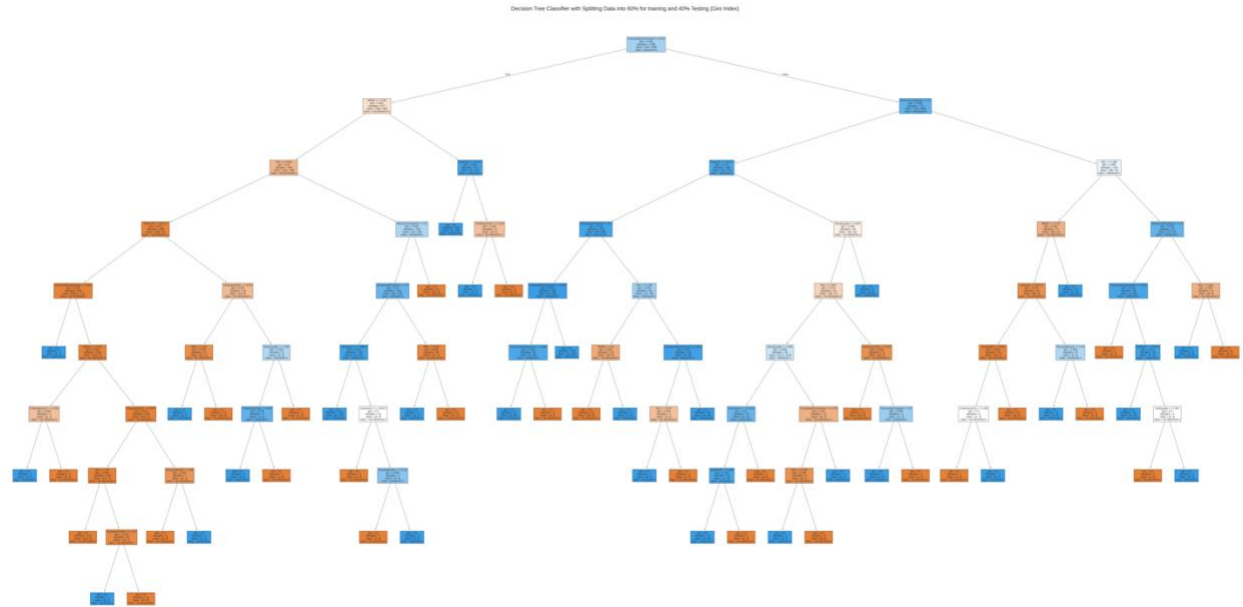
- **Confusion Matrix**



- **True Positives (TP):** 271 (Correctly predicted Alzheimer's)
- **False Negatives (FN):** 38 (Alzheimer's patients misclassified as non-Alzheimer's)
- **False Positives (FP):** 22 (Non-Alzheimer's patients misclassified as having Alzheimer's)
- **True Negatives (TN):** 529 (Correctly predicted non-Alzheimer's)

This confusion matrix shows that the model maintained strong overall performance, though the number of false negatives slightly increased compared to other splits. It still correctly identified most cases in both classes using the Gini Index with a 60/40 split.

- **Decision Tree**



- **Accuracy**

The model achieved an accuracy of **93.02%** using the 60/40 train-test split with Gini Index. While this is slightly lower than the accuracies achieved with other splits and with Information Gain, the model still demonstrated reliable performance, correctly classifying the majority of Alzheimer's and non-Alzheimer's cases.

Accuracy Comparison Across Data Splits Using Gini Index

Split	70% training 30% testing	80% training 20% testing	60% training 40% testing
Accuracy	93.64%	92%	93.02%

Findings:

The table above presents the model's accuracy across three different training-testing splits using Gini Index as the attribute selection method. Among the three, the **70/30 split achieved the highest accuracy at 93.64%**, making it the most effective configuration when using Gini Index. The **60/40** and **80/20** splits followed with accuracies of **93.02%** and **92%**, respectively. Although all splits showed strong and consistent performance, the 70/30 split provided the best results in terms of correctly classifying Alzheimer's and non-Alzheimer's cases.

- **Accuracy Comparison Across Data Splits Using Gini Index and Information Gain**

Split	70% training 30% testing		80% training 20% testing		60% training 40% testing	
Method	Information Gain	Gini Index	Information Gain	Gini Index	Information Gain	Gini Index
Accuracy	94.88%	93.64%	94%	92%	94.42%	93.02%

Findings:

The table above provides a comprehensive comparison of model accuracy using both Information Gain and Gini Index across three different train-test splits. Overall, **Information Gain consistently outperformed Gini Index** in each split.

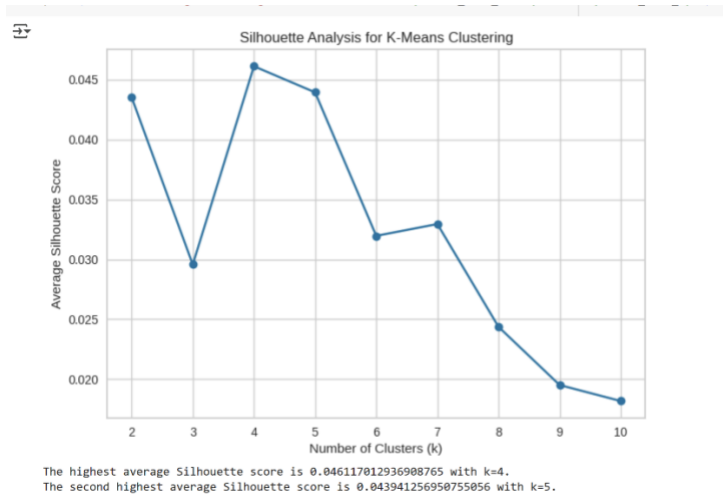
- In the **70/30 split**, Information Gain had the highest accuracy of **94.88%**, which is the best result overall.
- In the **80/20 split**, Information Gain also performed better (**94%**) than Gini (**92%**).
- In the **60/40 split**, Information Gain again had higher accuracy (**94.42%**) compared to Gini (**93.02%**).

From all the results, we can clearly see that the **70/30 split with both Information Gain and Gini Index** gave the best performance among all methods and splits.

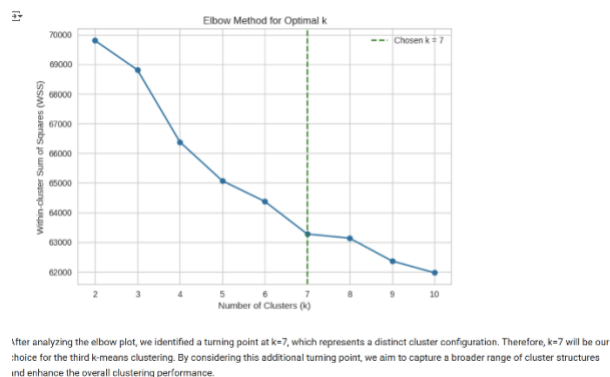
- Clustering

We chose 3 different sizes [4, 5, 7] based on the result of the validation methods that we used. Then, we used these sizes to perform the K-means clustering.

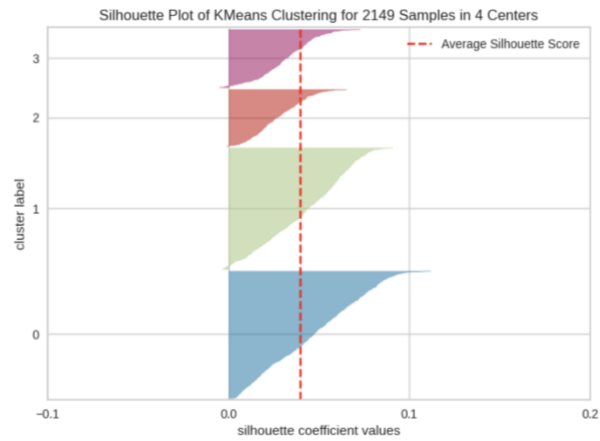
Silhouette Analysis: evaluates how similar a data point is to its own cluster compared to other clusters. A higher silhouette score indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters.



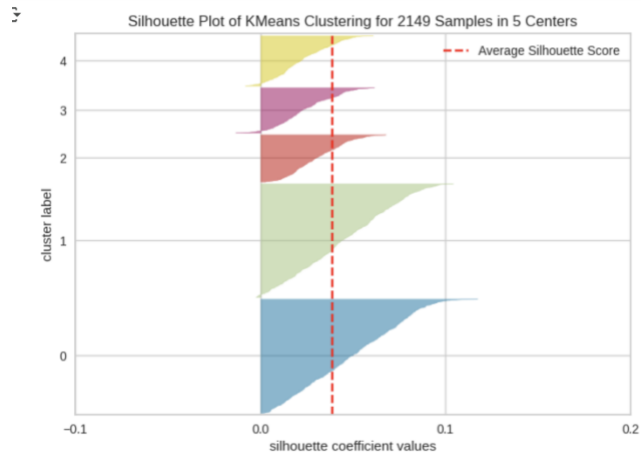
Elbow Method: is used to determine the most suitable number of clusters for K-means by plotting the within-cluster sum of squares (inertia) against the number of clusters and identifying the point where the rate of decrease sharply changes, forming an elbow.



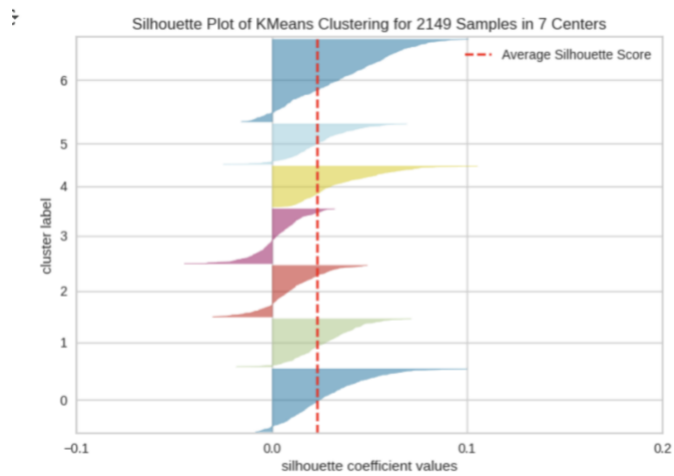
silhouette scores [K=4]



silhouette scores [K=5]



silhouette scores [K=7]



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



	K	Average Silhouette Score	WSS
0	4	0.0461	66380.18
1	5	0.0439	65069.33
2	7	0.0329	63276.61

7. Findings

In the beginning, we selected a dataset that represents patients' health, behavioral, and cognitive characteristics to predict the probability of having Alzheimer's Disease. The main goal is to support early detection and provide medical insights that contribute to improving patients' quality of life.

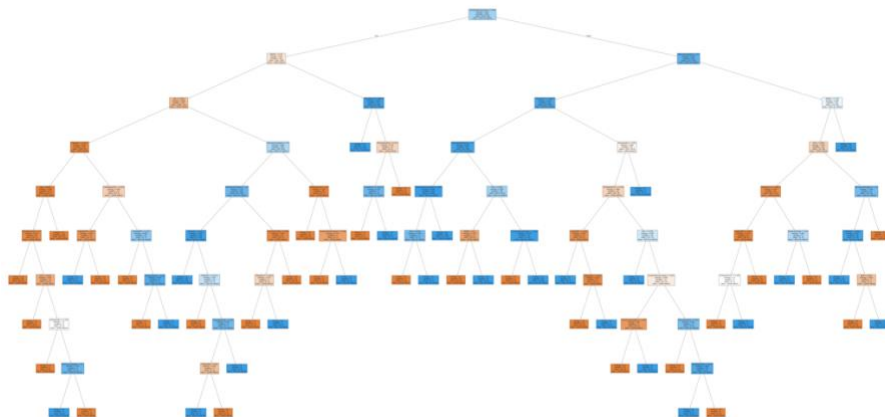
To ensure accurate and reliable model performance, we applied several preprocessing techniques to enhance the data quality. We used visualization tools such as boxplots and histograms to better understand the distribution of the attributes and to detect outliers and missing values. Based on these visualizations and descriptive statistics, we removed null values and handled outliers that may negatively impact the model results.

We also applied data transformation techniques such as normalization and discretization. These steps helped unify the scales of numerical attributes (e.g., BMI, MMSE, Functional Assessment), and improved the overall handling of data during the modeling process.

As part of the data mining process, we applied two main techniques: classification and clustering. For classification, we used the Decision Tree algorithm with two different splitting criteria (Entropy and Gini Index), and tested the model using three different train/test splits:

- 70% Training, 30% Testing
- 80% Training, 20% Testing
- 60% Training, 40% Testing

Among these splits, the model using the 70/30 split with the Information Gain criterion achieved the highest accuracy (~94.88%), making it the most effective configuration.



Interpreting the Decision Tree

From the plot of the decision tree (Information Gain based 70/30), we observed the following:

- The root node is "Functional Assessment", which had the highest information gain. This variable reflects patients' ability to perform daily tasks and is highly indicative of cognitive decline.
- Lower Functional Assessment values were strongly associated with Alzheimer's diagnosis.
- The next important attribute was MMSE (Mini-Mental State Examination), which further separated patients based on cognitive scores.
- Additional splits were made on Memory Complaints, Smoking, Sleep Quality, and Cholesterol Total.
- Patients with low MMSE, memory issues, and negative lifestyle indicators were commonly classified as Alzheimer's positive.
- Features like Gender, Ethnicity, and Education Level had limited impact on the decision process.

This tree provided clear, medically relevant rules that align with clinical patterns in Alzheimer's progression.

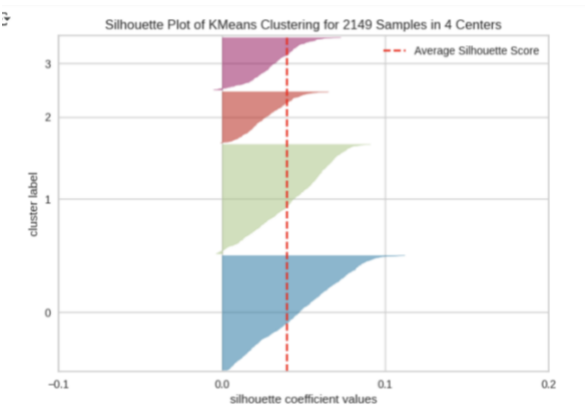
- Clustering:

From the analysis, we applied different clustering evaluation techniques to determine the optimal number of clusters (K) for our dataset. Based on the results of the evaluation metrics including WSS (Within-Cluster Sum of Squares) and the average silhouette score, we focused on the values of K = 4, 5, and 7. The following table summarizes the results:

	K	Average Silhouette Score	WSS
0	4	0.0461	66380.18
1	5	0.0439	65069.33
2	7	0.0329	63276.61

As shown in the table, **K = 4** yielded the highest average silhouette score, indicating that the clusters formed are more distinct and well-separated compared to other values. While the WSS continues to decrease as K increases (which is expected), the silhouette score helps in identifying the best K that maintains both cohesion and separation.

Additionally, we visualized the silhouette plot for K = 4 using 2149 samples. The chart below supports our choice, as the majority of the samples have positive silhouette scores, meaning they are appropriately grouped within their clusters and are well separated from other clusters.



The graph of K-Means Clustering for 2149 samples in 4 centers shows that the silhouette scores are mostly positive, suggesting that the samples are relatively well-clustered. The overall average silhouette score (around 0.0461) reflects moderate structure in the data.

However, it also indicates that while the clustering is acceptable, the clusters may not be perfectly distinct; some data points might fall near the border of two clusters, leading to ambiguity.

While clustering provided useful insights into how the data can be grouped, classification is more suitable for our dataset since it contains labeled outcomes. Clustering helped explore the structure, but classification offers more accurate and reliable predictions when labels like "Diagnosis" are available.

7.3 Final Evaluation and Comparison

Classification (Supervised):

- Achieved the highest accuracy (~94.88%) with the 70/30 split using Information Gain.
- Highlighted important features such as Functional Assessment and MMSE.
- Provided clear, interpretable rules that support medical diagnosis.

Clustering (Unsupervised):

- Useful for exploratory analysis and revealing hidden structures.
- Helped identify patient groups based on health and cognitive profiles.
- Moderate silhouette scores indicated acceptable but not perfect separation.

Conclusion: Since our dataset includes labeled outcomes ("Diagnosis"), classification is more suitable for solving the primary problem of predicting Alzheimer's risk. Clustering served as a complementary technique that added structural insights and supported early screening potential.

Together, both methods enhanced our understanding of the data and contributed to meaningful and medically relevant findings.

8. References:

- [1] R. El Kharoua, “Alzheimer’s Disease Dataset,” *Kaggle*, 2023. [Online]. Available: <https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset>. [Accessed: Jan. 20, 2025].
- [2] L. aldbays, “Phase2.ipynb,” *GitHub*, 2024. [Online]. Available: <https://github.com/leen204/mining-project/blob/main/Reports/%20Phase2.ipynb>. [Accessed: Jan. 20, 2025].
- [3]” Labs and Lecture Slides”, College of Computer Science, Department of Information Technology, King Saud University. [Accessed: Jan. 20, 2025].