Breast Cancer Data Mining

Outline

01 Problem

02 Dataset

03 Data Preprocessing

04 Data Mining Techniques

05 Results and Findings

01 Problem

In the world

1 in 4 new cancer cases is a breast cancer[1].

Kingdom of Saudi Arabia

19.8% of all cancer cases detected in the Kingdom[2].

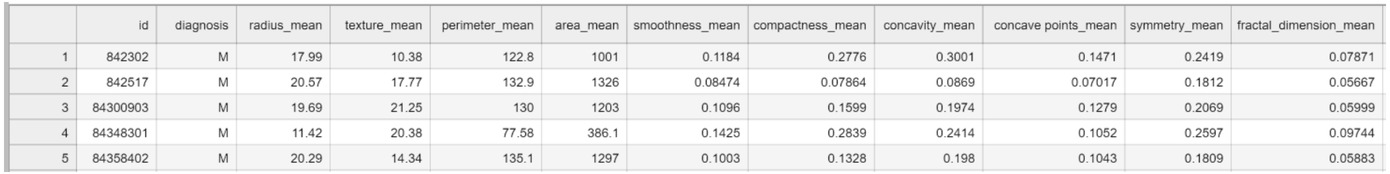
02 Dataset[3]

our dataset (breast-cancer) consist of:

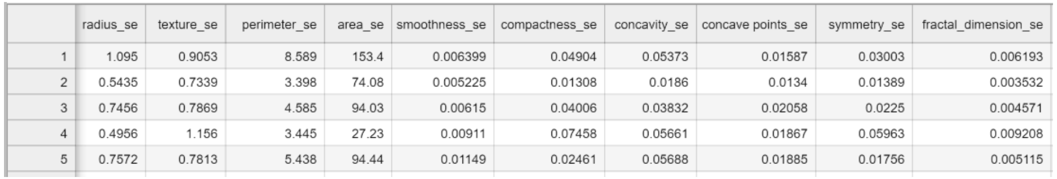
- 569 objects
- 32 attributes
- Class label is : diagnosis

02 Dataset[3]

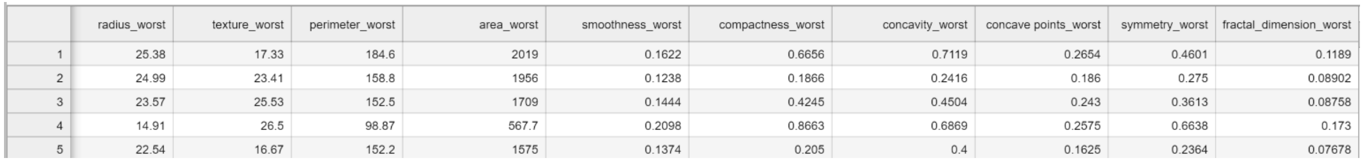
dataset Attributes:



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
1	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
2	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
3	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
4	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
5	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883

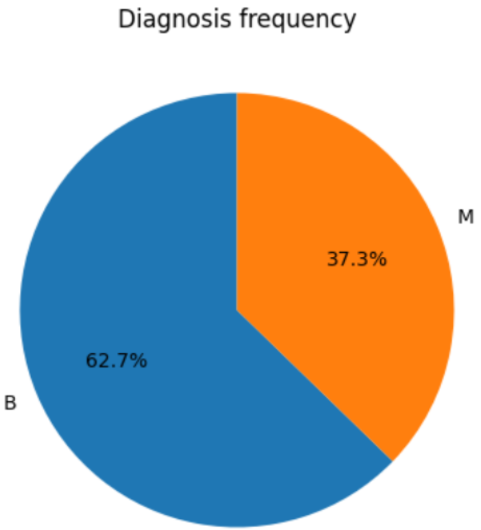


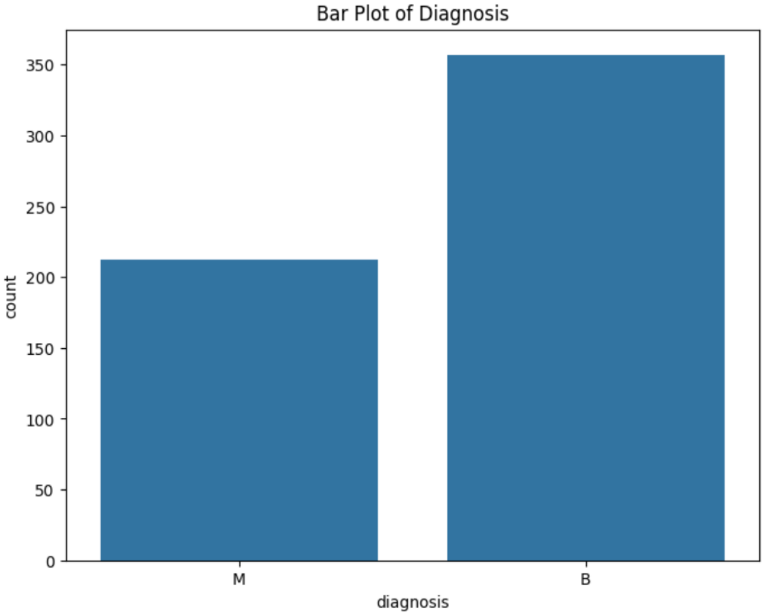
	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se
1	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193
2	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532
3	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571
4	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208
5	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115



	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
1	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189
2	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275	0.08902
3	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613	0.08758
4	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173
5	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	0.07678

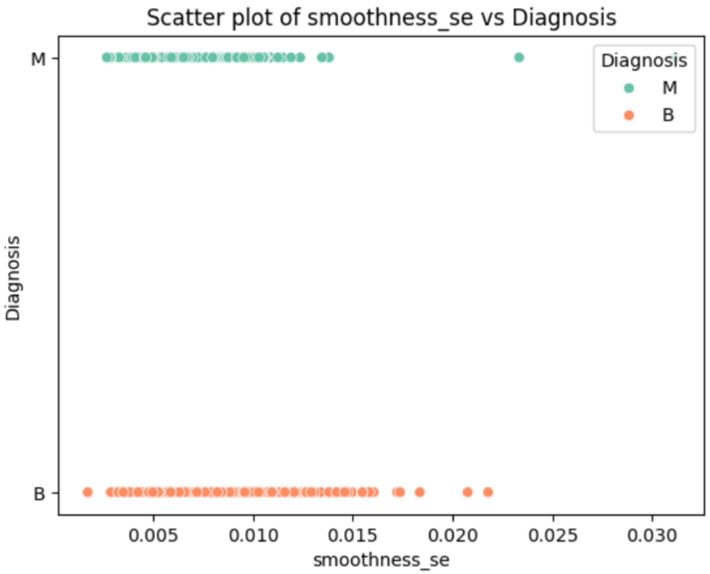
03 Data Graphs

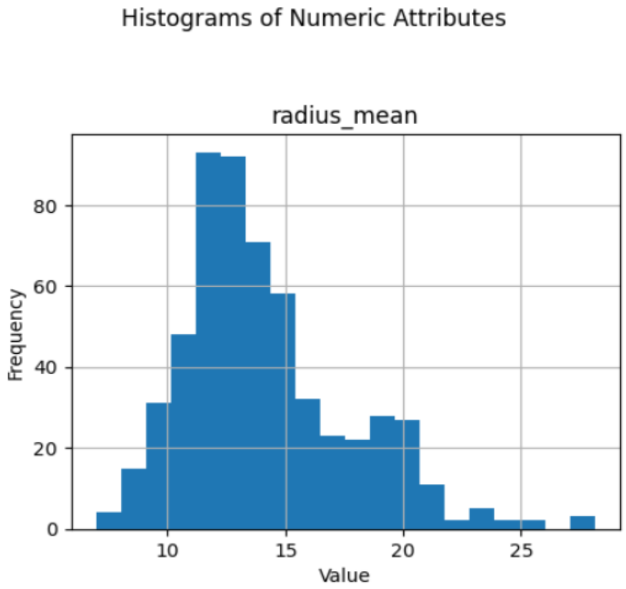




03 Data Graphs

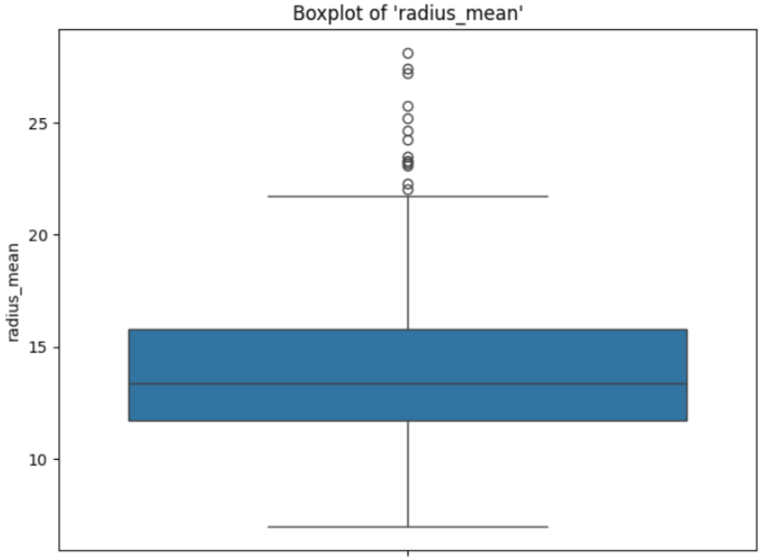
Data Graphs of some Attributes





03 Data Graphs

Data Graphs of some Attributes



03 Data Preprocessing

To make our data accurate and reliable and easy to use for analysis or machine learning, we used the following techniques:

Data Cleaning

the process of identifying and correcting errors and inconsistencies in a dataset.

Data Transformation

the process of changing the format, or content of data to make it more suitable for analysis or modeling.

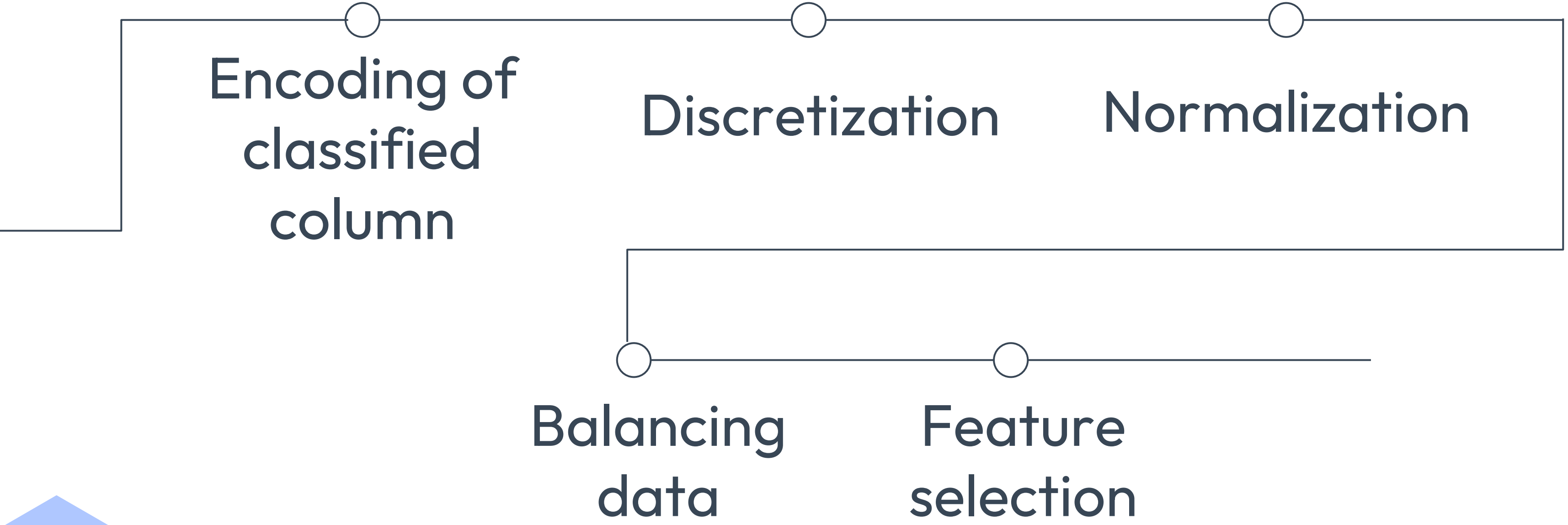
Data Cleaning

Checking missing value

checking duplicated rows

Removing outliers

Data Transformation



Encoding of classified column

Discretization

Normalization

Balancing data

Feature selection

04 Data Mining Techniques

Classification

Aplay supervised learning to detect the cancer stage.

Clustering

Aplay unsupervised learning to group patients.

Classification

- we used a decision tree which is a recursive algorithm produces a tree with a leaf nodes representing the final decisions.

- This technique includes dividing the dataset into Training dataset which Used for building the decision tree, and Testing dataset which Used to evaluate the constructed model.

- We tried 3 different sizes of testing size to get the best result

Classification-Gini index

splitting the dataset into training and testing sets

30%
test split

Accuracy
0.9056

20%
test split

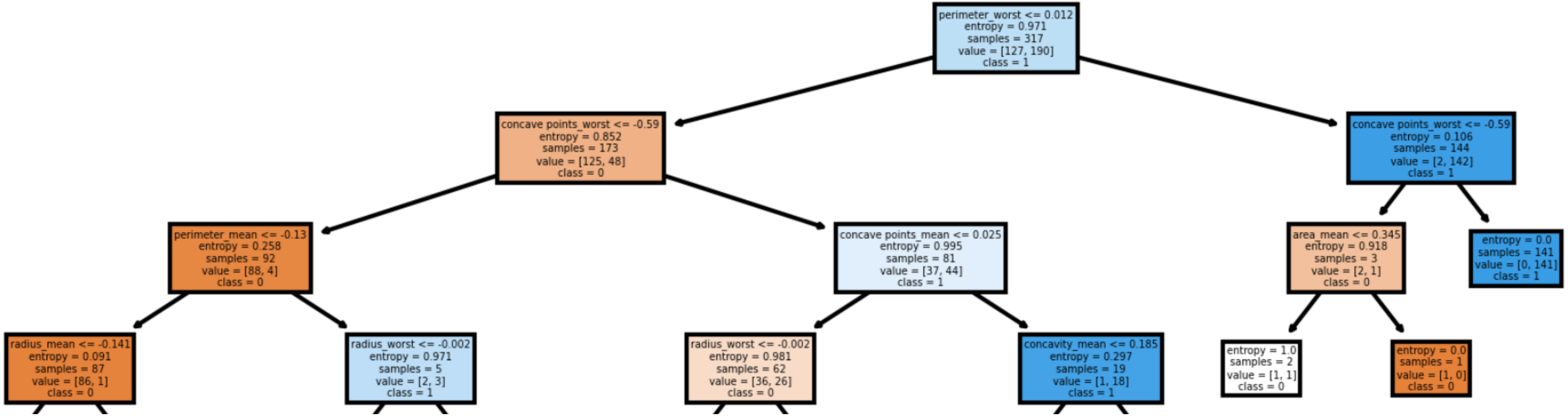
Accuracy
0.887

10%
test split

Accuracy
0.861

Classification

Illustration of the tree



Classification-IG(entropy)

splitting the dataset into training and testing sets

30%

test split

Accuracy

0.9056

20%

test split

Accuracy

0.887

10%

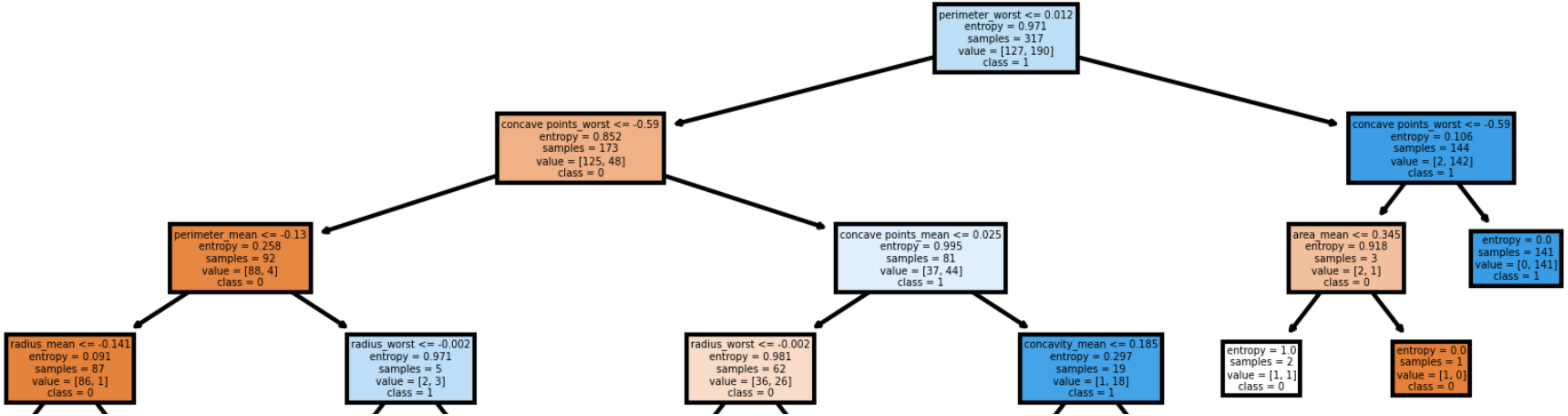
test split

Accuracy

0.861

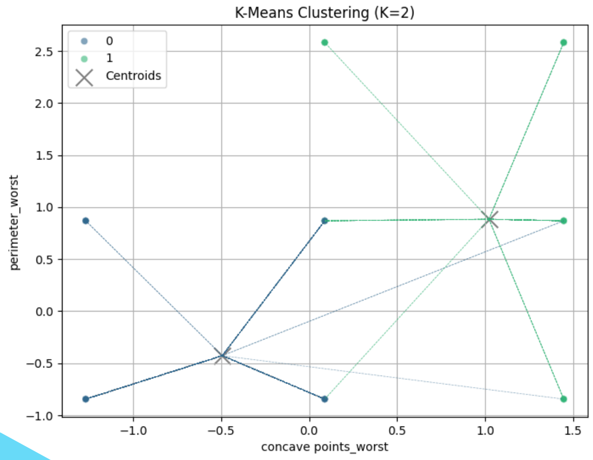
Classification

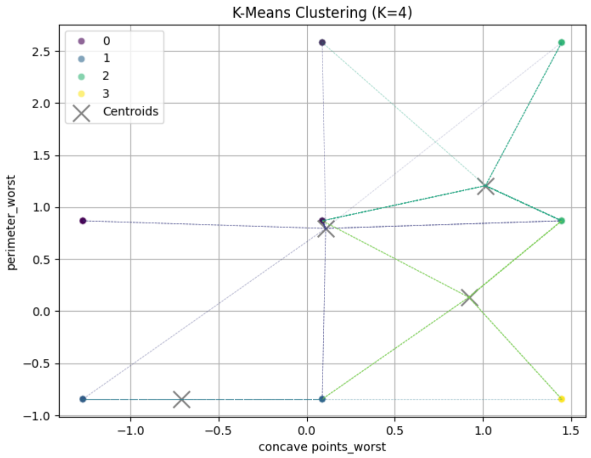
Illustration of the tree

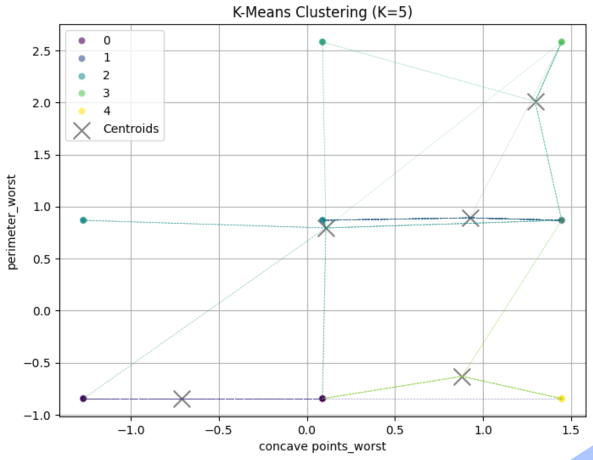


Clustering

1.Applying K-mean clustering



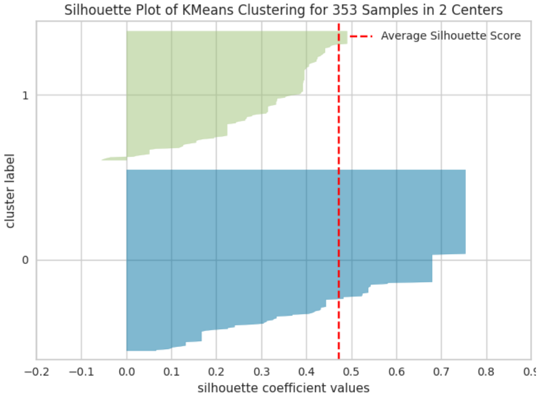


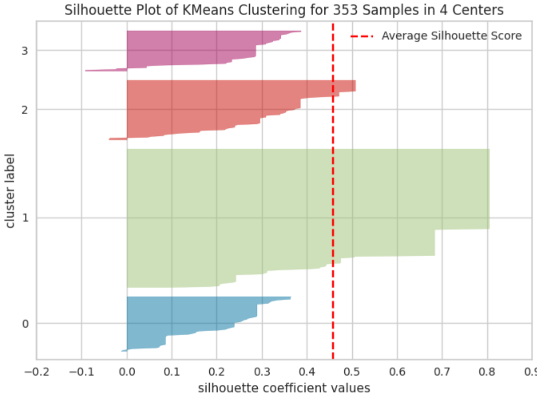


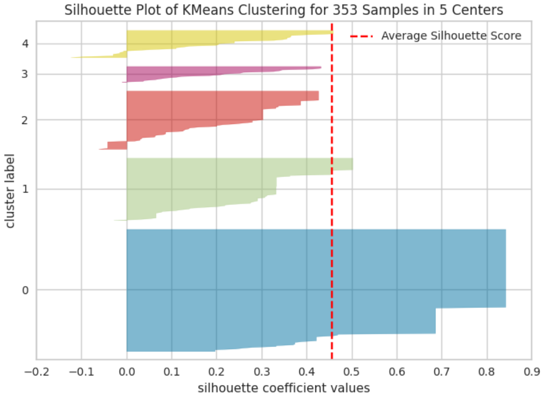
Clustering

2.determining Optimal K

a)Using Silhouette coefficient



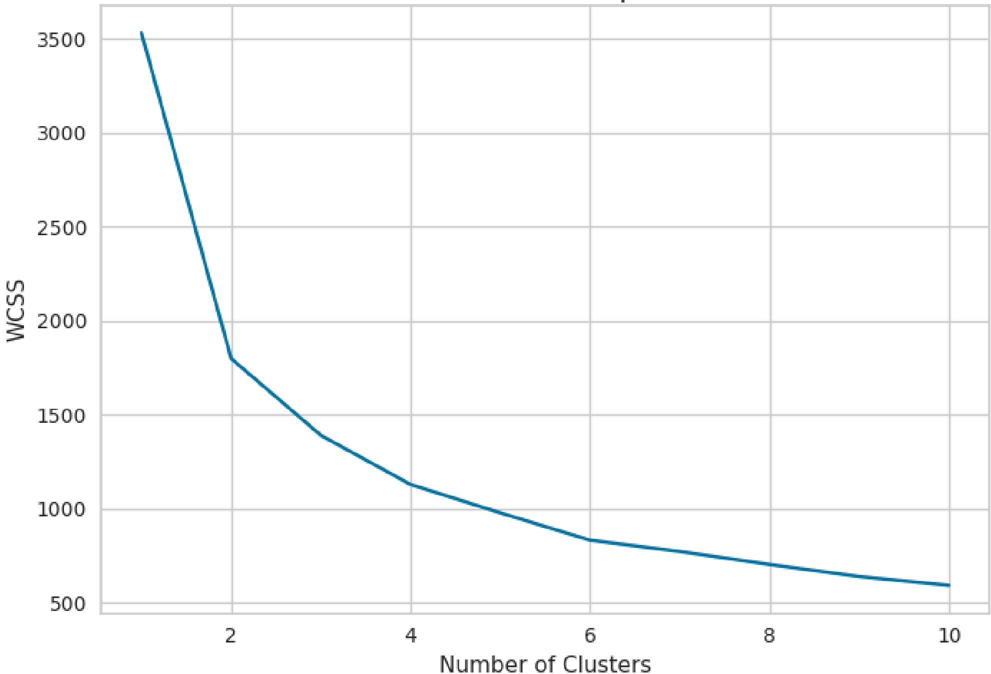




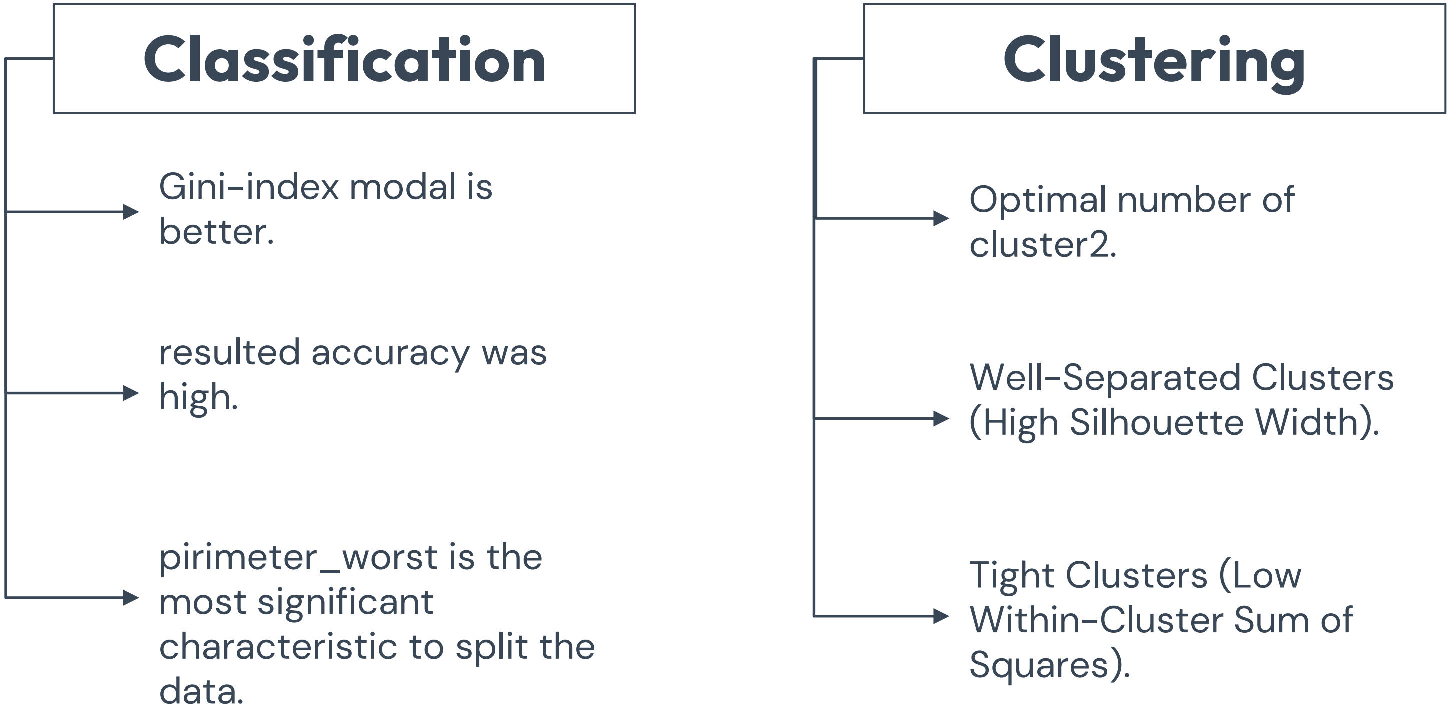
Clustering

2.determining Optimal K

b)Using elbow method (total within cluster sum of square)



05 Results and Findings



Classification

- Gini-index modal is better.
- resulted accuracy was high.
- pirimeter_worst is the most significant characteristic to split the data.

Clustering

- Optimal number of cluster2.
- Well-Separated Clusters (High Silhouette Width).
- Tight Clusters (Low Within-Cluster Sum of Squares).

Thanks!

Do you have any questions?

Prepared by:

Tarfah Al Ateeq	443200800
Doaa abdul hakim	443203882

Supervised by:
Dr.Sharefah A. Al-Ghamdi

github link

Resources

- [1] Breast cancer: Global patterns of incidence, mortality, and …, https://ascopubs.org/doi/10.1200/JCO.2023.41.16_suppl.

- [2]B. AlRajhi et al., “Breast cancer awareness among women in Saudi Arabia: A systematic review,” Breast cancer (Dove Medical Press), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10726713/.

- [3]Learning, U. M. (2016, September 25). Breast cancer wisconsin (diagnostic) data set. Kaggle. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data