Brest Cancer Detection

RITUPARNA DE

Introduction

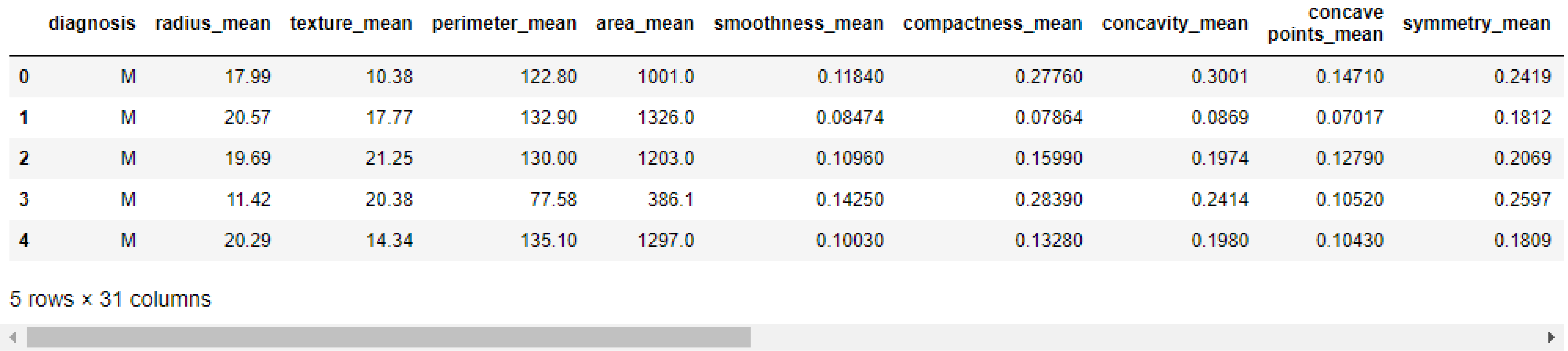
Breast cancer is a disease in which cells in the breast grow out of control in a rapidly. Breast cancer occurs when a malignant (cancerous) tumor originates in the breast cells. It is the most commonly occurring cancer in women and the second most common cancer overall. Around 2 million cases were observed in 2018. The early diagnosis of breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients affected. Further accurate classification from the data of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of breast cancer and the classification of patients into malignant or benign groups is the subject of all research done and observed. Because of its unique advantages in critical features detection from complex breast cancer datasets, machine learning (ML) is widely recognized as the methodology of choice in Breast Cancer pattern classification.

Objective

This project is a relative study of the implementation of models using **Decision Tree Algorithm** and **K-Nearest Neighbors Algorithm**, which is done on the data set taken from the UCI repository.

Data Overview

The whole data can be used from here; https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

5 rows × 31 columns

Data Overview

Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things. This data has –

- Total 33 columns .
- Total 669 rows .
- Total 669 null values .

Attribute Information:

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) Area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

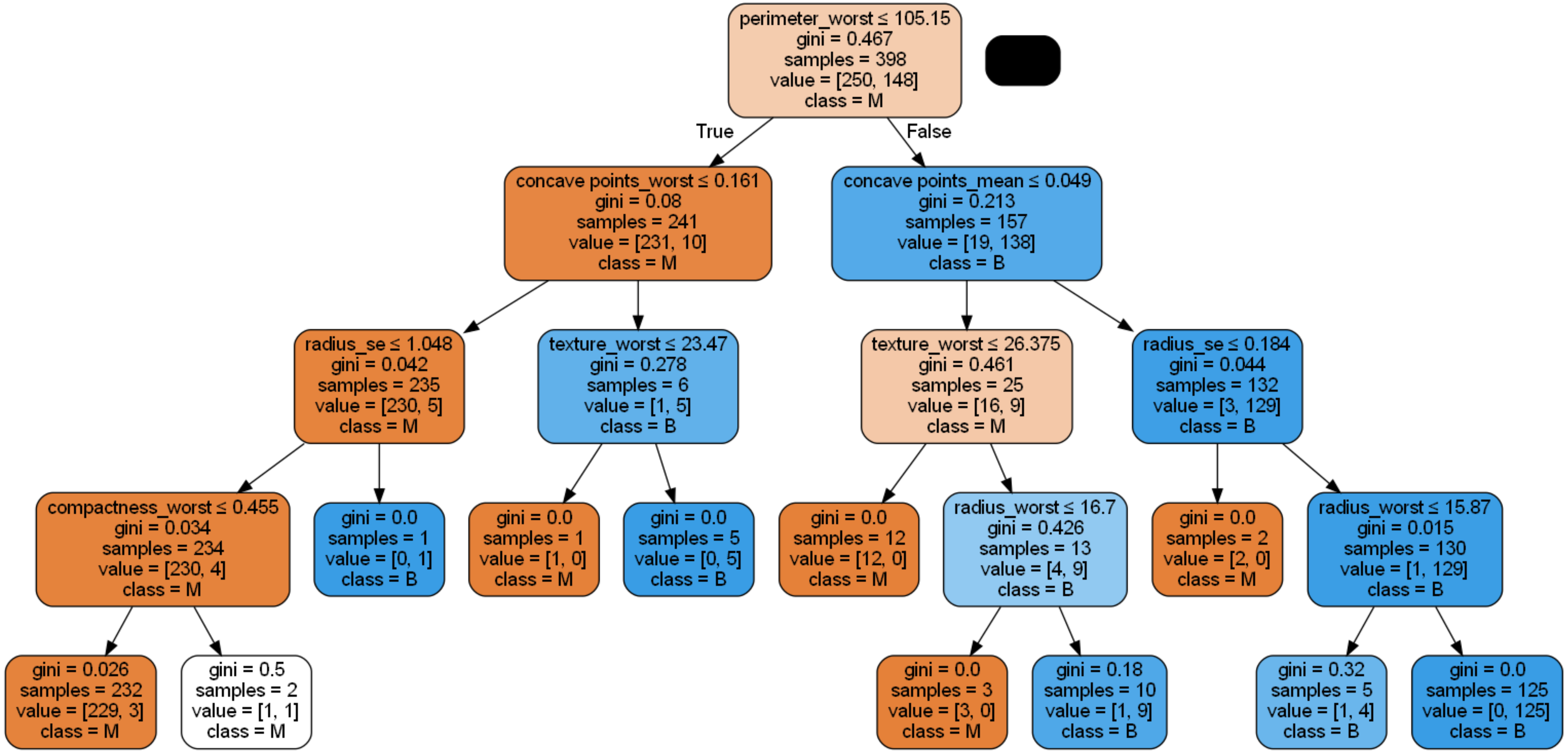
The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

Decision Tree



[redacted]

perimeter_worst ≤ 105.15
gini = 0.467
samples = 398
value = [250, 148]
class = M

True

False

concave points_worst ≤ 0.161
gini = 0.08
samples = 241
value = [231, 10]
class = M

concave points_mean ≤ 0.049
gini = 0.213
samples = 157
value = [19, 138]
class = B

radius_se ≤ 1.048
gini = 0.042
samples = 235
value = [230, 5]
class = M

texture_worst ≤ 23.47
gini = 0.278
samples = 6
value = [1, 5]
class = B

texture_worst ≤ 26.375
gini = 0.461
samples = 25
value = [16, 9]
class = M

radius_se ≤ 0.184
gini = 0.044
samples = 132
value = [3, 129]
class = B

compactness_worst ≤ 0.455
gini = 0.034
samples = 234
value = [230, 4]
class = M

gini = 0.0
samples = 1
value = [0, 1]
class = B

gini = 0.0
samples = 1
value = [1, 0]
class = M

gini = 0.0
samples = 5
value = [0, 5]
class = B

gini = 0.0
samples = 12
value = [12, 0]
class = M

radius_worst ≤ 16.7
gini = 0.426
samples = 13
value = [4, 9]
class = B

gini = 0.0
samples = 2
value = [2, 0]
class = M

radius_worst ≤ 15.87
gini = 0.015
samples = 130
value = [1, 129]
class = B

gini = 0.026
samples = 232
value = [229, 3]
class = M

gini = 0.5
samples = 2
value = [1, 1]
class = M

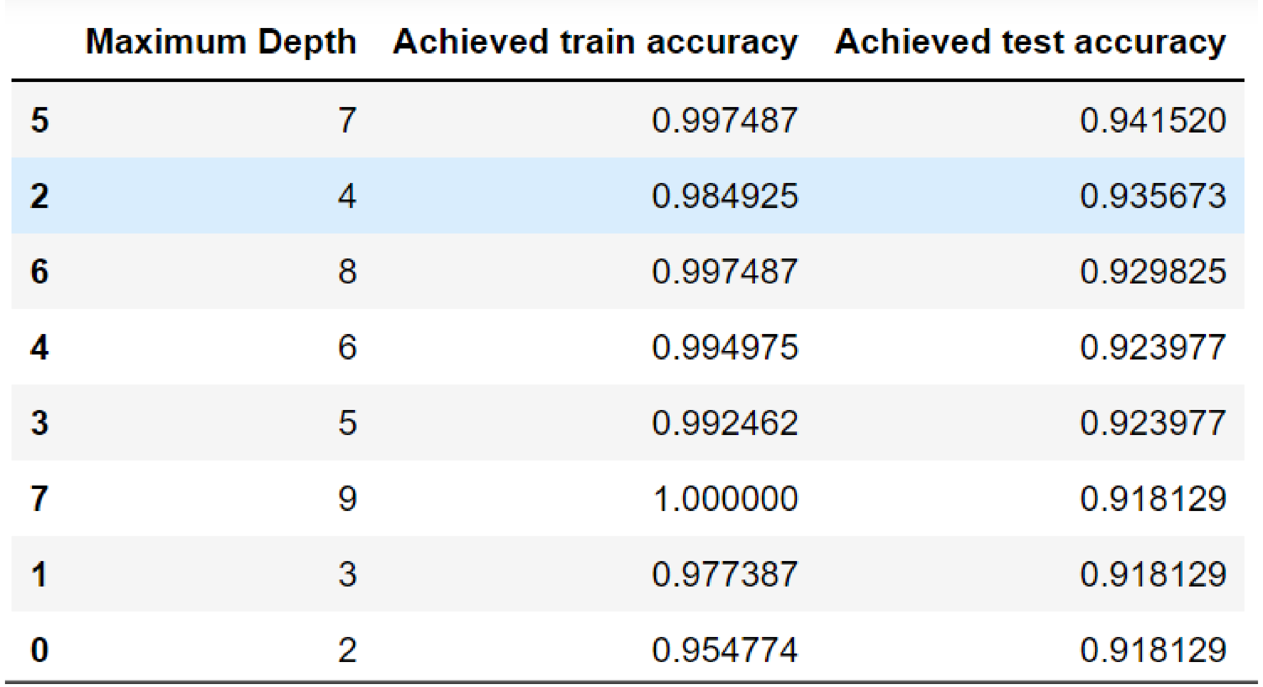
gini = 0.0
samples = 3
value = [3, 0]
class = M

gini = 0.18
samples = 10
value = [1, 9]
class = B

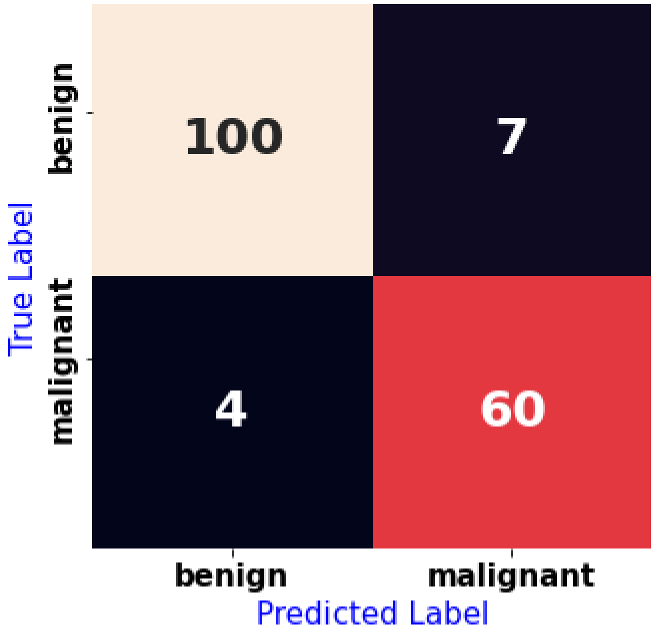
gini = 0.32
samples = 5
value = [1, 4]
class = B

gini = 0.0
samples = 125
value = [0, 125]
class = B

Model Performance (Decision Tree)



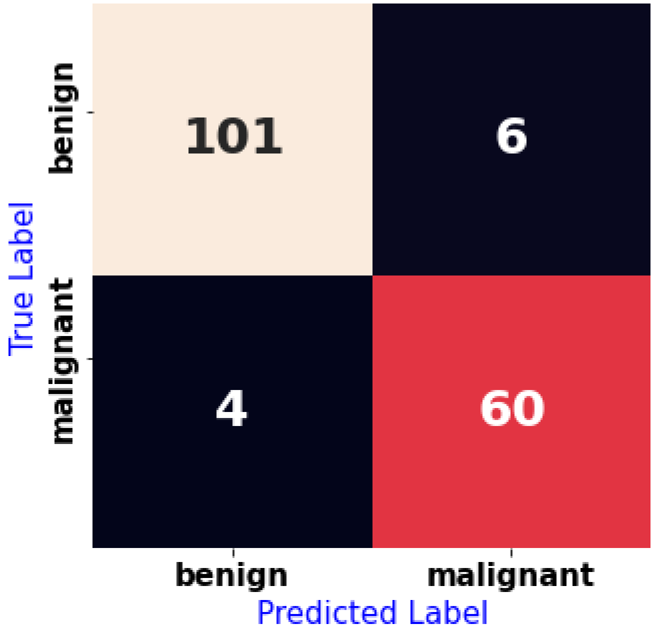
	Maximum Depth	Achieved train accuracy	Achieved test accuracy
5	7	0.997487	0.941520
2	4	0.984925	0.935673
6	8	0.997487	0.929825
4	6	0.994975	0.923977
3	5	0.992462	0.923977
7	9	1.000000	0.918129
1	3	0.977387	0.918129
0	2	0.954774	0.918129

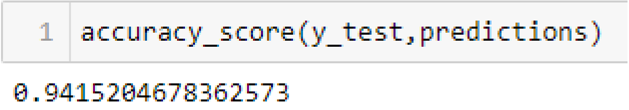


After sorting the result table we clearly see that test accuracy is quite same for maximum depth 4,8 and 6. But there can overfitting issue with with maximum depth 8 and 6 as training accuracy is too high. So, we can choose the decision tree with maximum depth 4, where train and test accuracy trade off is quire nice.

Model Performance (KNN & SVM)

KNN

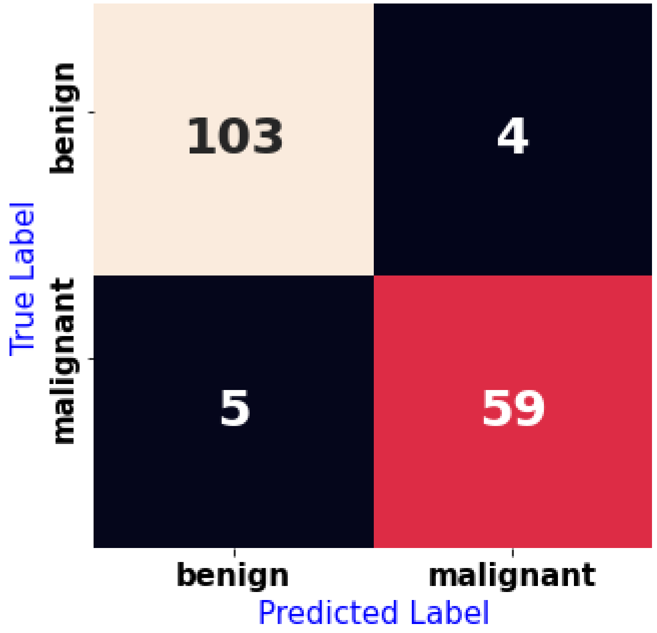


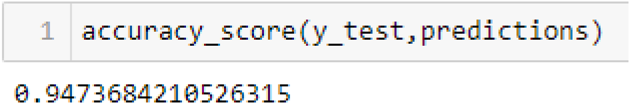


```
accuracy_score(y_test,predictions)
```

```
0.9415204678362573
```

SVM





```
accuracy_score(y_test,predictions)
```

```
0.9473684210526315
```

Conclusion

We choose Support Vector Machine (SVM) algorithm, as it is giving best accuracy 95%. If we detect Brest cancer depends on these independent variables in early stage, then it can be very helpful to society.

Thank You