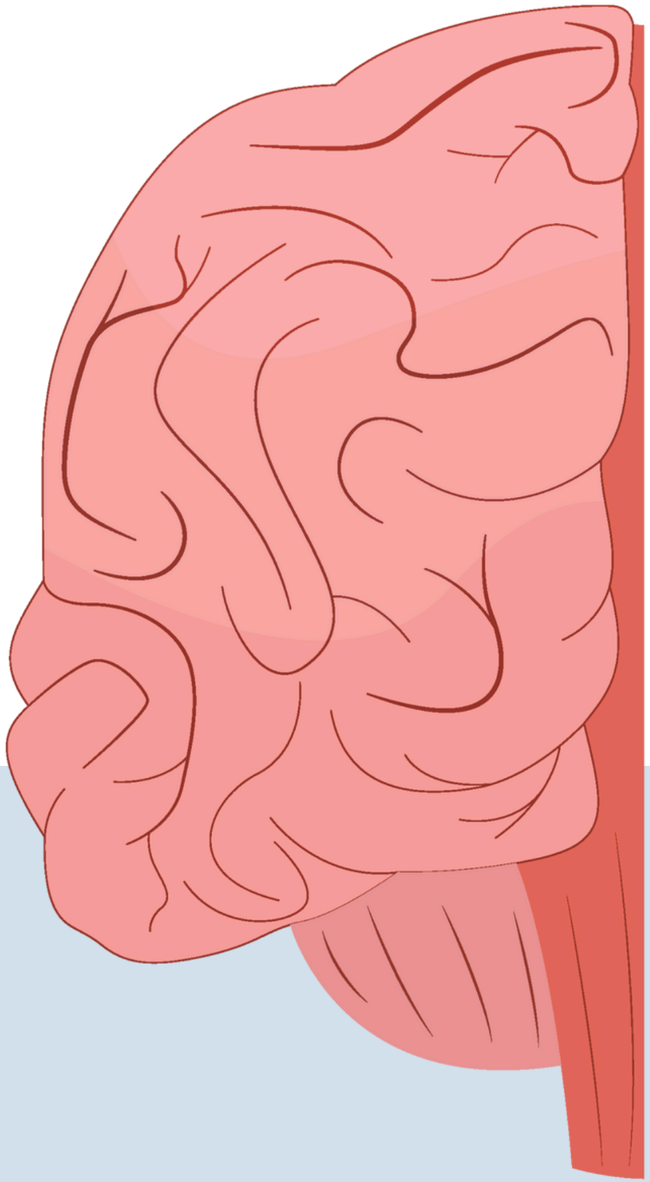


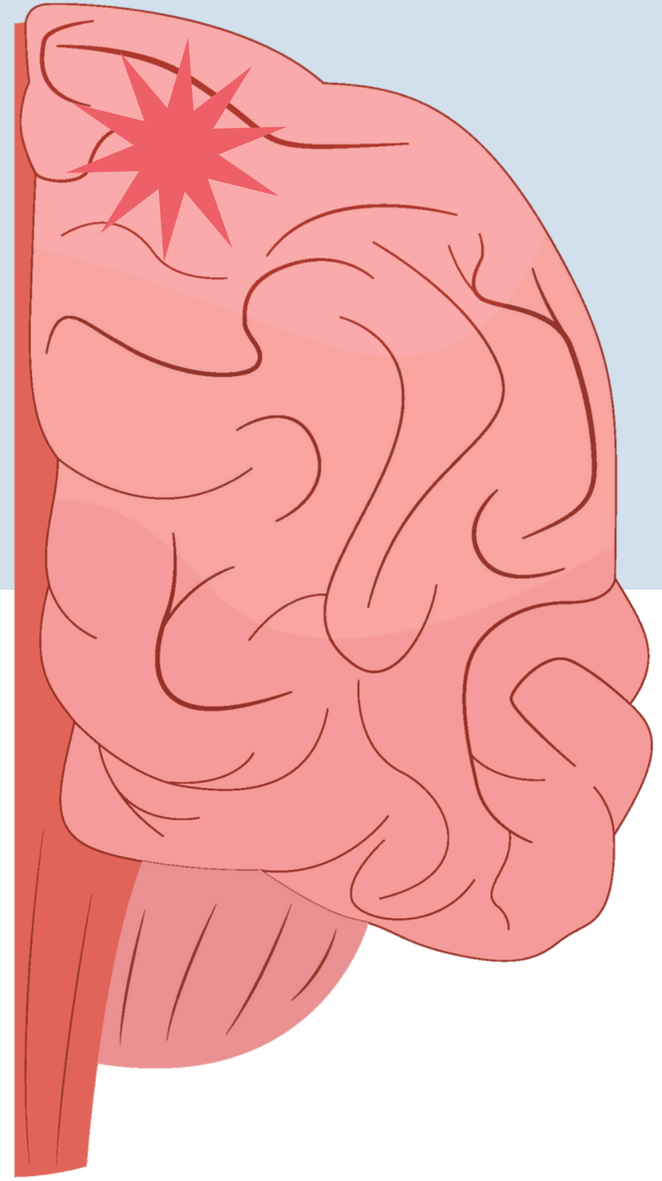
Brain Stroke *Predicting*

Supervised by: Hessah Alsaaran

Table of contents

01

Introduction



Introduction

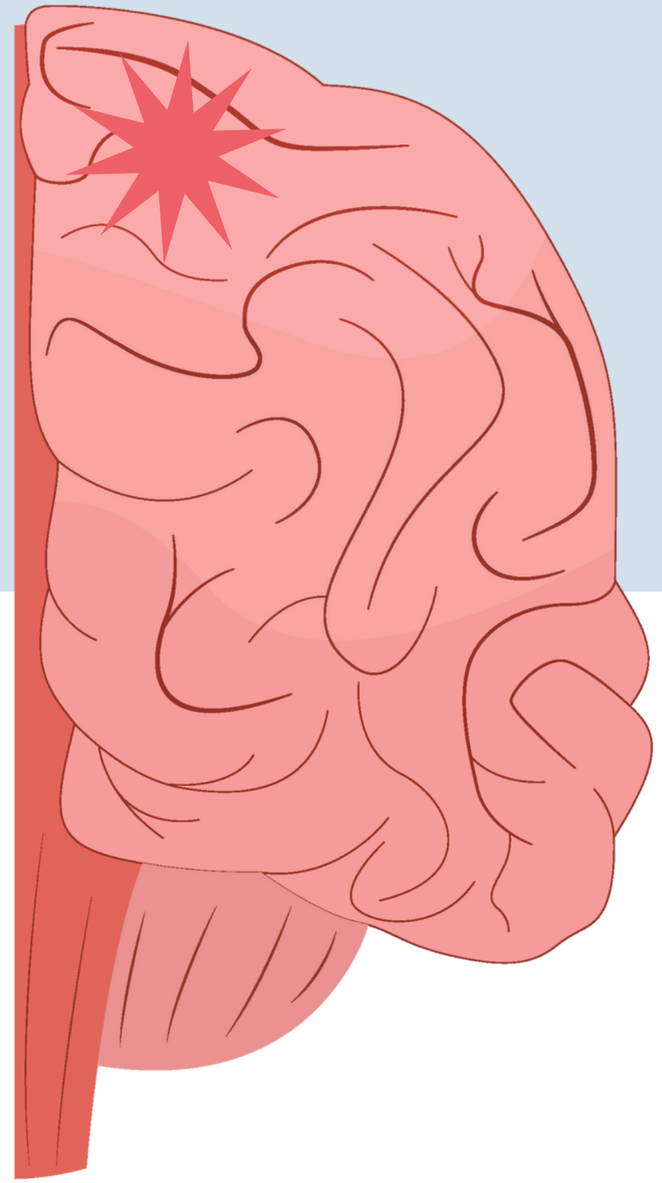
Problem statement

Recently, the incidence of brain strokes has been on the rise, becoming increasingly common among individuals. leading to severe health issues.

Our project aims to analyze patient data to identify risk factors and predict stroke likelihood, enabling individuals to take preventive measures.

02

Dataset



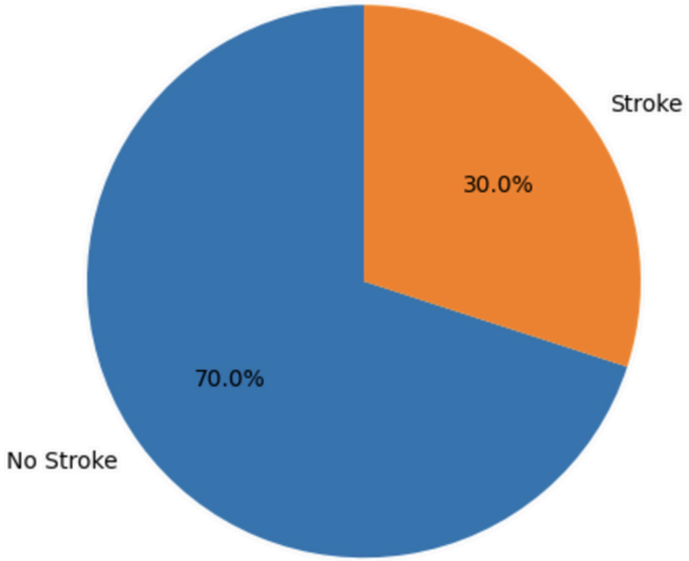
General Information about the dataset

- Number of objects in original dataset: 4981.
- Data distribution: non stroke= 4733, stroke= 248.
- Number of attributes: 11.
- Class labels: stroke.
- Missing values: there is no missing values.

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	67	0	1	0	0	0	228.69	36.6	0	1
0	80	0	1	0	0	1	105.92	32.5	1	1
1	49	0	0	0	0	0	171.23	34.4	2	1

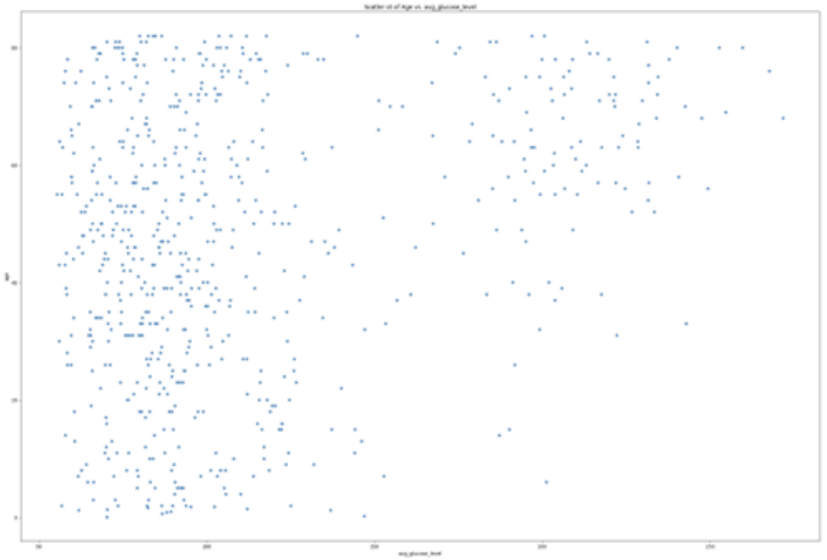
General Information about the dataset

Distribution after sampling

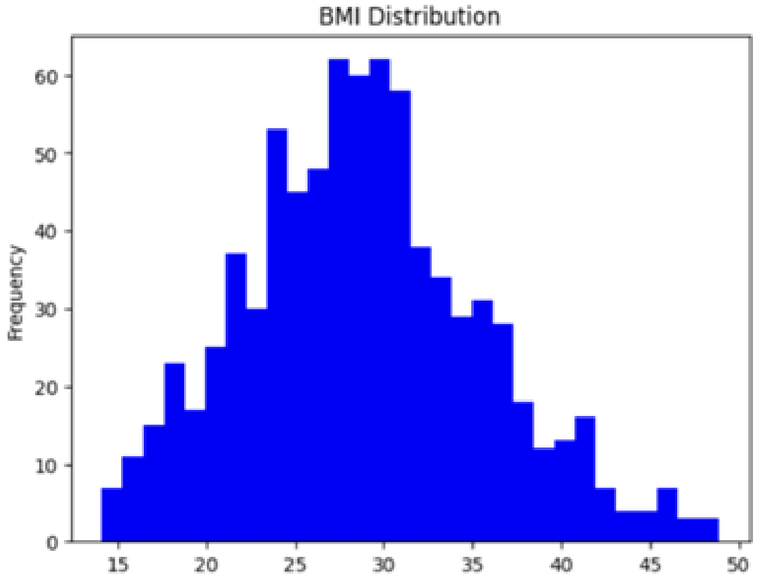


- Number of objects in sampled dataset: 800.
- Data distribution: non stroke= 560, stroke= 240.

General Information about the dataset



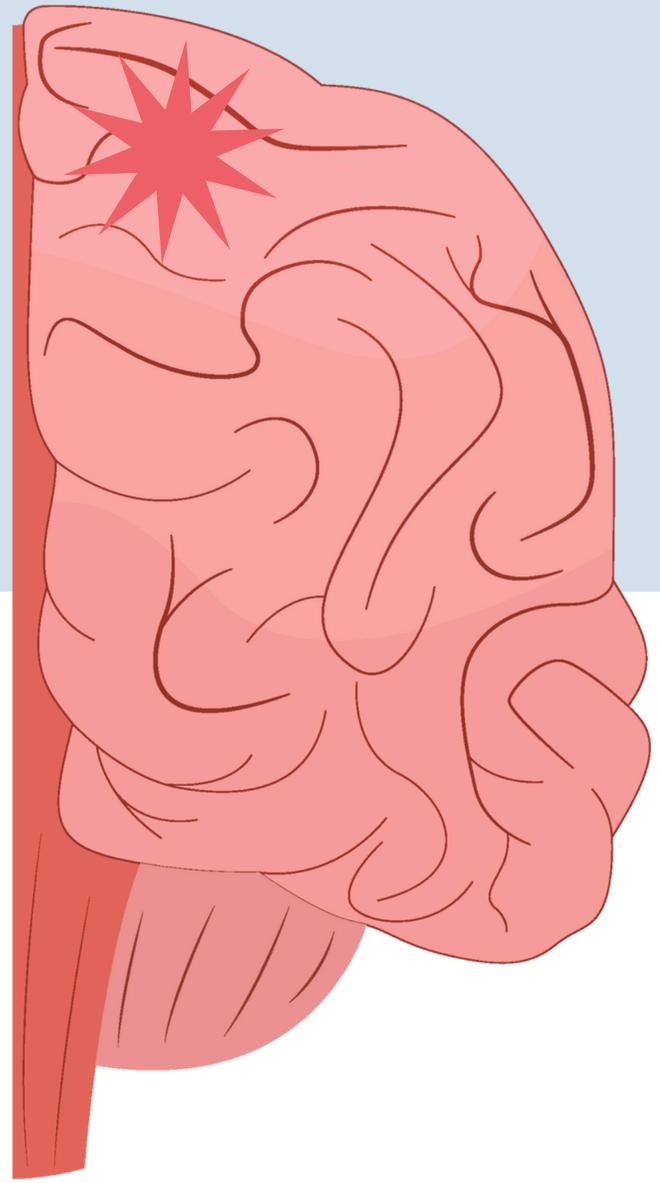
- The majority of individuals have an average glucose level within the range of 60 to 125. A significant portion of those with glucose levels exceeding 125 are older adults, typically aged 50 and above.



- The histogram shows that most individuals have a BMI between 25 and 35, indicating that the majority fall within the normal to overweight range. The peak BMI is around 25-30 which has the highest frequency with 60. There are relatively few cases of underweight (BMI below 20) or extreme obesity (BMI above 40), suggesting that outliers on both ends are rare.

03

Classification



Classification

- **Supervised learning** is a technique used to classify data into predefined categories. It plays a crucial role in improving decision-making.

- In our case, we trained our model to predict whether a patient is categorized as having a **stroke** or **non-stroke** based on the dataset we prepared. This classification allows us to provide targeted interventions and improve patient outcomes.

Classification

This technique involves dividing the dataset into two sets:

- **Training Dataset**: Used to train the model and learn patterns from the input data.
- **Testing Dataset**: Used to evaluate the model's performance on unseen data.

To achieve the best results, we experimented with three different dataset splits, applying both Entropy and Gini Index as splitting criteria for our model:

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

Evaluation of Classification

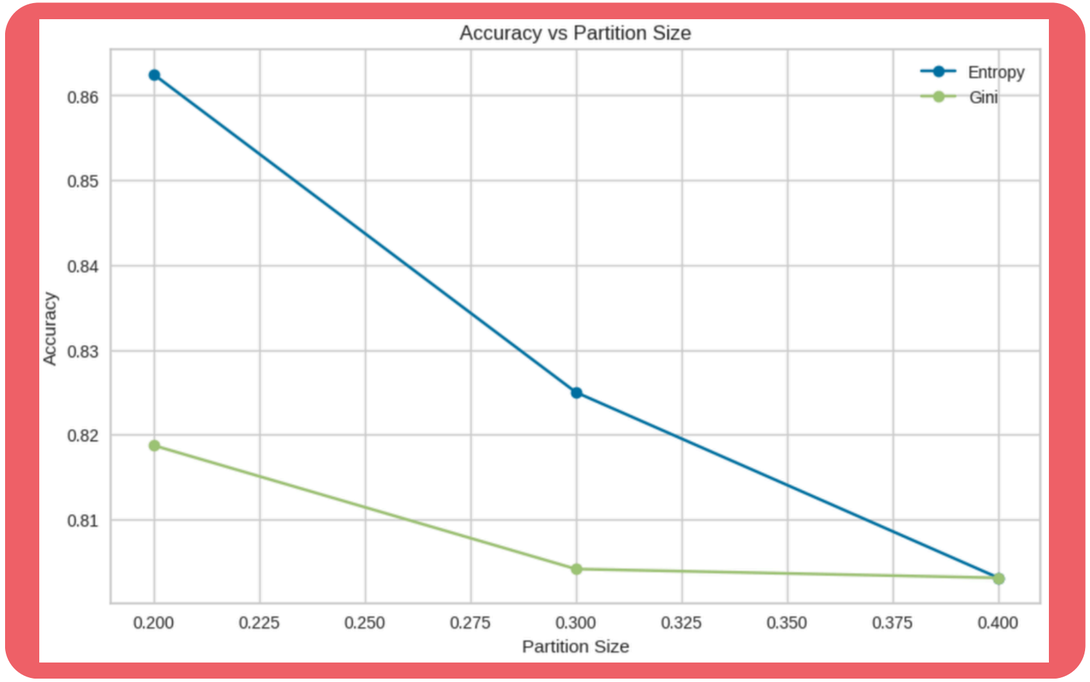
Information Gain (Entropy) Results:

Metric	70% Training, 30% Testing	80% Training, 20% Testing	60% Training, 40% Testing
Accuracy	83%	80%	80%
Sensitivity	72%	80%	69%
Specificity	87%	87%	85%
Precision	70%	72%	68%

Gini Index Results:

Metric	70% Training, 30% Testing	80% Training, 20% Testing	60% Training, 40% Testing
Accuracy	80%	81%	80%
Sensitivity	72%	79%	72%
Specificity	83%	82%	84%
Precision	66%	63%	67%

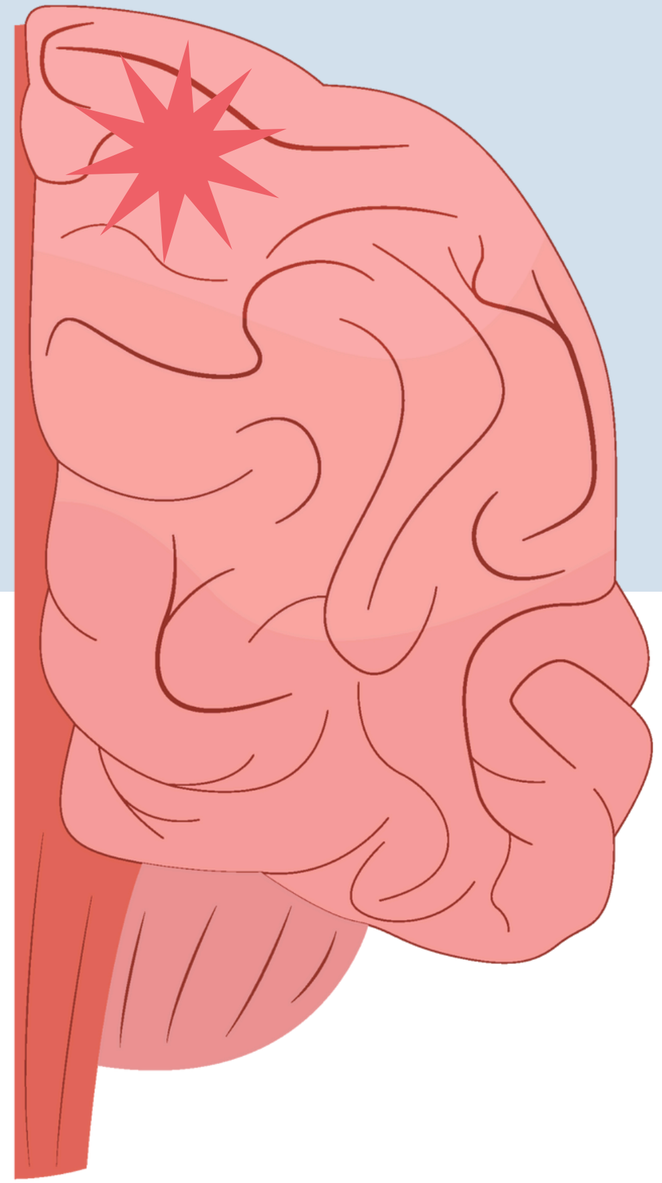
Evaluation of Classification



The graph shows the relationship between partition size and classification accuracy. The “Entropy” method consistently performs better than the “Gini” method across different partition sizes. However, accuracy decreases as the test partition size increases.

04

Clustering



Clustering

- Clustering is unsupervised learning; it will group objects in a cluster based on similarity and dissimilarity.

- Our model will create a set of clusters for the patient who has similar characteristics, then these clusters will be used to predict new patients' results.

Clustering - K means

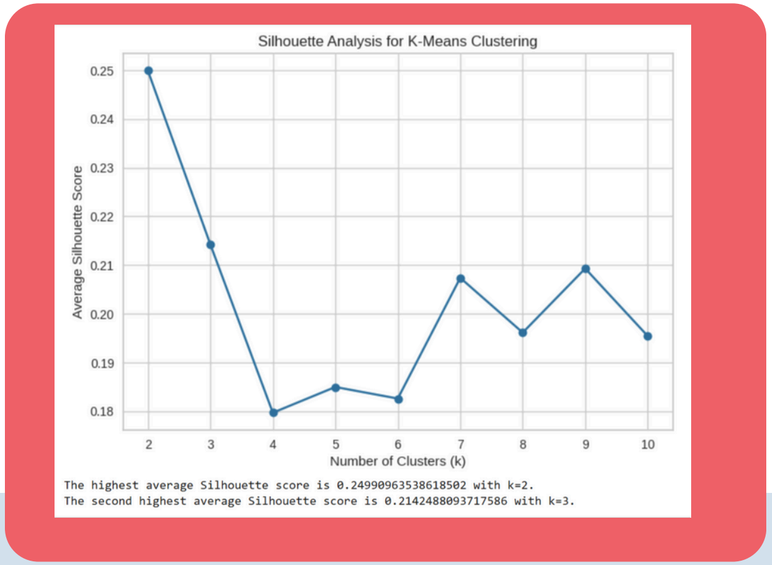
- We used the K-mean algorithm, which is an algorithm that produces K clusters, where each cluster is represented by the center point of the cluster and assigns each object to the nearest cluster.

- Then iteratively recalculates the center and reassigns the object until the center point of each cluster does not change, meaning the object is in the right cluster.

Evaluation of Clustering- Metrics

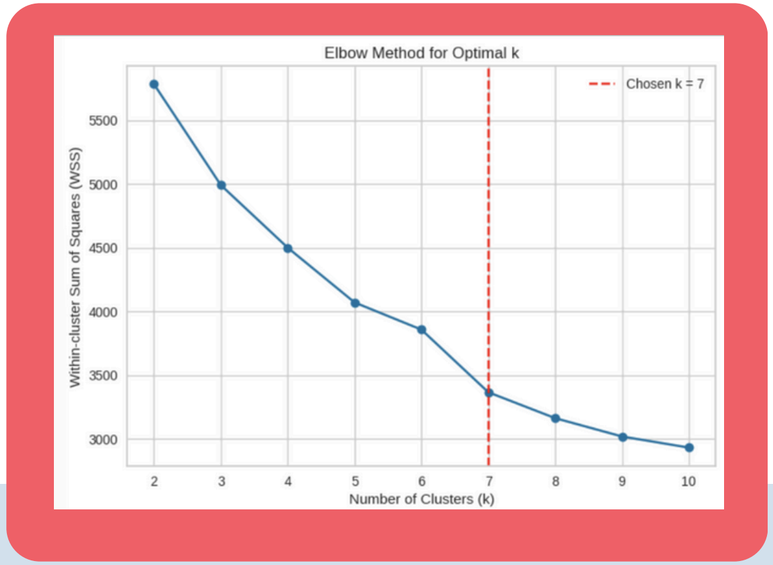
1.Silhouette Score:

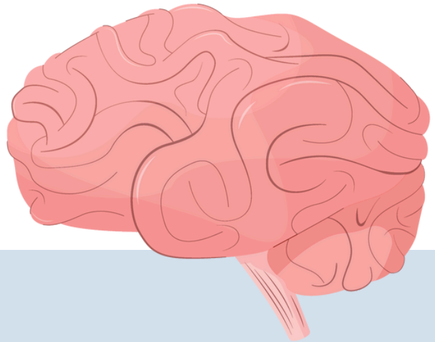
- Measures cluster compactness and separation.
- Higher score indicates well-defined clusters.
- Best silhouette score observed at K = 2.



2. Within-Cluster Sum of Squares (WSS):

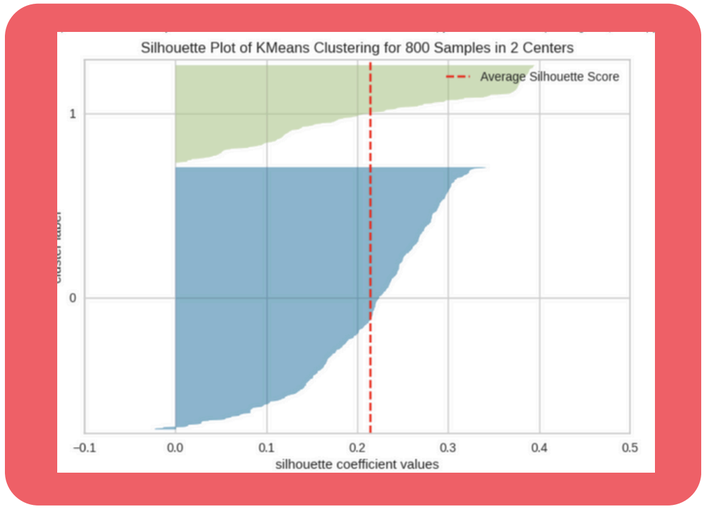
- Measures compactness of clusters.
- Decreases as K increases, balancing compactness and separation.

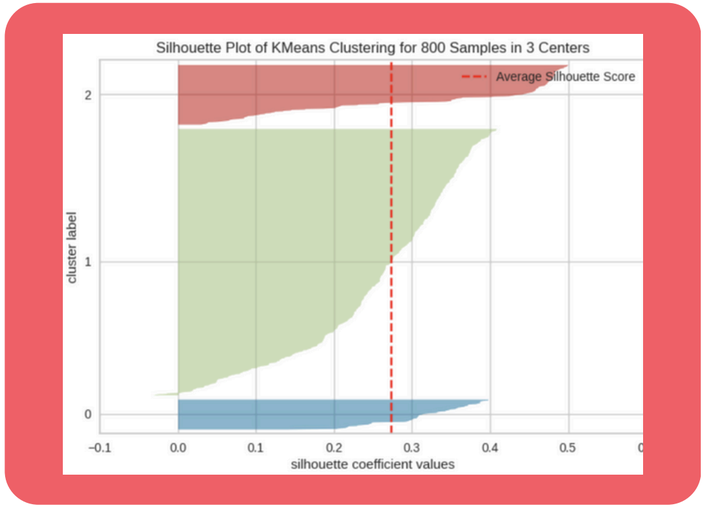


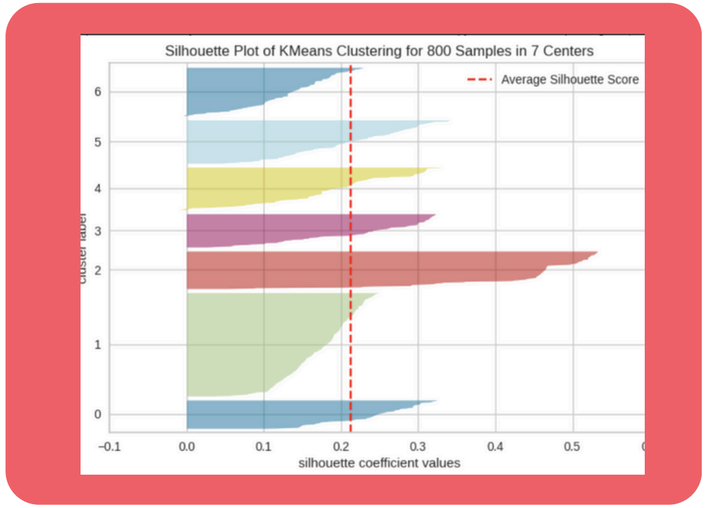


Evaluation of Clustering- Comparison

After comparing K=2, K=3, and K=7 and calculating the average silhouette width for each k



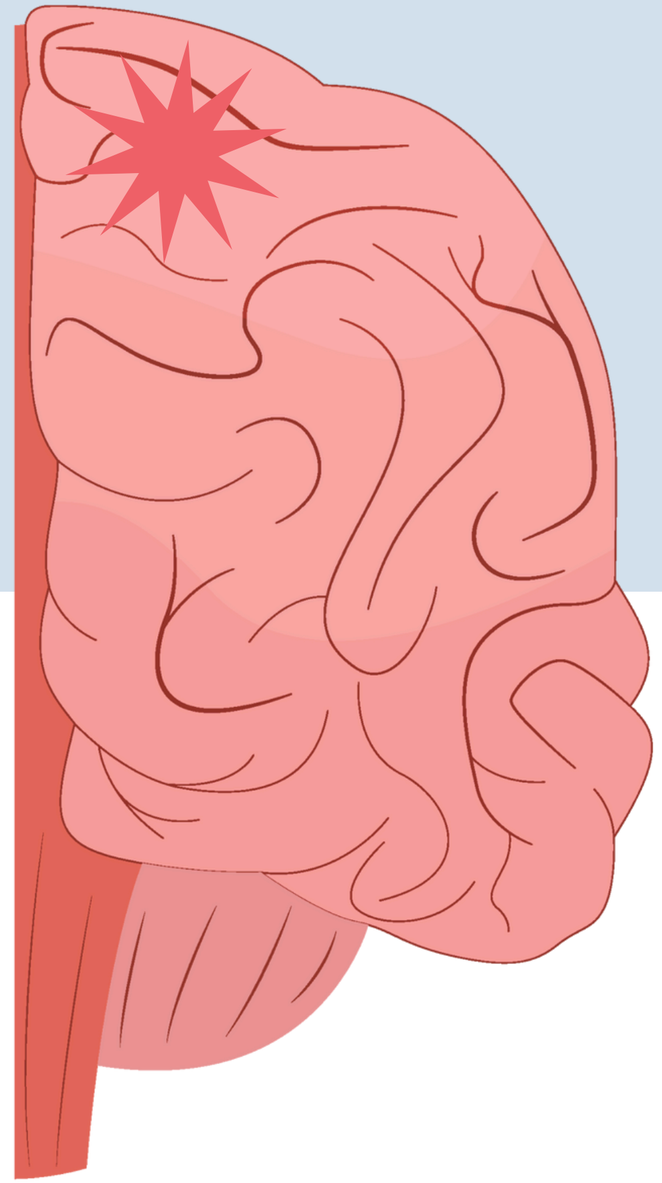




The model that has the optimal number of clusters is 2-Mean since it has the least overlapping between clusters compared to the other models.

05

Conclusions

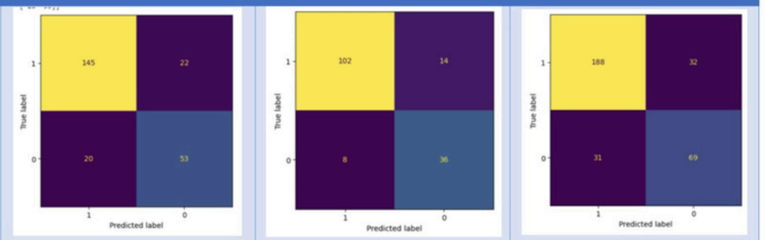
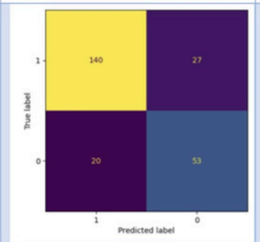
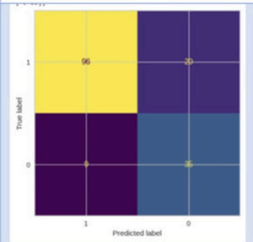
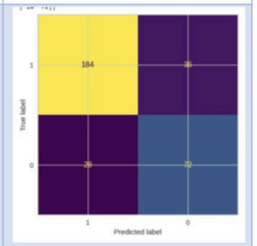


Key Findings and Insights

1. **Evaluation of Results**

• The Decision Tree with the 70% Training, 30% Testing split using Information Gain showed the best accuracy and balanced metrics, making it the most suitable classification model.

• K-Means clustering with K=2 provided the optimal clustering results based on silhouette width and WSS, identifying two distinct groups.

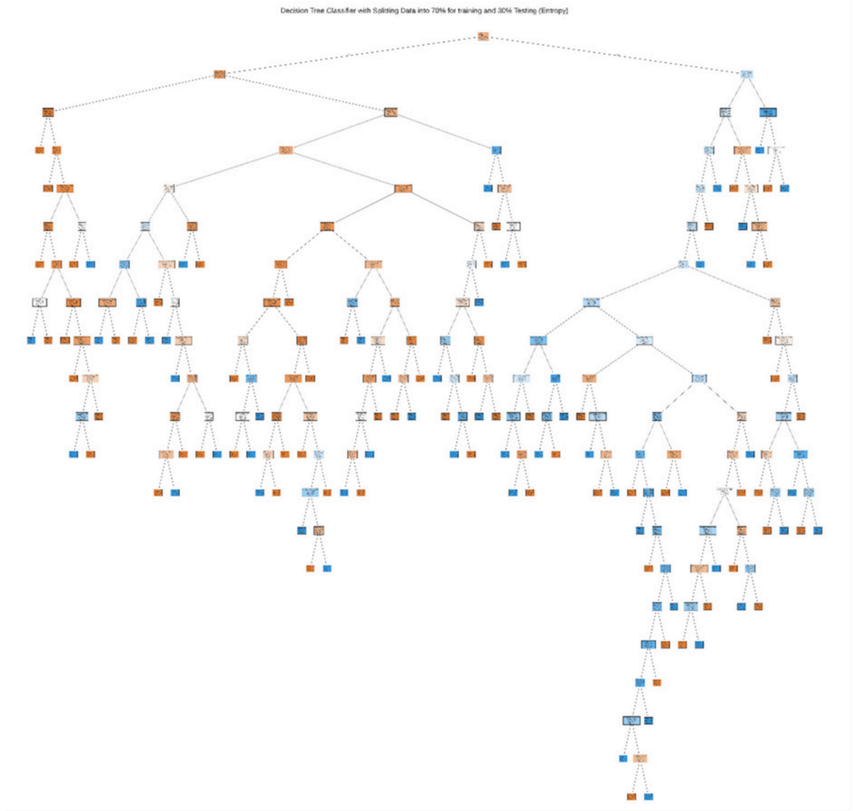
		70% training set 30% test set	80% training set 20% test set	60% training set 40% test set
Entropy				
	Accuracy	83%	80%	80%
Gini Index				
	Accuracy	80%	81%	80%

Number of clusters	K=2	K=3	K=7
Average Silhouette width	0.250	0.214	0.207
Total within-cluster sum of square	5781.88	4990.2	3365.38

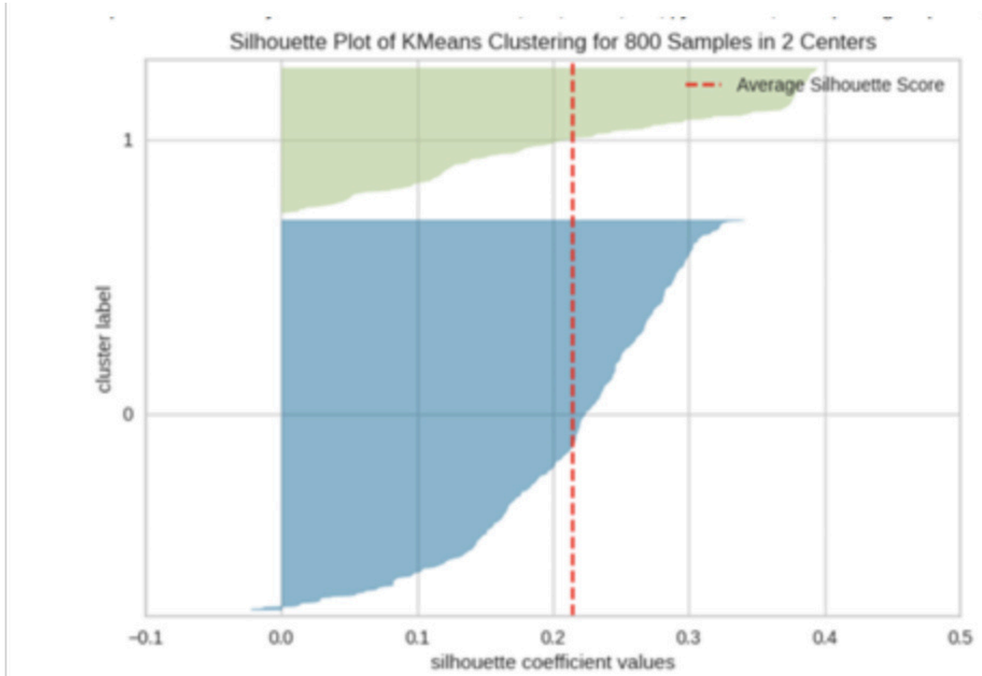
Key Findings and Insights

2. **Best Model**

• Classification: The Decision Tree effectively identified key features such as age, BMI, and glucose levels, allowing accurate stroke risk predictions.



• Clustering: The two clusters revealed shared characteristics but lacked the precision of supervised learning for stroke prediction.



Key Findings and Insights

3. **Insights and Problem Solutions**

• **Classification** proved more effective because of labeled data (“stroke”), enabling actionable predictions to aid in stroke prevention.

• **Clustering** provided insights into patterns and similarities but was less suitable for precise predictions.

4. **Interpretation**

• The **Decision Tree** structure highlighted interpretable paths to predict stroke risk using critical attributes.

• **K-Means clustering** results, visualized with silhouette plots, showed distinct clusters with some overlap.

Thanks!