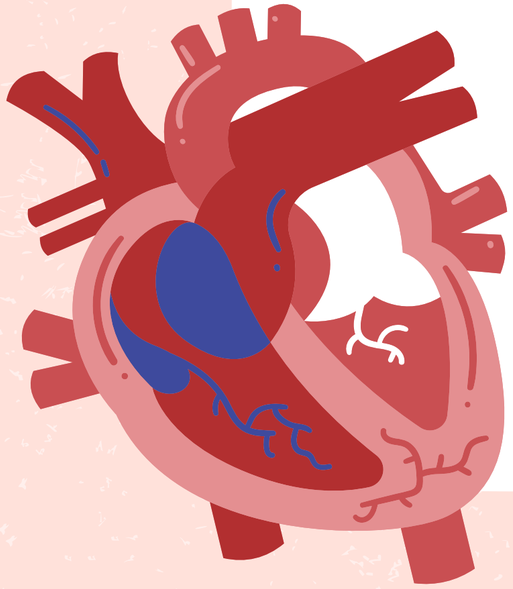
Heart Disease Predicting

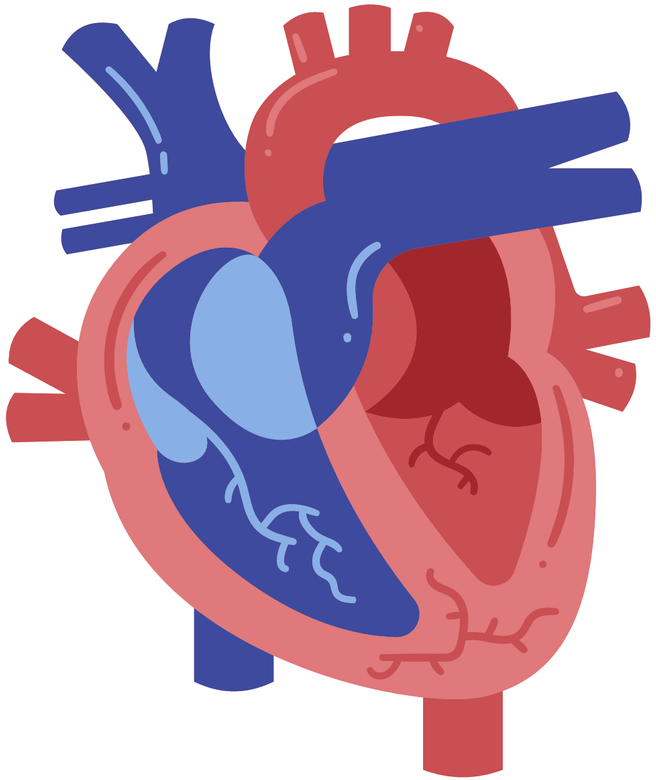


Supervised by: Dr.LAMA ALSUDIAS

Table of contents

1

Introduction



Introduction

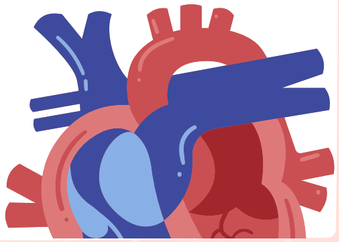
Problem:

Heart disease remains one of the leading causes of death worldwide, often due to delayed diagnosis or failure to recognize early symptoms. The Heart Disease Dataset contains detailed health-related information such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, and more. These features are important indicators that can help predict the presence or risk of heart disease.

Goal:

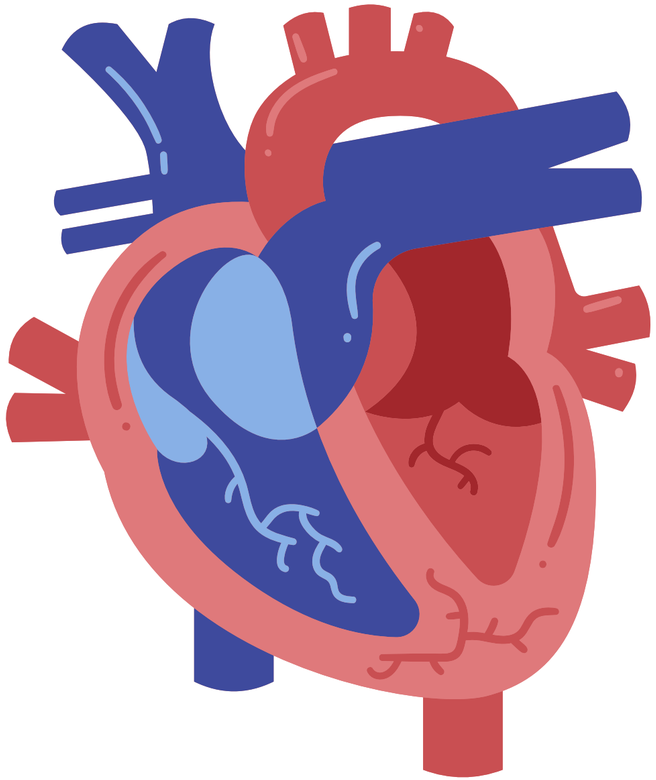
The goal is to analyze patient data to build a predictive model that accurately identifies individuals at risk of developing heart disease. Early detection is vital—it not only saves lives but also reduces long-term complications and eases the overall burden on healthcare systems.

By leveraging machine learning techniques, this project empowers individuals and healthcare professionals to make informed decisions, implement preventive strategies, and take timely medical action, ultimately leading to improved health outcomes and a better quality of life.



2

Dataset



Dataset

General Information about the dataset

Number of objects in original dataset: 1025

Number of attributes: 14

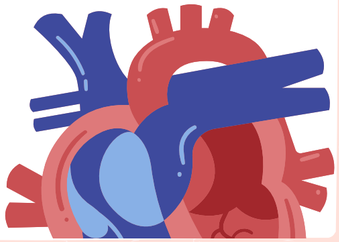
Class labels: Target

(1 = have heart disease, 0 = no heart disease)

Missing values: there is no missing values.

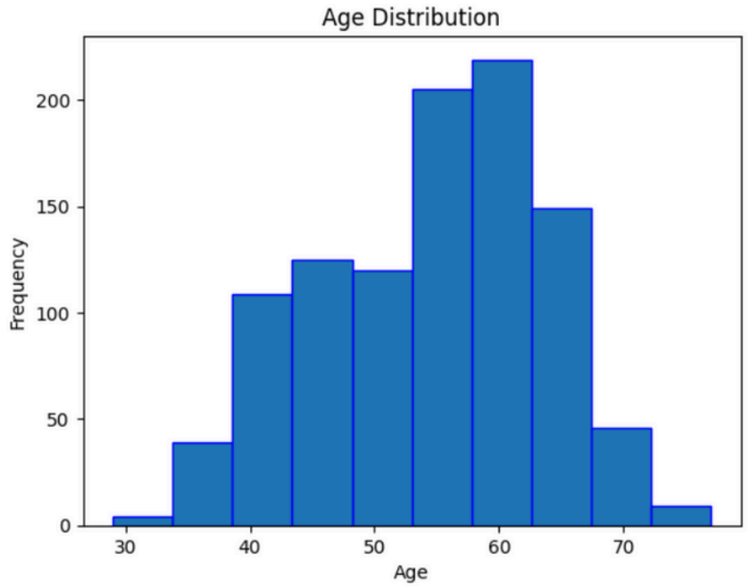
```
Missing values in each column:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64

Rows with missing values:
0       0
1       0
2       0
3       0
4       0
      ..
1020    0
1021    0
1022    0
1023    0
1024    0
Length: 1025, dtype: int64
```



Dataset

General Information about the dataset



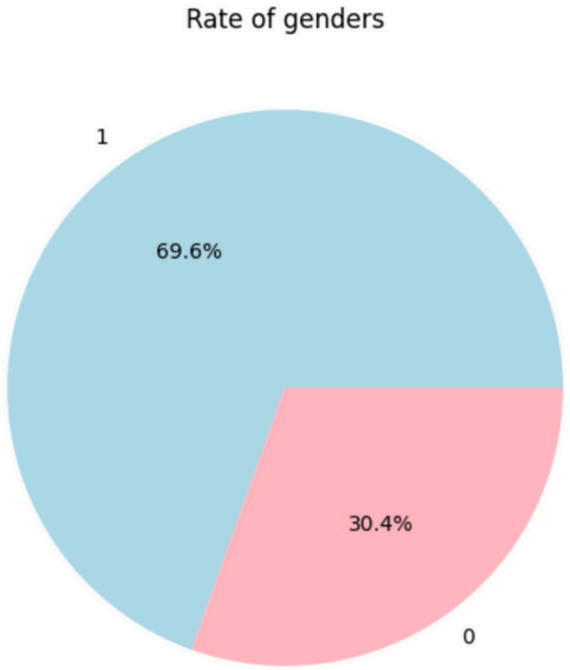
Number of objects in sampled dataset: 20.

Majority Age Group: Most individuals are aged between 50 and 60 years, indicating a focus on middle-aged adults.

Range: The ages range from approximately 30 to over 70, showing a diverse study population.

Dataset

General Information about the dataset

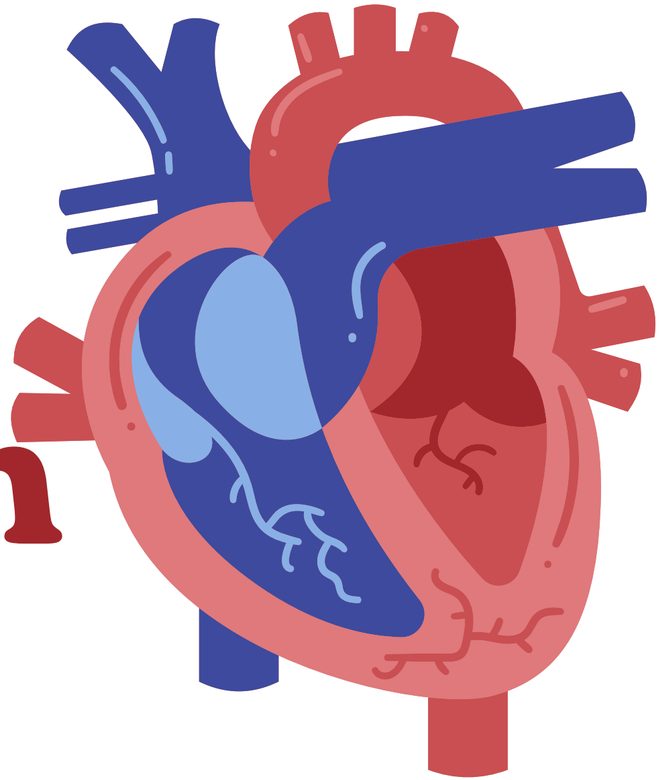


Number of objects in sampled dataset: 20.

We used a pie chart to illustrate the percentage distribution of each gender in the total data set , this results in 69.6% of men and 30.4% of female.

3

Classification

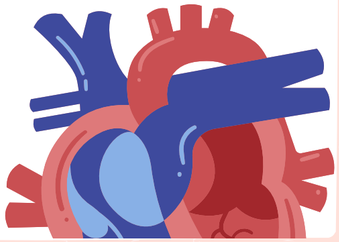


Classification

What is 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 an individual is at risk of heart disease or not based on medical and lifestyle features such as age, cholesterol, blood pressure, heart rate, and diabetes status. This classification allows us to identify individuals at risk and provide targeted interventions to improve health outcomes.



Classification

In our case, we applied the technique of splitting 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.

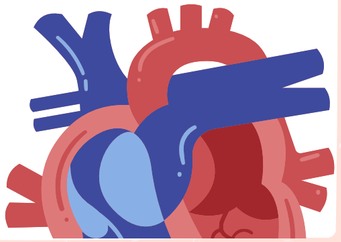
In our case, we applied the Gini Index and Information Gain as splitting criteria for our model, experimenting with different dataset splits to optimize performance:

Gini Index:

- 80% Training - 20% Testing
- 70% Training - 30% Testing
- 50% Training - 50% Testing

Information Gain:

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



Evaluation of Classification

information gain

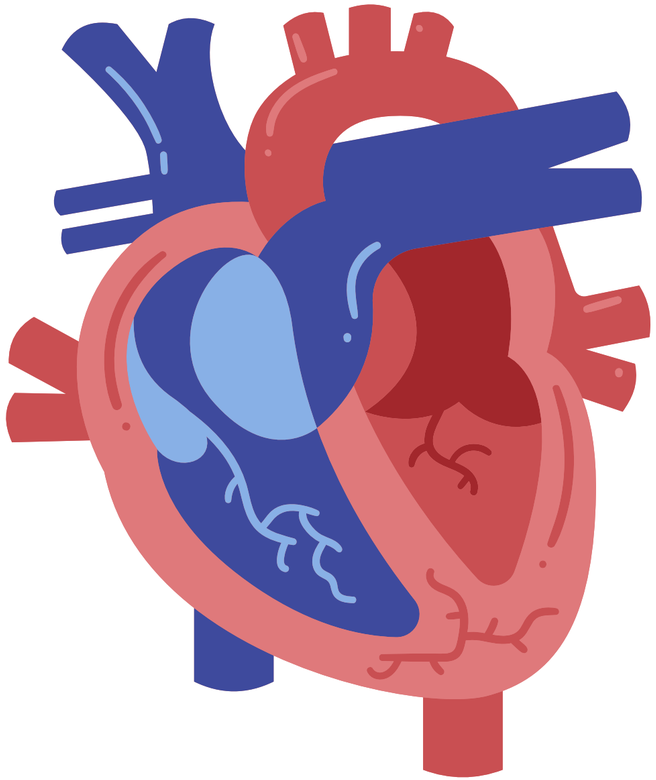
	Split	60% training, 40% testing	70% training, 30% testing	80% training, 20% testing
0	Accuracy	0.9658	0.9772	1.0
1	Error Rate	0.03414	0.0227	0.0
2	Sensitivity	0.9695	0.9795	1.0
3	Specificity	0.9624	0.9751	1.0
4	Precision	0.9597	0.9729	1.0

Gini Index

	Split	50% training, 50% testing	70% training, 30% testing	20% training, 80% testing
0	Accuracy	0.9045	0.8766	0.7988
1	Error Rate	0.0955	0.1234	0.2012
2	Sensitivity	0.9582	0.9388	0.8789
3	Specificity	0.8480	0.8199	0.7143
4	Precision	0.8690	0.8263	0.7645

4

Clustering



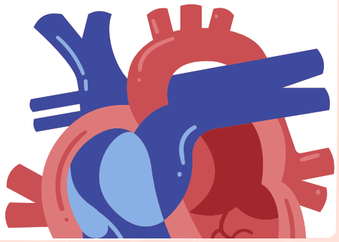
Clustering

What is Clustering ?

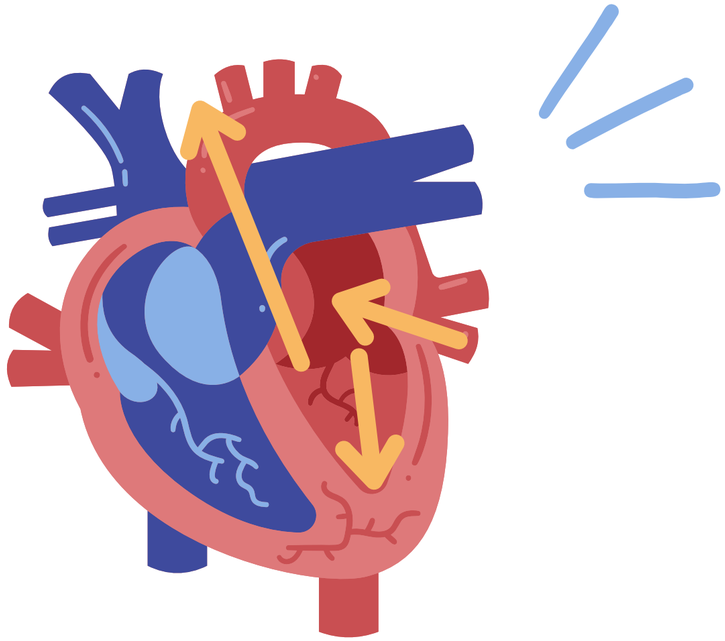
Clustering is unsupervised learning;
it will group objects in a cluster based on similarity and dissimilarity.

Goal:

Our model uses K-means clustering to group individuals with similar health characteristics. These clusters help reveal hidden patterns in the data and can provide valuable insights into factors associated with heart disease risk.



K means - Clustering



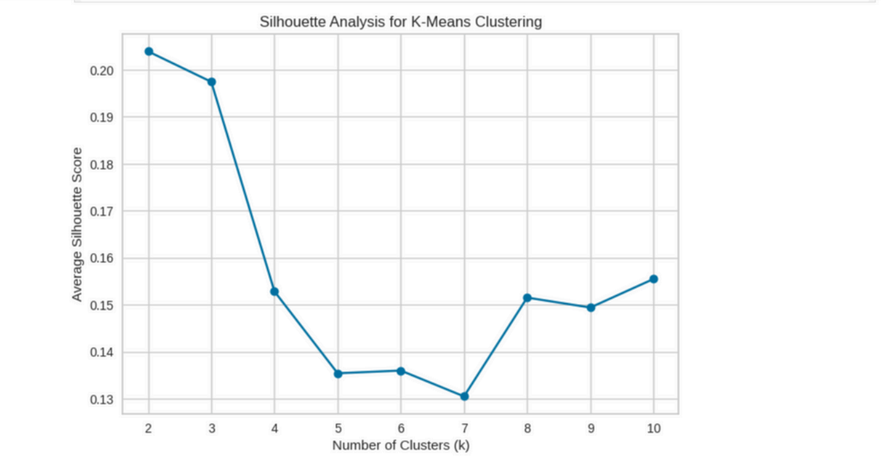
We applied the K-means algorithm, which partitions the data into K clusters. Each cluster is represented by a central point, and each data point is assigned to the nearest cluster.

The algorithm then iteratively recalculates the center of each cluster and reassigns the data points until the cluster centers no longer change, indicating that the points are correctly grouped.

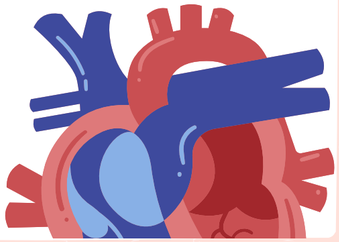
Evaluation of Clustering

Silhouette method

- evaluates the quality of clustering by measuring how well each data point fits within its assigned cluster compared to other clusters.
- higher values indicate more well-defined and distinct clusters.



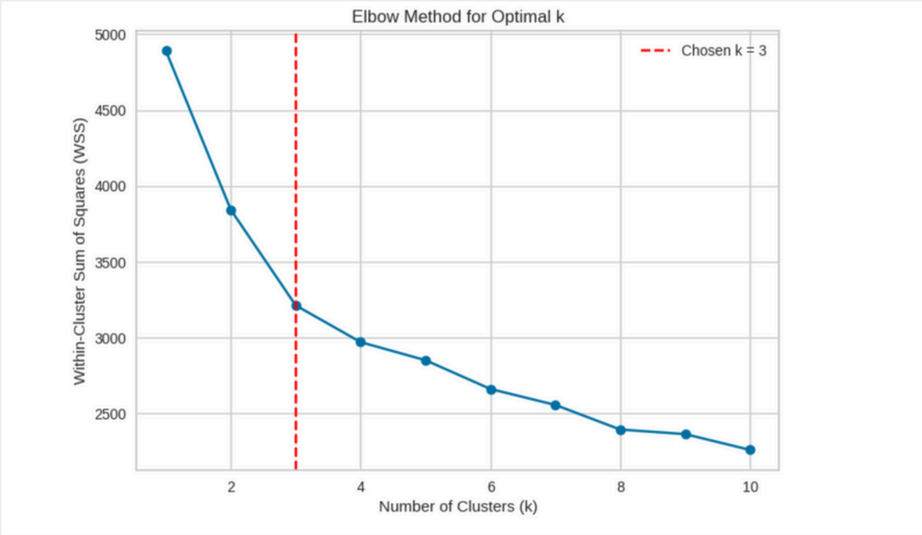
As observed above, the highest average Silhouette score is **0.2039** for **(K = 2)**, indicating that **(K = 2)** is the most optimal choice for our K-means clustering.



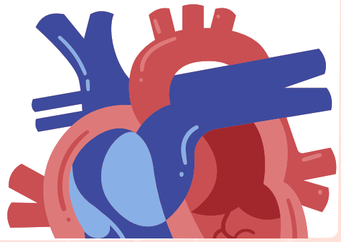
Evaluation of Clustering

Elbow method

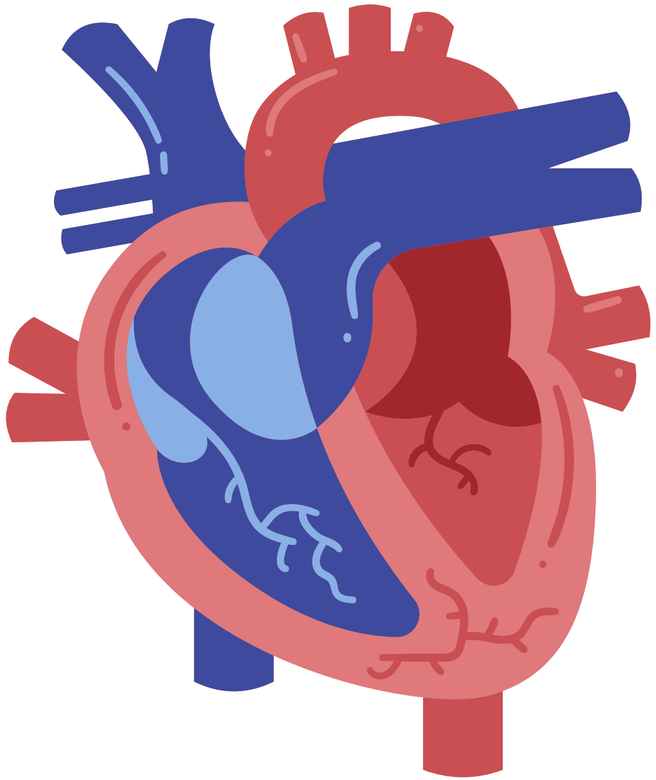
- popular approach for identifying the ideal number of clusters in K-means clustering.



We noticed that the elbow point appears at K = 3, where the drop in the WSS starts to slow down. This means that adding more clusters after K = 3 doesn't make a big difference in improving how tight the clusters are.



5 Findings and Insights



1- Classification

70% training and 30% testing

accuracy 97%

20% training and 80% testing

accuracy 79%

50% training and 50% testing

accuracy 90%

GINI Index

70% training and 30% testing

accuracy 97%

20% training and 80% testing

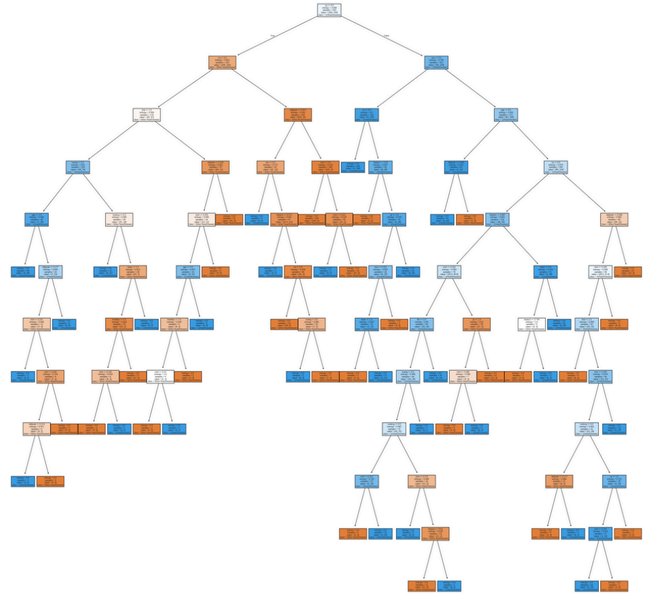
accuracy 100%

60% training and 40% testing

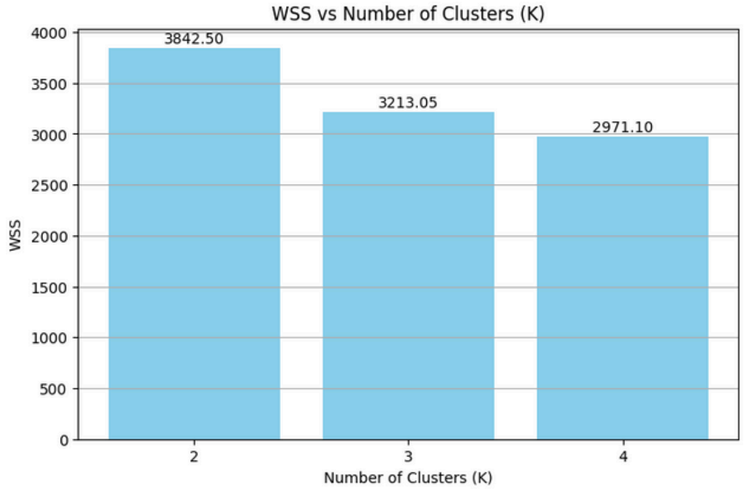
accuracy 96%

INFORMATION GAIN

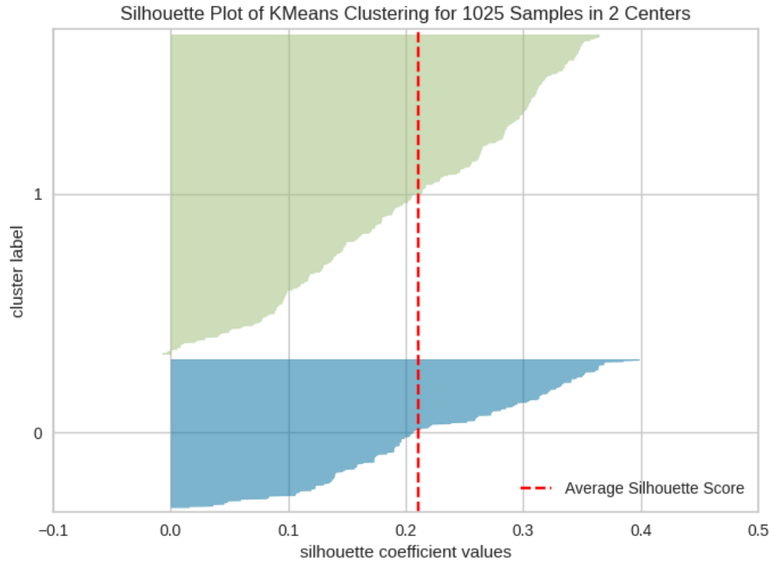
- The Decision Tree with the 80% Training, 20% Testing split using Information Gain showed the perfect accuracy and balanced metrics, making it the most suitable classification model.



2- Clustering



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



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

conclusion

- The Information Gain Decision Tree (20-80 split) is the most reliable model for heart disease prediction balancing high accuracy with generalizability.

- The K=2 clustering configuration offers a meaningful division of the dataset into groups, potentially useful for further profiling of patient types.

To gain a comprehensive understanding of heart disease prediction, both classification and clustering should be useed. By using both methods, you can achieve precise predictions while uncovering deeper patterns in the data.

Thanks !