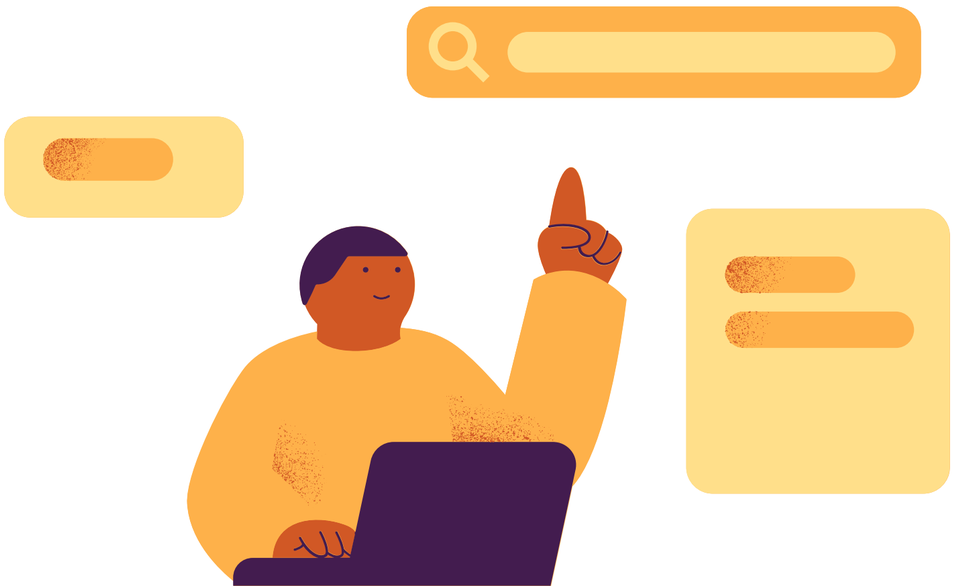
Data mining

BREAST CANCER WISCONSIN MODEL

TABLE OF CONTENT



01.

PROBLEM DESCREPTION

The problem is accurately diagnosing breast masses as malignant or benign using computed features .

accurate diagnosis of breast masses is important for early detection, personalized treatment planning, improved survival rates, enhanced patient care, and optimal allocation of healthcare resources

GENERAL INFORMATION ABOUT THE DATASET:

The task involves training models and algorithms on the features of breast masses to predict whether a specific mass is malignant(M) or benign(B).

the class label is: Diagnostic.

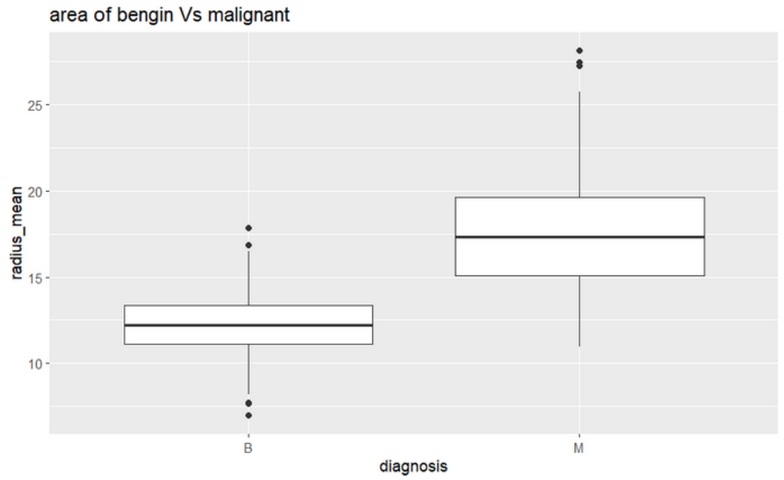
the data set contains: 32 columns and 569 records.

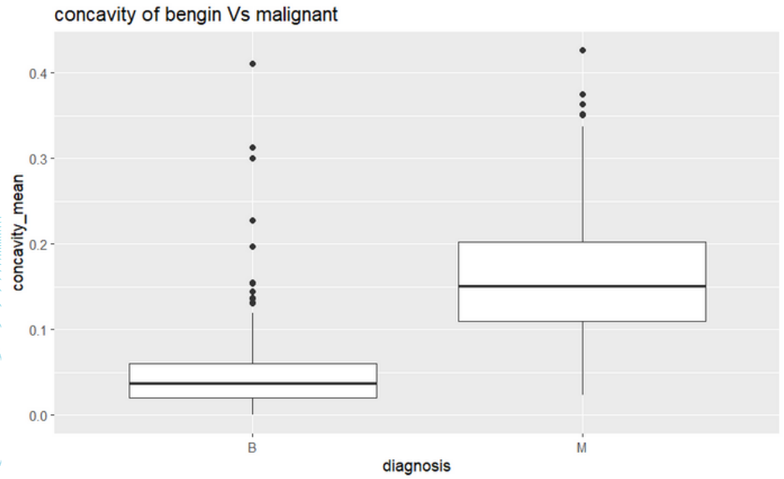
DATA SET:

1	id	diagnosis	radius_mea	texture_me	perimeter_	area_mean	smoothnes	compactne	concavity_	concave po	symmetry_	fractal_dim	radius_se	texture_se	perimeter_	area_se	smoothnes	compactne	concavity_s
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553

1	concave po	symmetry_	fractal_dim	radius_wor	texture_wc	perimeter_	area_worst	smoothnes	compactne	concavity_v	concave po	symmetry_	fractal_dimension_worst
2	0.01587	0.03003	0.006193	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189
3	0.0134	0.01389	0.003532	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275	0.08902
4	0.02058	0.0225	0.004571	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613	0.08758
5	0.01867	0.05963	0.009208	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173
6	0.01885	0.01756	0.005115	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	0.07678
7	0.01137	0.02165	0.005082	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985	0.1244
8	0.01039	0.01369	0.002179	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063	0.08368
9	0.01448	0.01486	0.005412	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196	0.1151
10	0.01226	0.02143	0.003749	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378	0.1072

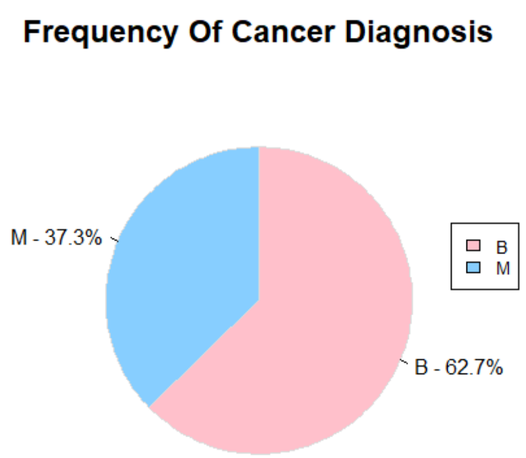
BOX PLOT:





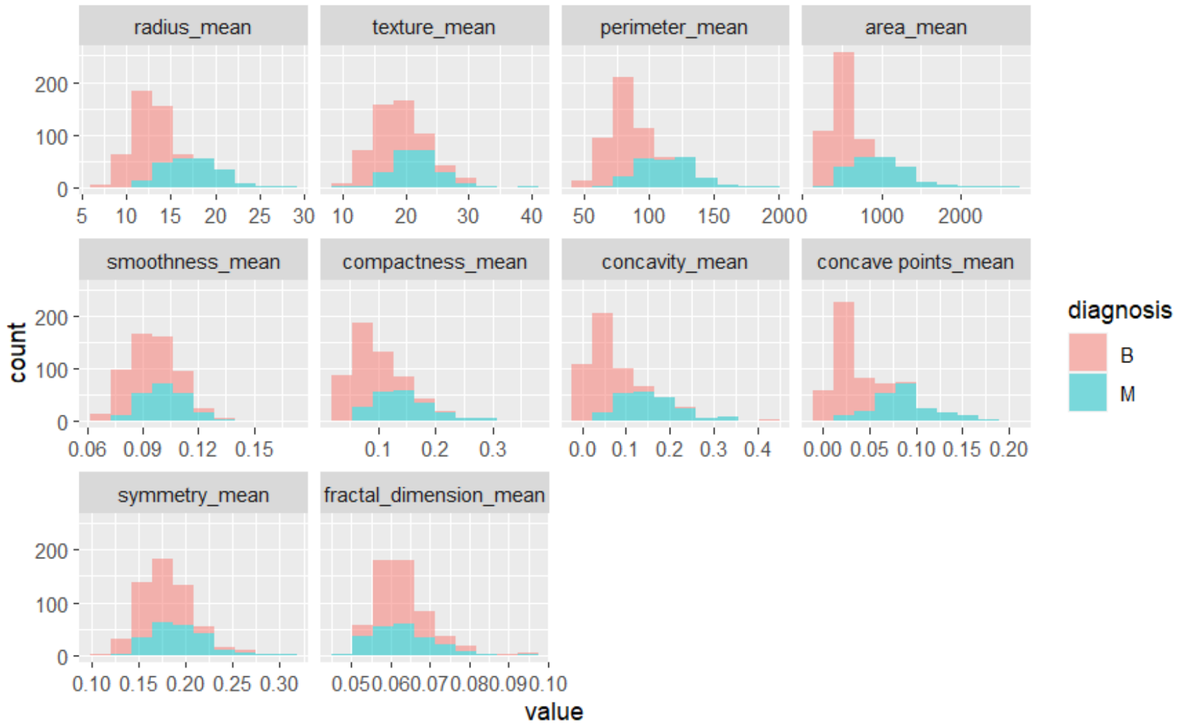
Comparing the raidus, area and concavity of benging malignant stage By constructing a box plot with the function ggplot()+geom_boxplot(), we noticed malignant cells have higher radius, area and concavity mean than benign cell which is an information we can use later for predicting and training the data.

PIE CHART



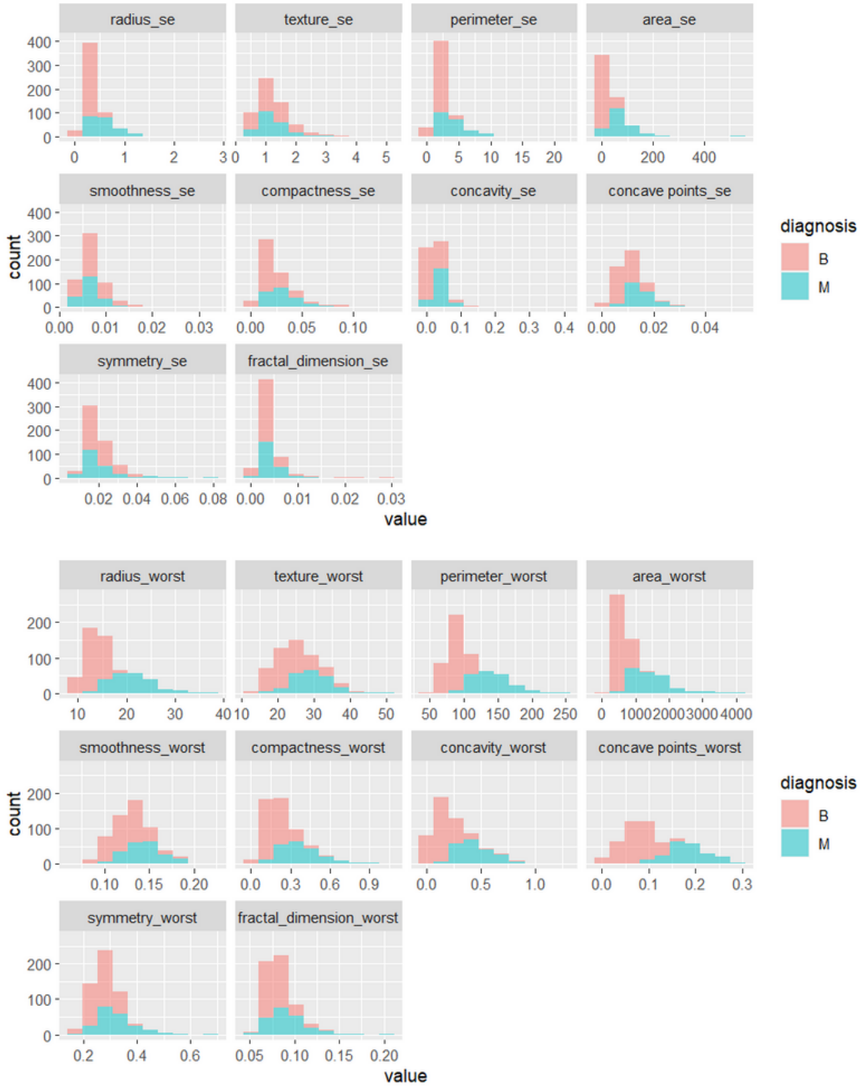
M= Malignant (cancerous); B= Benign (non-cancerous) 357 observations which account for 62.7% of all observations indicating the absence of cancer cells, 212 which account for 37.3% of all observations shows the presence of cancerous cell.

HISTOGRAMS OF DATA MEAN:



1. we can see that radius_mean of malignant tumors are bigger than radius_mean of benign tumors mostly.
2. The benign distribution (blue in graph) is approcimately bell-shaped that is shape of normal distribution. However the same is not true for the malignant class data.
3. Also the mean value of malignant is higher than that of benign class.

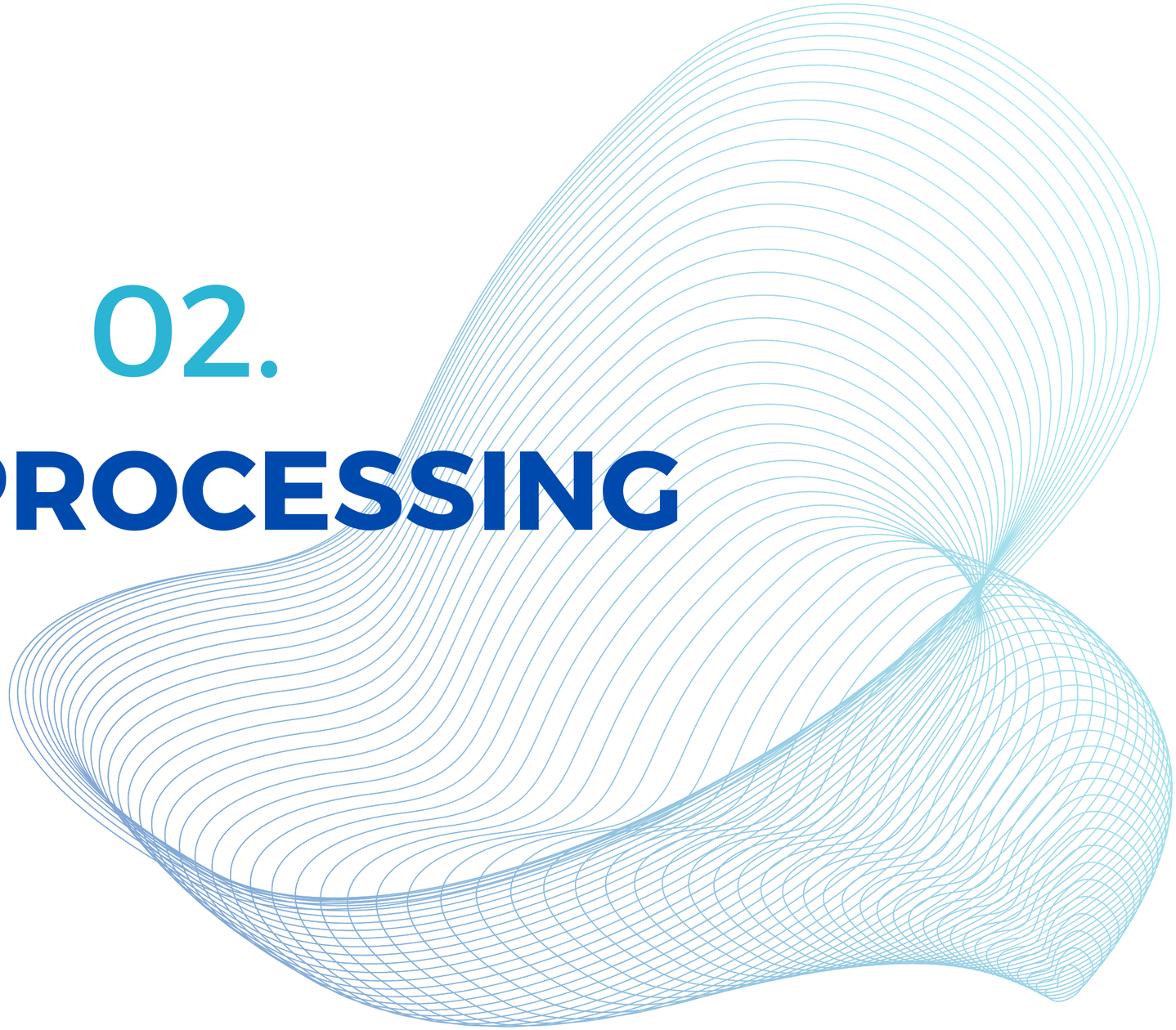
HISTOGRAMS OF DATA “_SE”,“_WORST”:



Most of the features are normally distributed. Comparison of radius distribution by malignancy shows that there is no perfect separation between any of the features; we do have fairly good separations for concave.points_worst, concavity_worst, perimeter_worst, area_mean, perimeter_mean.

02.

PRE-PROCESSING



DATA CLEANING

removing noise:

Data reduction using Dimensionality Reduction:

- “id” and column “Unnamed”

Removing Outliers using Z-score:

- every object that was an outlier more than five times will be deleted, therefore we deleted a total of 13 object.

DATA TRANSFORMATION

Attribute transformation

Encoding:

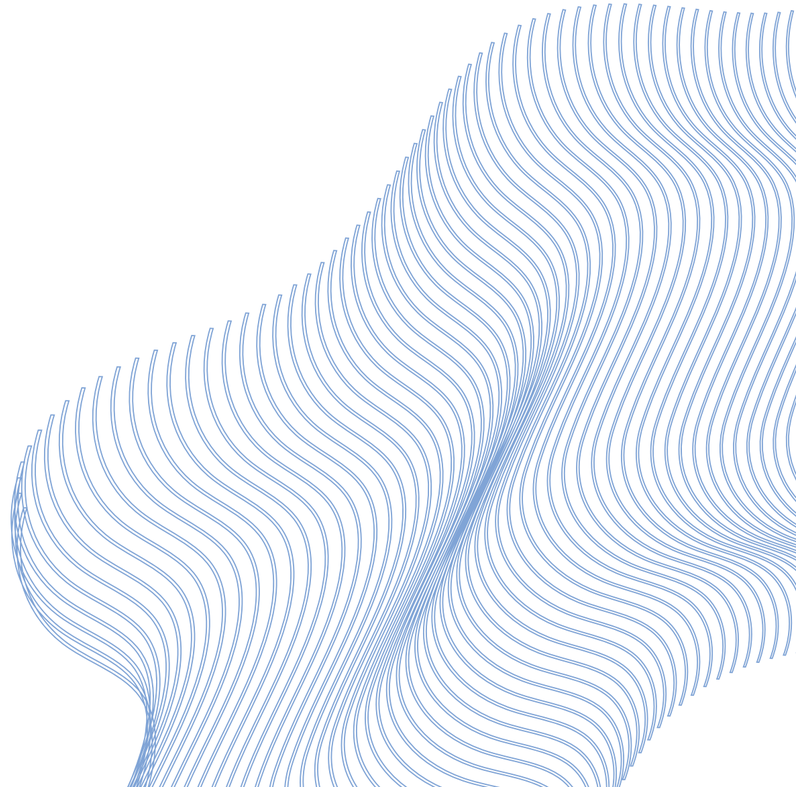
- B (benign) became 0
- M (malignant) became 1

Discretization:

- Dividing all numeric continues attributes using equal width with 20 bins.

Normalization:

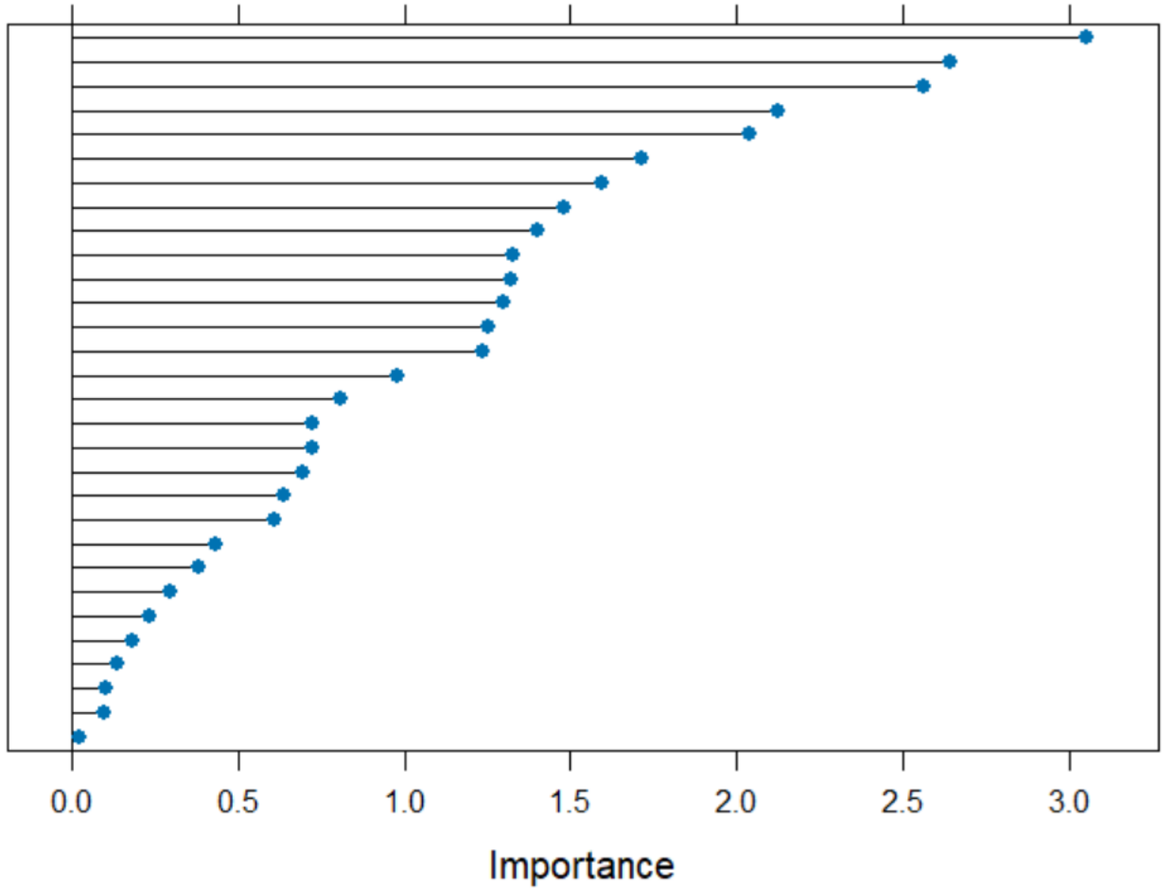
- applying the normalize() function to the selected columns to normalize them into 0 and 1.



FEATURE SELECTION

Attribute transformation

- calculates the importance of each attribute.
- The least important 9 attributes were identified by their importance being under 0.5.
- highest importance being 3.05 and belonging to “compactness_mean”.



03.

CLASSIFICATION

DECISION TREE

Definition: A tree-like model for classification tasks.

Structure:

- Nodes: Represent features or attributes.
- Branches: Decision rules based on feature values.
- Leaves: Outcome or class label.

Working Principle:

- Recursive splitting based on best features.

Decision Rules:

- Transparent and easy-to-understand.
- Interpretability is a key strength.

PARTITIONING METHOD

1 Holdout Partitioning

Split the dataset into two subsets: one for training the model and the other for evaluating its performance.

2 Holdout Features

a- simplicity.
b- Ease of implementation.
c- provide a quick assessment of a model's performance.

3 Partition splits

70-30, 60-40, and 80-20.

ATTRIBUTE SELECTING MEASURES

1 Information Gain

Algorithm: C50 algorithm
Function: The C5.0() function has Information Gain as the default attribute selecting measure.

2 Gini Index

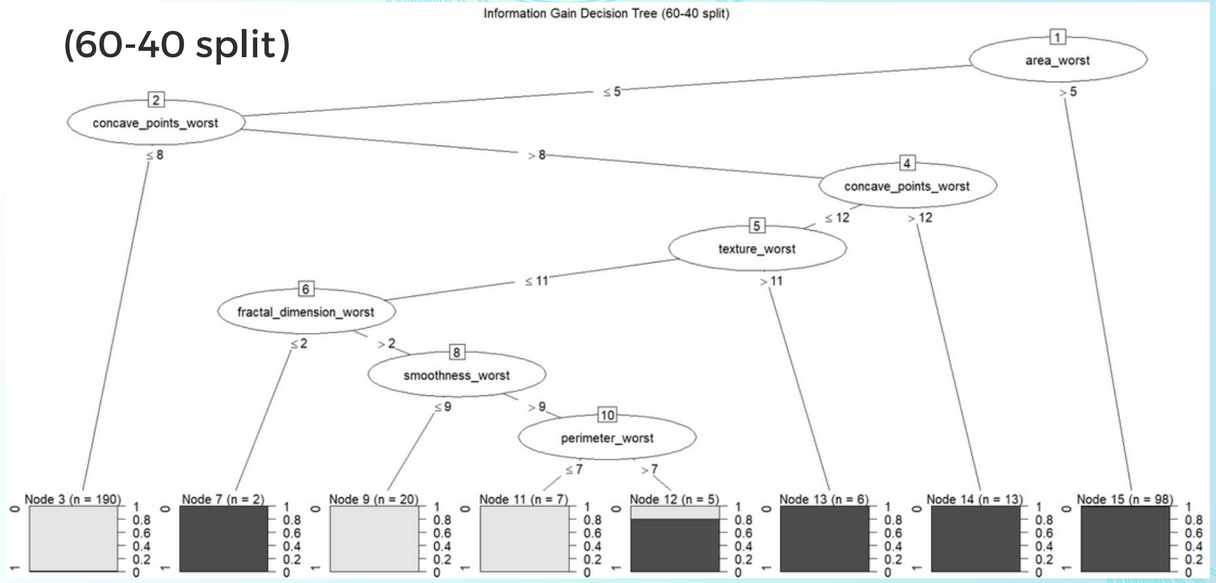
Algorithm: rpart algorithm
Function: The rpart function has a direct argument to specify Gini Index as Attribute selecting, using The parms argument with list(split = “gini”)

3 Gain Ratio

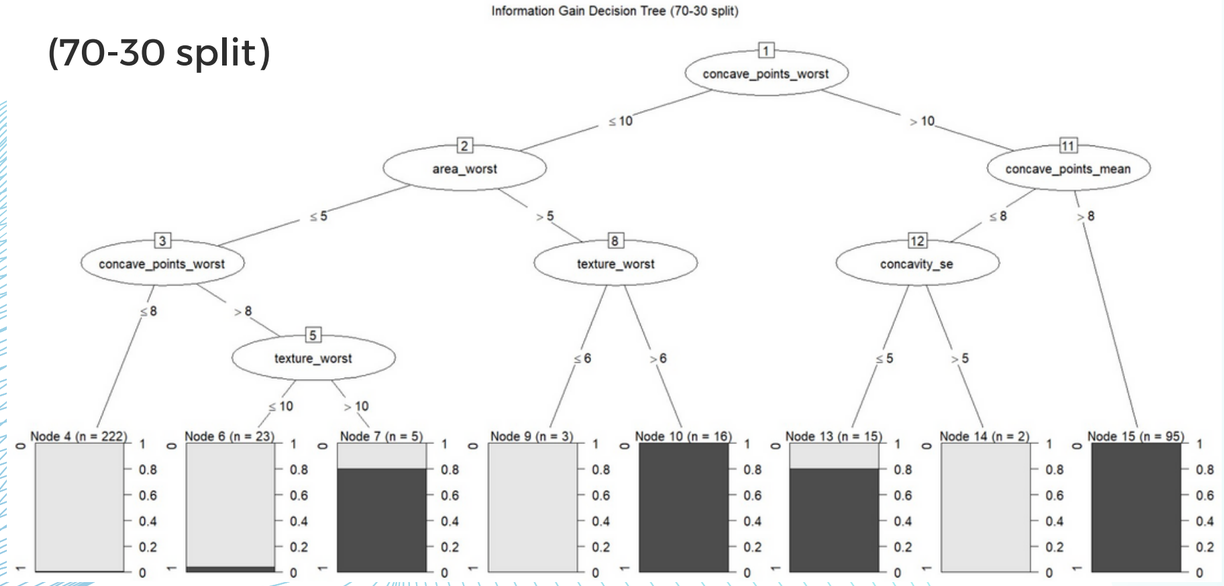
Algorithm: C4.5 algorithm using RWeka package
Function: The J48() function has Gain Ratio as the default attribute selecting measure.

INFORMATION GAIN DECISION TREES

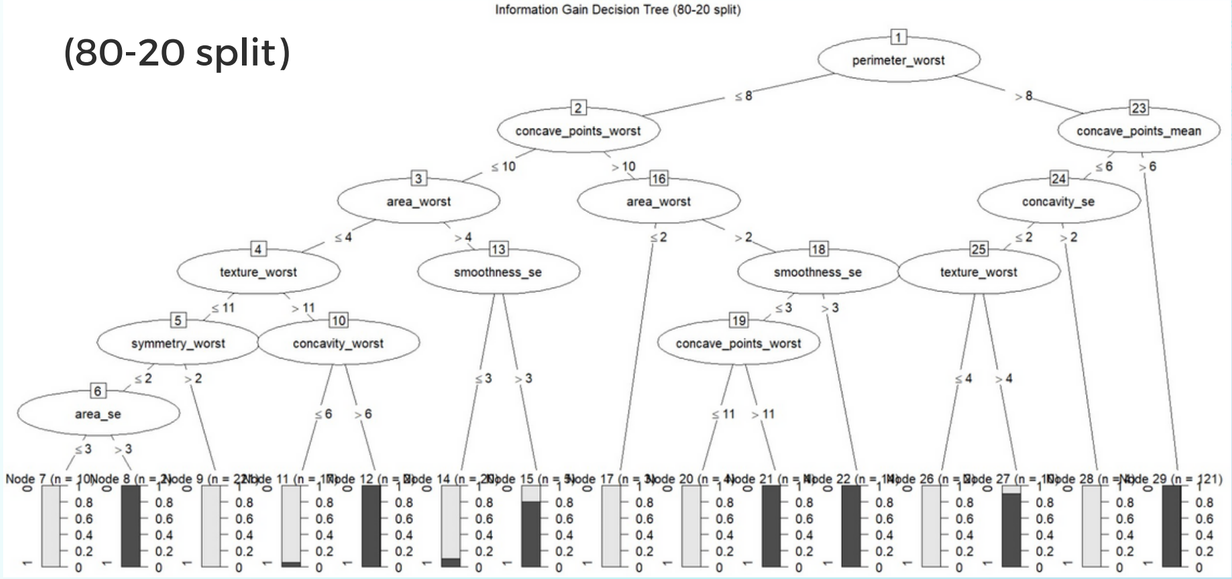
(60-40 split)



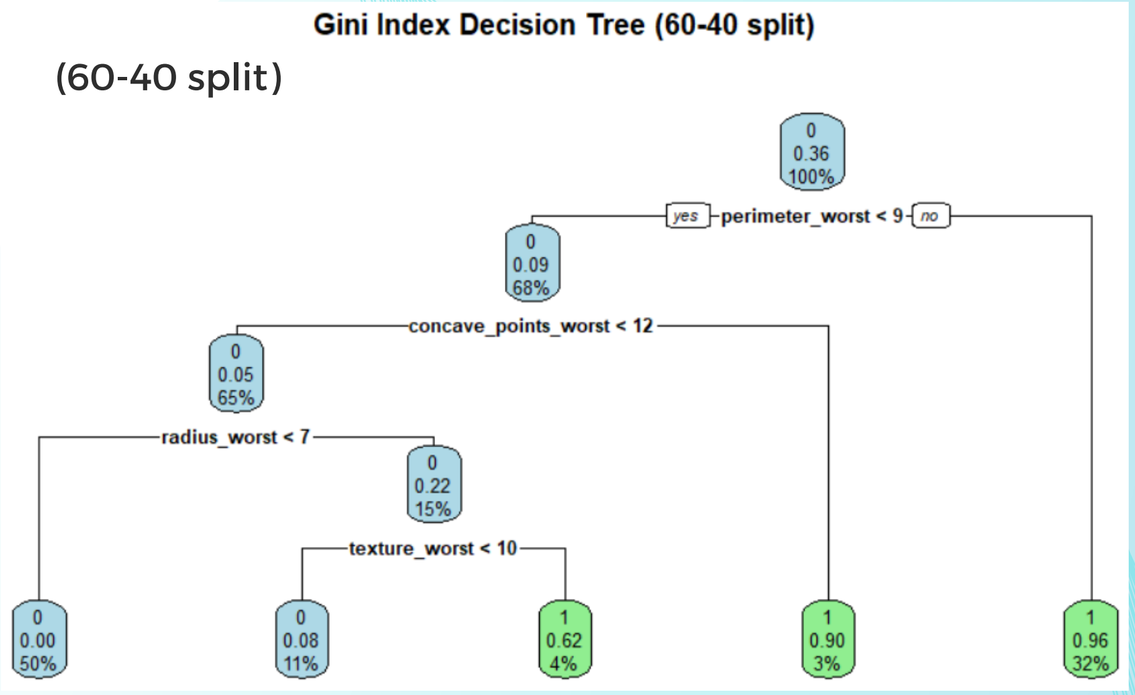
(70-30 split)

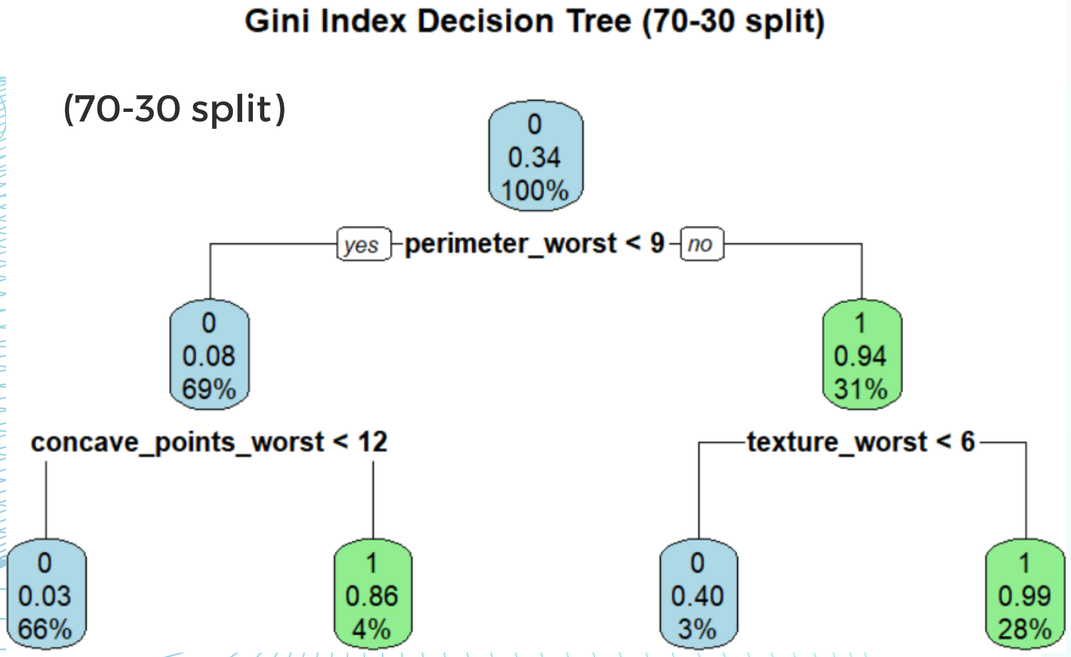


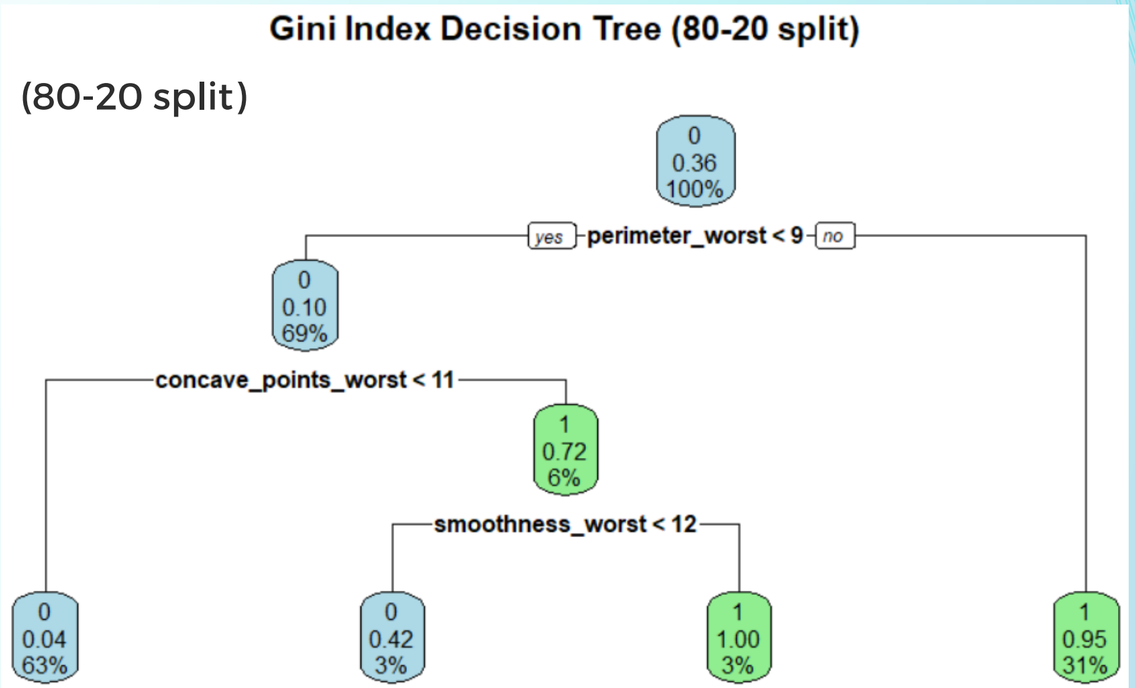
(80-20 split)



GINI INDEX DECISION TREES

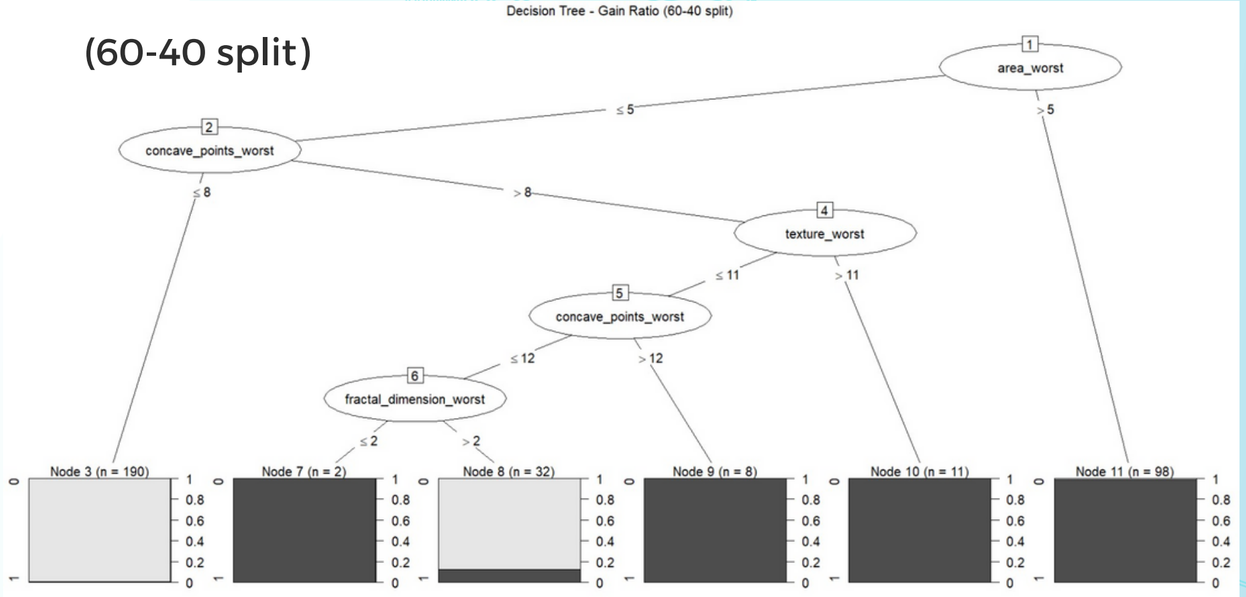




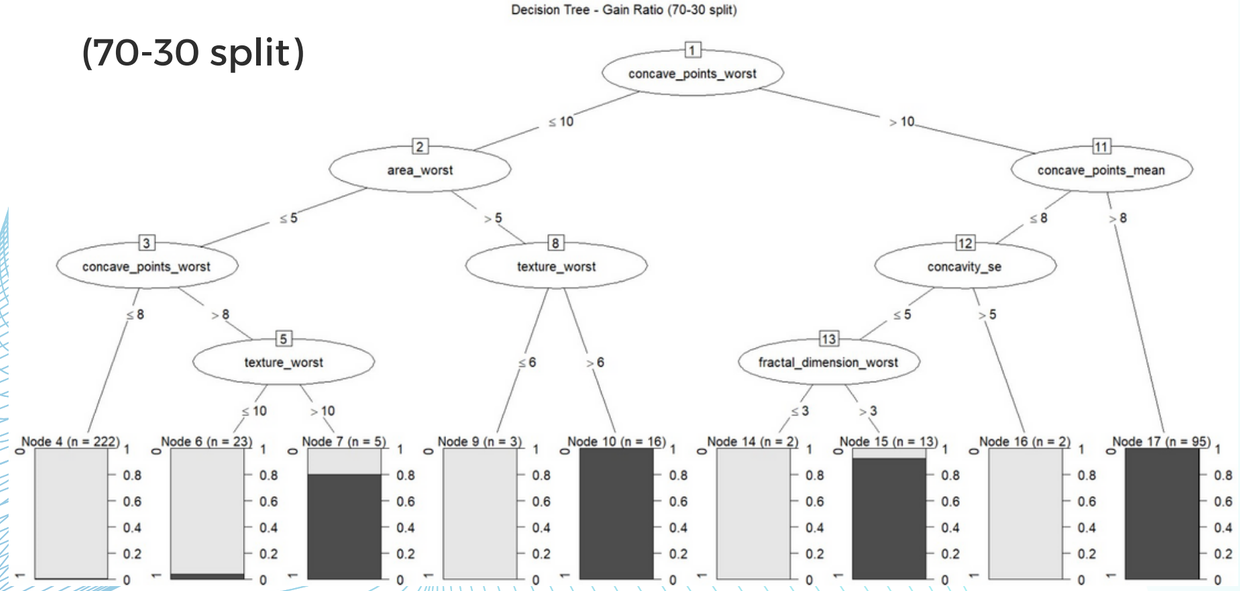


GAIN RATIO DECISION TREES

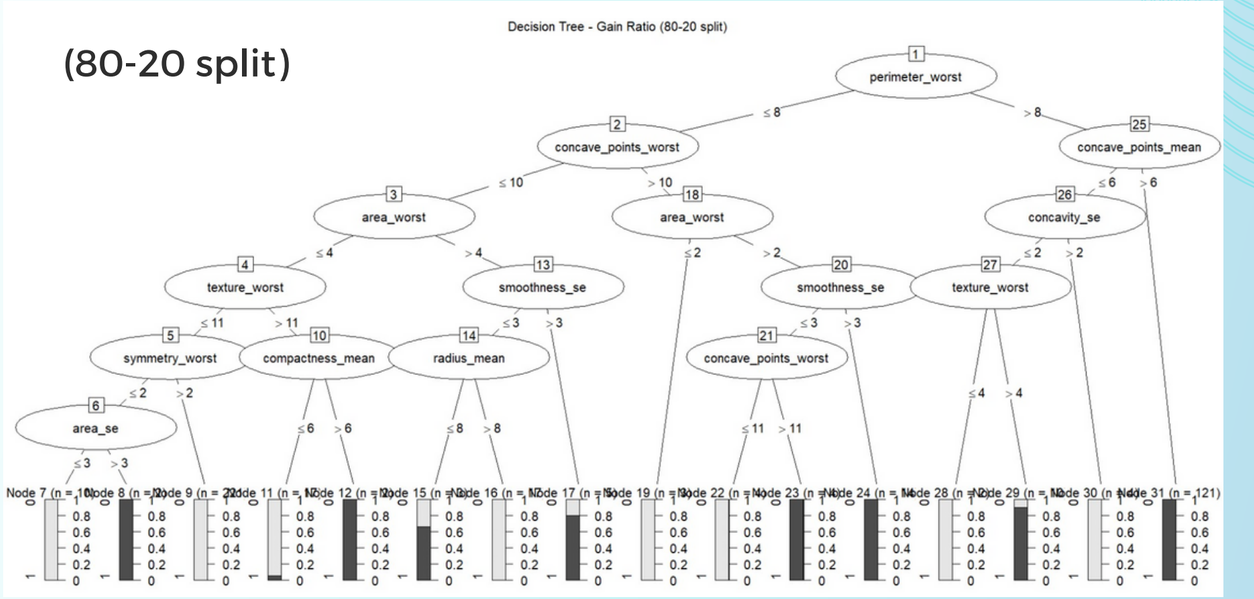
(60-40 split)



(70-30 split)



(80-20 split)



Model	Data Split	Accuracy	Precision	Sensitivity	Specificity
Information Gain	70-30 Split	89.66%	88.57%	86.11%	92.16%
Information Gain	60-40 Split	93.93%	92.11%	90.91%	95.62%
Information Gain	80-20 Split	92.24%	95.12%	84.78%	97.14%
Gini Index	70-30 Split	88.51%	77.14%	93.10%	86.21%
Gini Index	60-40 Split	90.19%	92.11%	82.35%	95.35%
Gini Index	80-20 Split	93.97%	95.12%	88.64%	97.22%
Gain Ratio	70-30 Split	89.08%	86.21%	91.55%	87.36%
Gain Ratio	60-40 Split	91.38%	97.56%	81.63%	98.51%
Gain Ratio	80-20 Split	88.08%	81.82%	89.47%	85.71%

different evaluation metrics and criteria are used to assess the performance of models like accuracy, precision, sensitivity, and specificity for models using different data splits and different splitting criteria (Information Gain, Gini Index, and Gain Ratio).

04.

CLUSTRING

Clustering k-means

CLUSTERING: THE PROCESS OF PARTITIONING A SET OF DATA OBJECTS INTO SUBSETS (CLUSTERS)

FOR CLUSTERING WE USED THE K-MEANS METHOD WITH THREE DIFFERENT RANDOM K SIZES

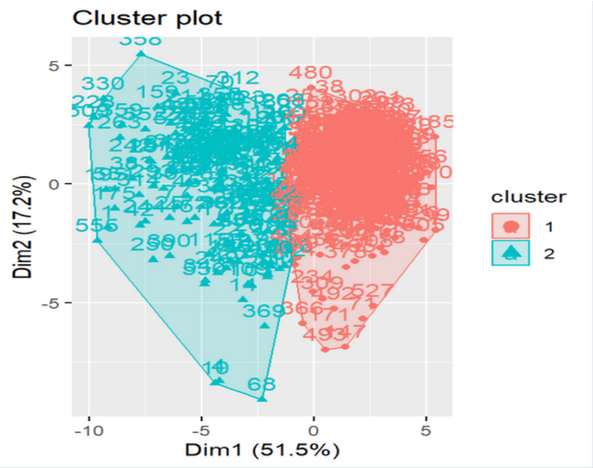
FOR K SIZES 3 K SIZES WERE PICKED RANDOMLY

IN THE CLUSTERING TASK, DIFFERENT METRICS ARE USED TO EVALUATE THE QUALITY OF CLUSTERING

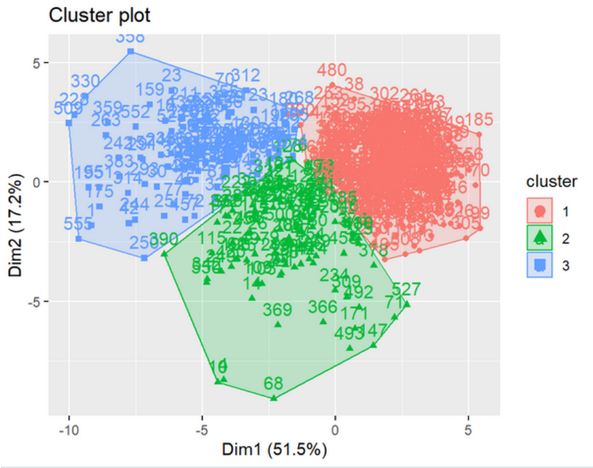
	K=2	K=3	K=4
Average Silhouette width	0.39	0.34	0.31
Total within-cluster sum of square	7121.733	6059.179	5670.971
Bcubed percision	0.8865035	0.8552597	0.8575291
Bcubed recall	0.8940466	0.699343	0.657348

COMPARING BETWEEN VISUALIZATION RESULTS

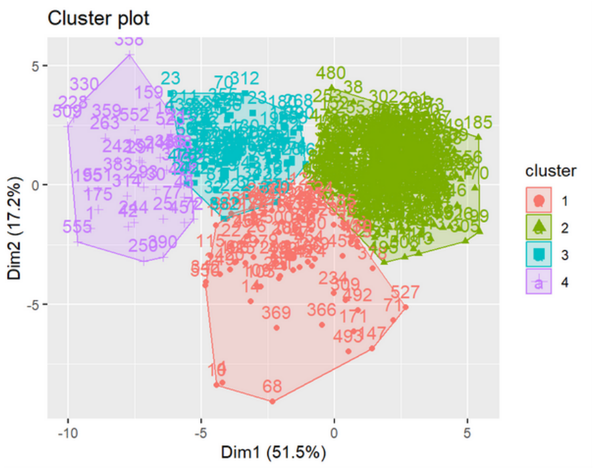
K=2



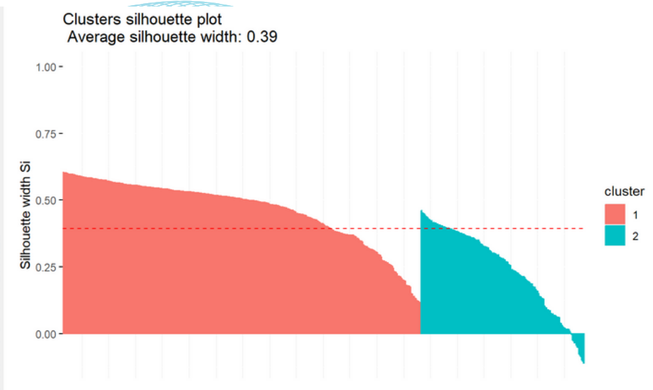
K=3

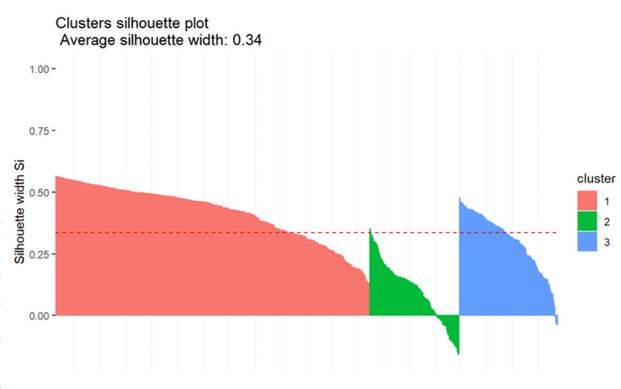


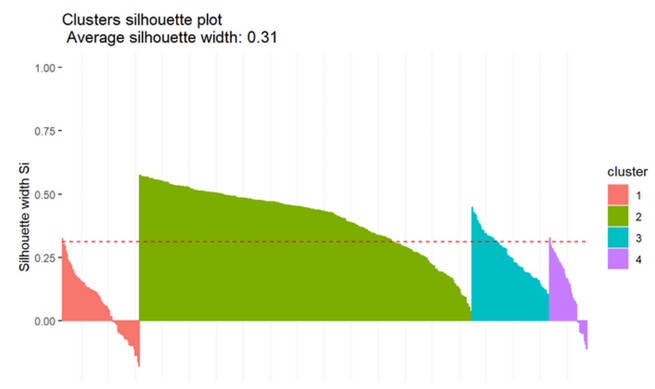
K=4



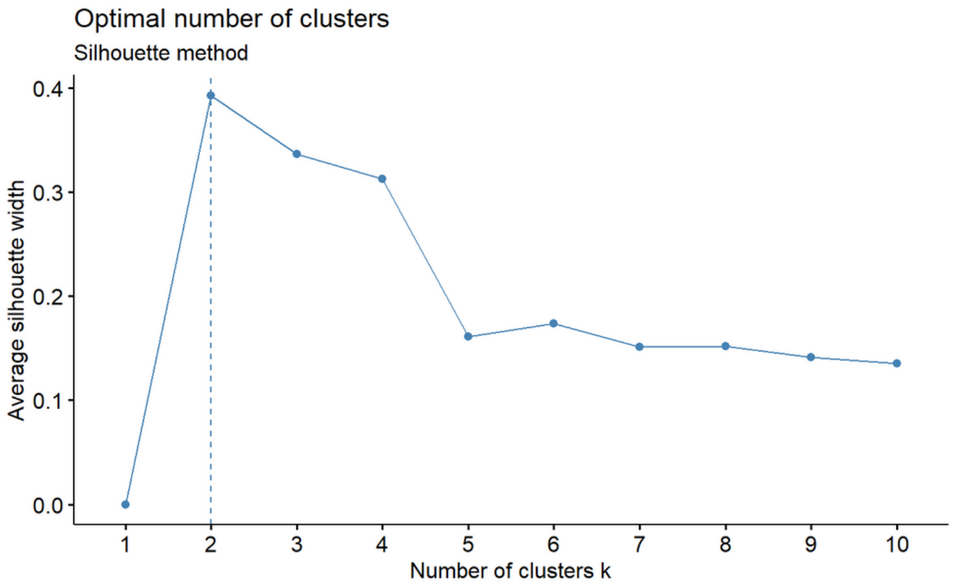
AVERAGE SILHOUETTE FOR EACH CLUSTERS



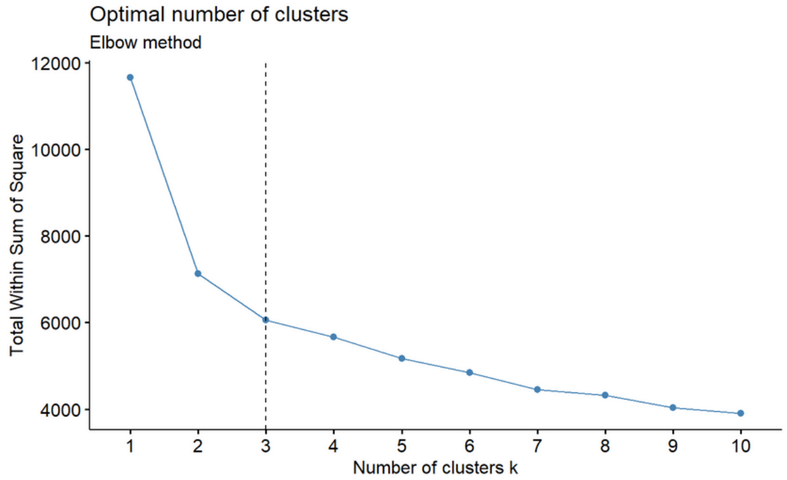




FIND OPTIMAL NUMBER OF CLUSTERS:



ELBOW METHOD



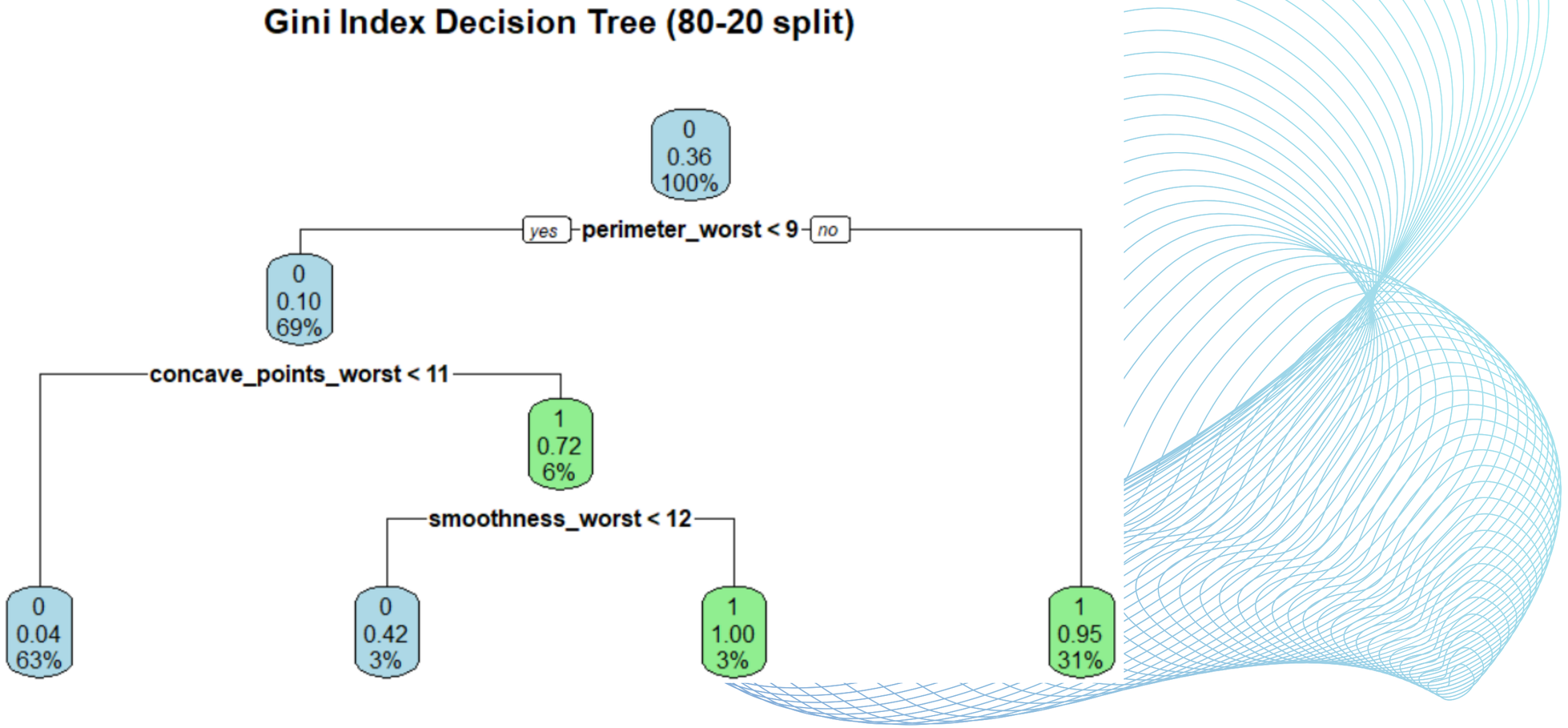

```
### b- Majority rule

According to the majority rule, the best number of clusters is 2
```

05.

FINDINGS

CLASSIFICATION RESULTS (DECISION TREE - GINI INDEX - 80-20 SPLIT):



FINDINGS:

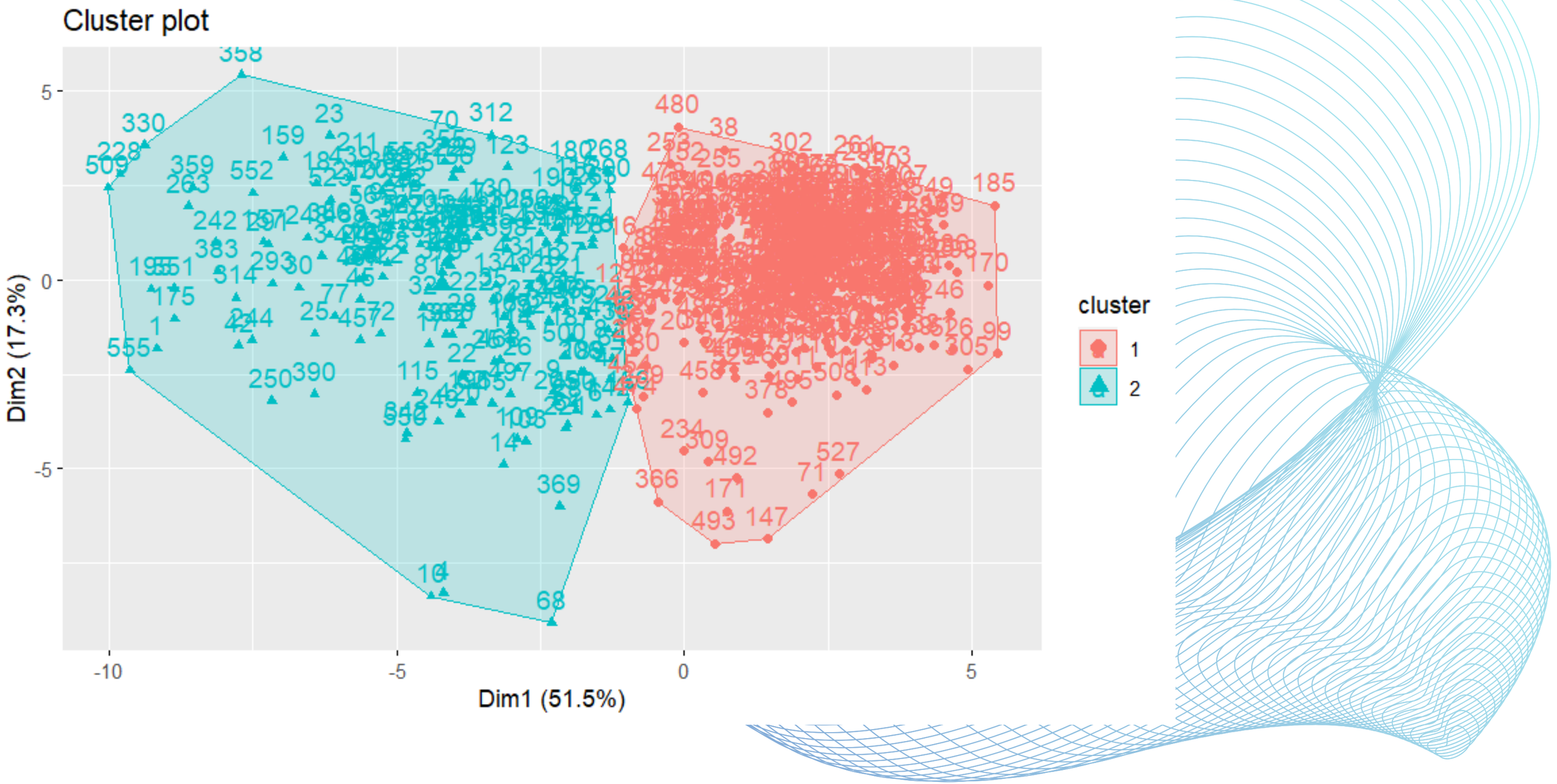
- Classification Results:
  - Accuracy: 93.97%
  - Precision: 95.12%
  - Sensitivity (Recall): 88.64%
  - Specificity: 97.22%

- Assessment of Classification Results:
  - High accuracy and specificity
  - Balanced precision and recall
  - Validation through Gini Index

- Problem Solutions:
  - Identified significant characteristics and extraction of diagnostic rules

- Conclusion:
  - Classification model achieves high accuracy, precision, sensitivity, and specificity
  - Provides clear decision boundary for distinguishing between malignant and benign masses

CLUSTERING RESULTS (K=2)



FINDINGS:

- Clustering Results:
  - Average Silhouette Width: 0.39
  - Total Within-Cluster Sum of Squares: 7121.733
  - Bcubed Precision: 0.8865035
  - Bcubed Recall: 0.8940466

- Assessment of Clustering Results:
  - Well-separated clusters (high silhouette width)
  - Tight clusters (low within-cluster sum of squares)
  - High Bcubed precision and recall

- Problem Solutions:
  - Identified differences between clusters in tumor measurements

FINDINGS:

- Overall Assessment:
  - Clinical significance and relevance
  - Decision Tree interpretability

THANK YOU FOR LISTENING

ANY QUESTIONS?

Prepared by:

ARWA MESLOUB: 443203895

WAREF ALYOUSEF: 442200377

LEEN AL-HARBI: 443201050

MUNIRA ALMOGREN: 443203895

Supervised by:

Dr. Mashael Aldayel