Assignment 2

Adharsh Sundaram Soudakar (23796349)

D1: *P.S: please zoom in to clearly view the contents of the image.*

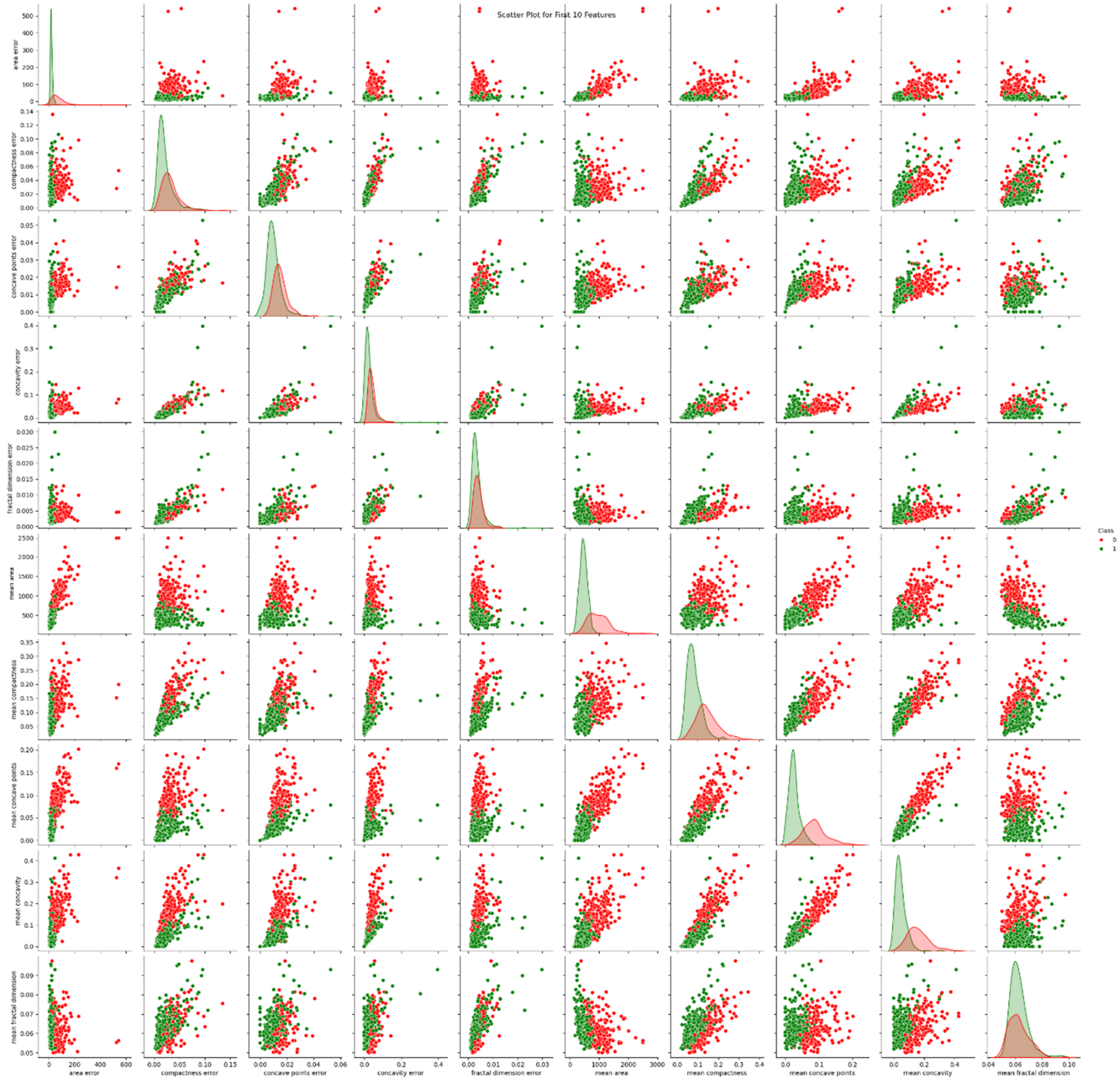


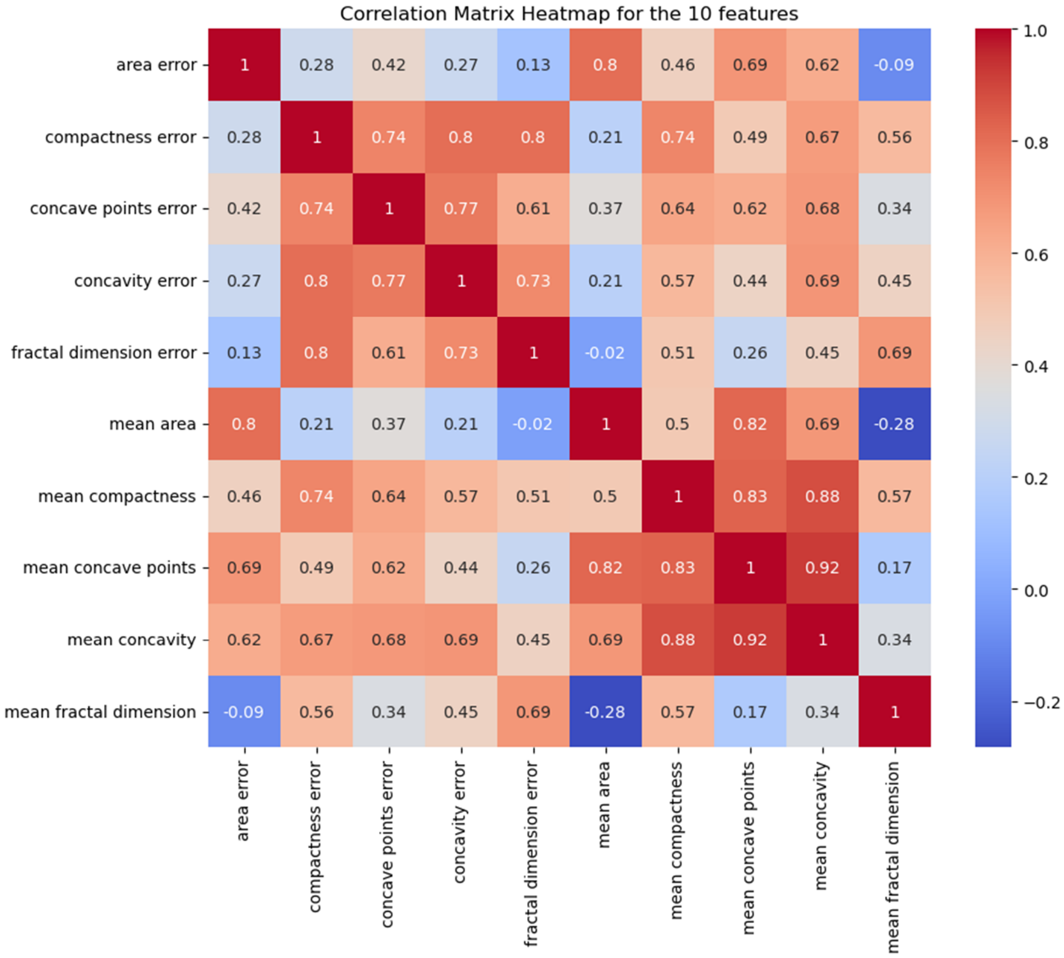
Fig. Scatter plot for the first 10 features.

- Class 0 – Malignant.
- Class 1 – Benign.

D2: Comments on the scatterplot:

- While there are many relationships that are linear and certain features that are positively correlated, there are few relationships that are complex and non-linear. For example, mean fractal dimension and mean compactness have a positive trend whereas mean fractal dimension and mean area have a negative trend.
- Yes, clusters of groups are present. The cluster formed by red points relates to Class (0) (Malignant) and by green points relates to Class (1) (Benign). Some features have distinct clusters (clear separation between the classes), these could help with classification tasks.
- Yes, there are some instances that could be outliers, points that appear away from the clusters. For example, relationship between mean area and mean compactness, there are several outliers for the class (0).
- Yes, certain features can be removed, that do not form a relationship that's positive and their clusters don't have clear separation between the classes with several outliers. For example, mean concavity, most of its relationship with other features are very linear (redundancy), the clusters are not well separated, there are several outliers as well. But further analysis is required before these decisions.

D3:



	area error	compactness error	concave points error	concavity error	fractal dimension error	mean area	mean compactness	mean concave points	mean concavity	mean fractal dimension
area error	1	0.28	0.42	0.27	0.13	0.8	0.46	0.69	0.62	-0.09
compactness error	0.28	1	0.74	0.8	0.8	0.21	0.74	0.49	0.67	0.56
concave points error	0.42	0.74	1	0.77	0.61	0.37	0.64	0.62	0.68	0.34
concavity error	0.27	0.8	0.77	1	0.73	0.21	0.57	0.44	0.69	0.45
fractal dimension error	0.13	0.8	0.61	0.73	1	-0.02	0.51	0.26	0.45	0.69
mean area	0.8	0.21	0.37	0.21	-0.02	1	0.5	0.82	0.69	-0.28
mean compactness	0.46	0.74	0.64	0.57	0.51	0.5	1	0.83	0.88	0.57
mean concave points	0.69	0.49	0.62	0.44	0.26	0.82	0.83	1	0.92	0.17
mean concavity	0.62	0.67	0.68	0.69	0.45	0.69	0.88	0.92	1	0.34
mean fractal dimension	-0.09	0.56	0.34	0.45	0.69	-0.28	0.57	0.17	0.34	1

Fig. Correlation matrix with correlation coefficient.

D4: Interpretation:

- Yes, this supports the previous observations made using the scatterplot. For example, the very linear relationship observed between mean concavity and mean concave points is supported here by the high correlation coefficient.

D6: Tables. Performance metrics and confusion matrix for Decision tree model with default hyperparameters

Performance metrics	Training Set	Test Set
Accuracy	1	0.96
Precision	1	0.97
Recall	1	0.97

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	39	2
Benign class (1)	2	71

D7: Comments on results:

- Yes, the classifier is overfitting, perfect results on the training set but not as good on the test set. With regards to confusion matrix, there are two false positives and two false negatives. Generalization capacity of the classifier is not up to the mark considering the problem (classifying cancer cells).
- Possible reasons could be that several instances were outliers and the size of the dataset (only 569 instances) which is very small, not enough for training the classifier.

D8: *P.S: please zoom in to clearly view the contents of the image.*

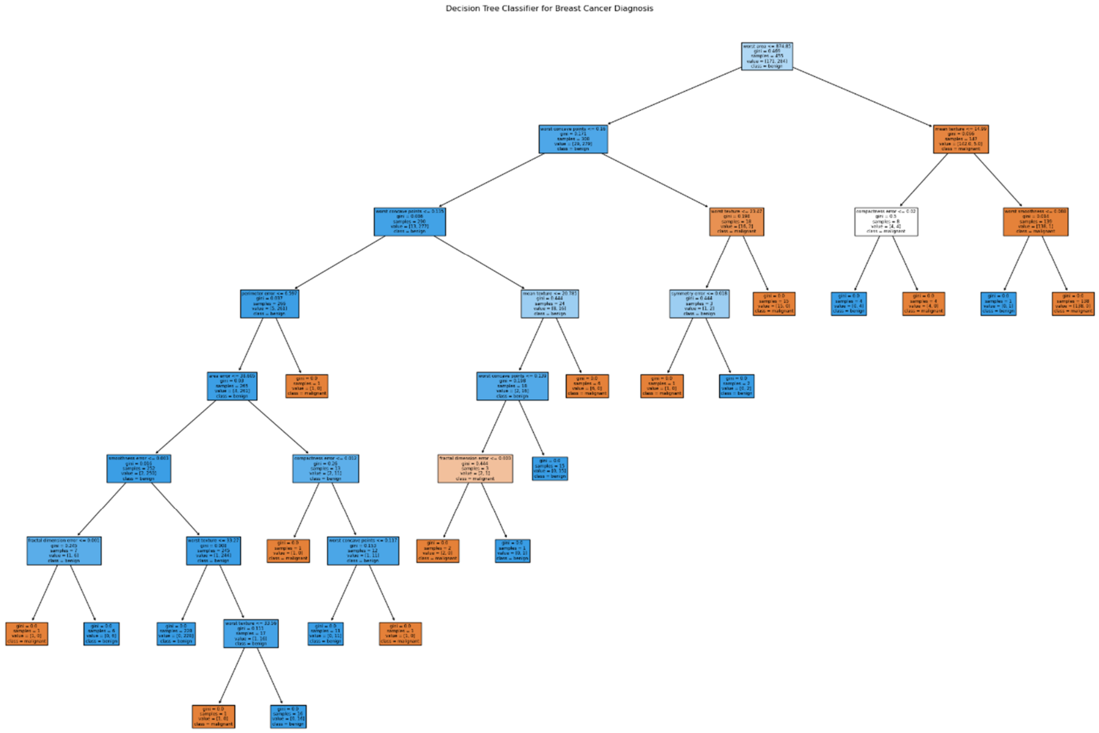

Decision Tree Classifier for Breast Cancer Diagnosis
worst area <= 874.85 gini = 0.469 samples = 455 value = [171, 284] class = benign
worst concave points <= 0.16 gini = 0.171 samples = 308 value = [29, 279] class = benign
mean texture <= 14.99 gini = 0.066 samples = 147 value = [142.0, 5.0] class = malignant
worst concave points <= 0.135 gini = 0.086 samples = 290 value = [13, 277] class = benign
worst texture <= 23.47 gini = 0.198 samples = 18 value = [16, 2] class = malignant
compactness error <= 0.02 gini = 0.5 samples = 8 value = [4, 4] class = malignant
worst smoothness <= 0.088 gini = 0.014 samples = 139 value = [138, 1] class = malignant
perimeter error <= 6.597 gini = 0.037 samples = 266 value = [5, 261] class = benign
mean texture <= 20.785 gini = 0.444 samples = 24 value = [8, 16] class = benign
symmetry error <= 0.018 gini = 0.444 samples = 3 value = [1, 2] class = benign
gini = 0.0 samples = 15 value = [15, 0] class = malignant
gini = 0.0 samples = 4 value = [0, 4] class = benign
gini = 0.0 samples = 4 value = [4, 0] class = malignant
gini = 0.0 samples = 1 value = [0, 1] class = benign
gini = 0.0 samples = 138 value = [138, 0] class = malignant
area error <= 38.605 gini = 0.03 samples = 265 value = [4, 261] class = benign
gini = 0.0 samples = 1 value = [1, 0] class = malignant
worst concave points <= 0.139 gini = 0.198 samples = 18 value = [2, 16] class = benign
gini = 0.0 samples = 6 value = [6, 0] class = malignant
gini = 0.0 samples = 1 value = [1, 0] class = malignant
gini = 0.0 samples = 2 value = [0, 2] class = benign
smoothness error <= 0.003 gini = 0.016 samples = 252 value = [2, 250] class = benign
compactness error <= 0.012 gini = 0.26 samples = 13 value = [2, 11] class = benign
fractal dimension error <= 0.003 gini = 0.444 samples = 3 value = [2, 1] class = malignant
gini = 0.0 samples = 15 value = [0, 15] class = benign
fractal dimension error <= 0.001 gini = 0.245 samples = 7 value = [1, 6] class = benign
worst texture <= 33.27 gini = 0.008 samples = 245 value = [1, 244] class = benign
gini = 0.0 samples = 1 value = [1, 0] class = malignant
worst concave points <= 0.117 gini = 0.153 samples = 12 value = [1, 11] class = benign
gini = 0.0 samples = 2 value = [2, 0] class = malignant
gini = 0.0 samples = 1 value = [0, 1] class = benign
gini = 0.0 samples = 1 value = [1, 0] class = malignant
gini = 0.0 samples = 6 value = [0, 6] class = benign
gini = 0.0 samples = 228 value = [0, 228] class = benign
worst texture <= 33.56 gini = 0.111 samples = 17 value = [1, 16] class = benign
gini = 0.0 samples = 11 value = [0, 11] class = benign
gini = 0.0 samples = 1 value = [1, 0] class = malignant
gini = 0.0 samples = 1 value = [1, 0] class = malignant
gini = 0.0 samples = 16 value = [0, 16] class = benign

Fig. Decision tree built from the training process.

D9: Comments:

- The number of levels / depth of the tree is 8. (Number of leaves is 20).
- Yes, the large number of levels and leaves, considering the size of the dataset, is a sign of overfitting.
- Certain leaves have very small sample size which leads to overfitting as the tree is kind of complex, and as each leaf represents a decision rule, the model has learned many rules meaning the training could have been affected by the noise/outliers in the dataset.
- To a certain extent, yes. The left side of the tree is complex to follow.

D10:

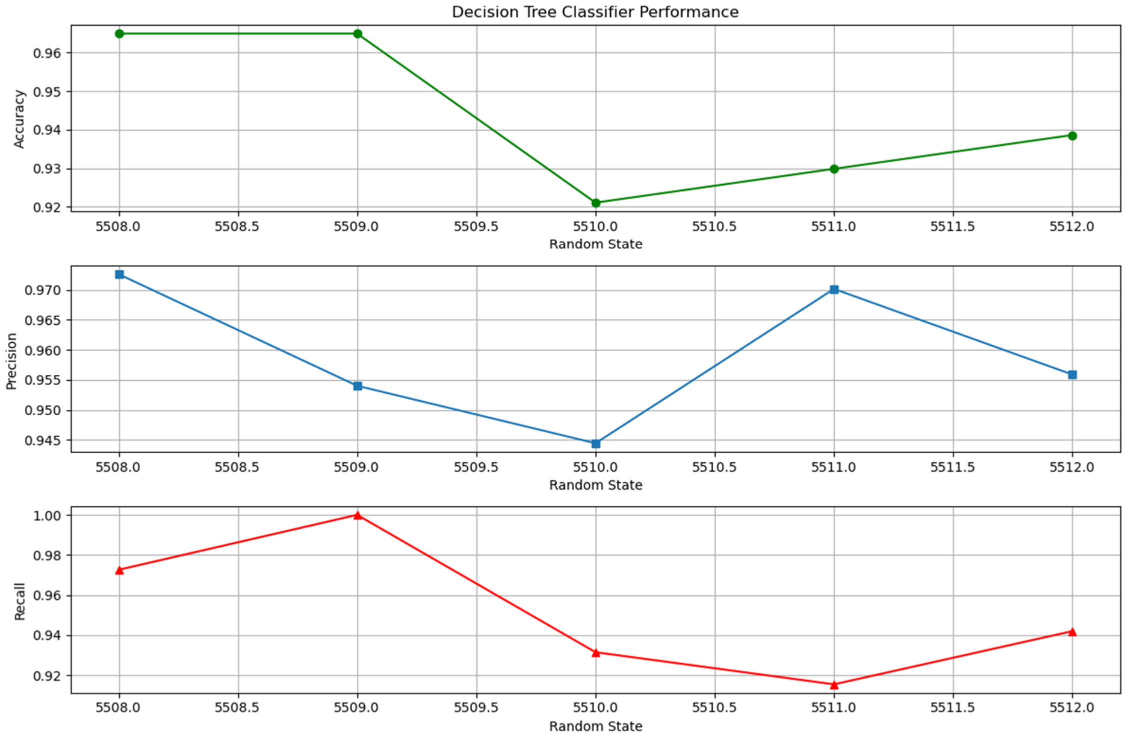


Fig. Decision tree classifier performance on test set for different random states.

Comments on consistency:

- **5508** – is accurate with best consistency across all metrics. Precision and recall are relatively high.
- **5509** – has high accuracy and recall but at the cost of precision.
- **5510** – has the lowest metrics of all random states. Both precision and recall are almost equal.
- **5511** – has moderate accuracy with high precision at the cost of recall.
- **5512** – is similar to 5510 random state with slightly better metrics.

D11:

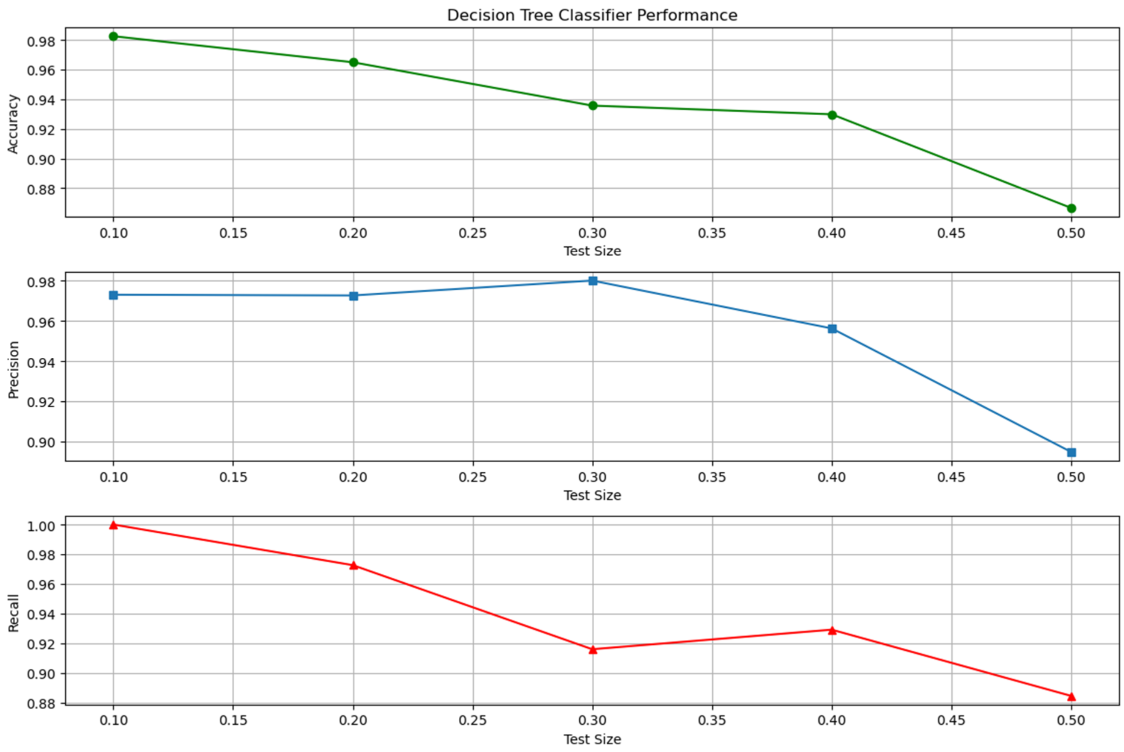


Fig. Decision tree classifier performance on test set for different data splits.

Comments on results:

- Accuracy has an expected negative trend, as the test size increases, accuracy decreases.
- Precision has a negative trend with an unexpected spike at 70%-30% split. In fact, this split has the best precision of all splits.
- Recall also has a negative trend. The dip at 70%-30% split (possibly has the greatest number of false negatives) is expected and proportional to high precision.
- Overall, the performance behaviour is as expected. The highest performance in all metrics was achieved with the expected 90%-10% split. Shows that the classifier requires a substantial amount of data to effectively identify all positive instances without mistakes.

D12: Tables. Performance metrics and confusion matrix for Decision tree model with optimal hyperparameters

Optimal Hyperparameters	
max_depth	3
min_samples_split	2
min_samples_leaf	2

Performance Metrics	Training Set	Test Set
Accuracy	0.96	0.94
Precision	0.95	0.96
Recall	0.99	0.95

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	38	3
Benign class (1)	4	69

D13: Comments on results of optimisation:

- Using optimal hyperparameters has not improved the generalization capacity of the classifier. In fact, all the performance metrics with optimal hyperparameters are not as good as seen on D6 over the test set.
- The confusion matrix also shows an increase in the number of false negatives and false positives.
- No, fine-tuning the hyperparameters has led to a slight decrement in the classifier's performance. But the overfitting problem in D6, is probably not present in D12 as the performance metrics over the training and test sets are almost similar.

D14: Tables. Results of different scoring options.

Optimal Hyperparameters for different scoring options	Accuracy	Precision	Recall
max_depth	3	4	3
min_samples_split	2	2	2
min_samples_leaf	2	2	5

Accuracy:

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	38	3
Benign class (1)	4	69

Precision:

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	39	2
Benign class (1)	4	69

Recall:

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	38	3
Benign class (1)	4	69

Comments:

- With scoring option set as Accuracy or Recall, false positive is one more than the option set to Precision. Considering the problem, tumour type, mistakes have big implications. So, scoring options Precision is better when compared to Accuracy and Recall as per the results from the confusion matrices.

D15: Table. Feature Importance of each feature obtained from the training process with scoring option as Accuracy with optimal hyperparameters. (The values are rounded off to two decimal places.)

Feature	Importance
Worst area	0.80
Worst concave points	0.15
Mean smoothness	0.02
Mean texture	0.01
Worst texture	0.01
Perimeter error	0.00
Mean area	0.00
Texture error	0.00
Worst symmetry	0.00
Worst smoothness	0.00
Worst fractal dimension	0.00
Worst concavity	0.00
Concave points error	0.00
Worst compactness	0.00
Concavity error	0.00
Symmetry error	0.00
Mean compactness	0.00
Smoothness error	0.00
Compactness error	0.00
Mean Symmetry	0.00
Fractal dimension error	0.00
Mean fractal dimension	0.00
Mean concavity	0.00
Mean concave points	0.00
Area error	0.00

D16: Table. Filtered features (>0.01%) and total feature importance value of retained features.

Retained Features	Removed Features	Total feature importance retained (rounded off to 2 decimal places)
Mean smoothness	Mean area	1.00 (actual 0.997)
Mean texture	Mean compactness	
Worst area	Mean concavity	
Worst concave points	Mean concave points	
Worst texture	Mean symmetry	
	Mean fractal dimension	
	Texture error	
	Perimeter error	
	Area error	
	Smoothness error	
	Compactness error	
	Concavity error	
	Concave points error	
	Symmetry error	
	Fractal dimension error	
	Worst smoothness	
	Worst compactness	
	Worst concavity	
	Worst symmetry	
	Worst fractal dimension	

D17: Tables. Comparison of performance metrics between all features and reduced features

Training set metrics	Accuracy	Precision	Recall
All features	0.96	0.95	0.99
Reduced features	0.98	0.97	0.99

Test set metrics	Accuracy	Precision	Recall
All features	0.94	0.96	0.95
Reduced features	0.95	0.97	0.95

All features:

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	38	3
Benign class (1)	4	69

Reduced features:

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	39	2
Benign class (1)	4	69

D18: Comments on the results (All features vs Reduced features)

- On both the training set and test set, performance metrics (Accuracy, Precision, and Recall) from reduced features are better. With regards to confusion matrix, all features have one more false positive than reduced features. Thus, reduced features have impacted the Generalization capacity of the classifier, improving it a little.

D19: Results of Random Forest model

Optimal number of estimators	100
Max depth	4

Performance Metrics	Training Set	Test Set
Accuracy	0.99	0.98
Precision	0.99	0.99
Recall	1	0.99

Confusion Matrix (test set)	Malignant class (0)	Benign class (1)
Malignant class (0)	40	1
Benign class (1)	1	72

D20: Performance comparison

- The performance metrics of random forest model are better than decision tree classifier with optimal hyperparamters. There are only 1 instance of false positive and false negative in the confusion matrix on the test set.
- The depth of the tree has reduced from 8 to 4.
- Yes, in general random forest classifier reduces overfitting as the depth of the tree is not allowed to grow and it also handles unbalanced datasets well, the dataset is unbalanced in this case. (Ensemble learning results in averaging the results of many trees, each of which are trained on a different subset of the data.)

D21: Thoughts on application of the models created.

- No, the models created are not perfect and can either falsely identify or miss an instance, both the cases have big consequences.
- Maybe, but model’s complexity does not necessarily improve the performance, in some instances, a complex model performs worse than simple model. In our case, if the dataset was larger, the model would have learned more complex patterns which could have resulted in better performance.
- No, the decision process cannot be automated, when it comes to the medical field, a doctor/surgeon’s intervention is necessary as computers/ML algorithms cannot be blindly trusted as its capacity to decide is limited to the data that it was trained on and can also be inaccurate. Considering the problem, a wrong decision, could result in the wrong medication, which could result in loss of life.
- Yes, the dataset is not balanced (Bias, Overfitting etc.) and not voluminous enough (Generalization capacity, Overfitting, Complexity).