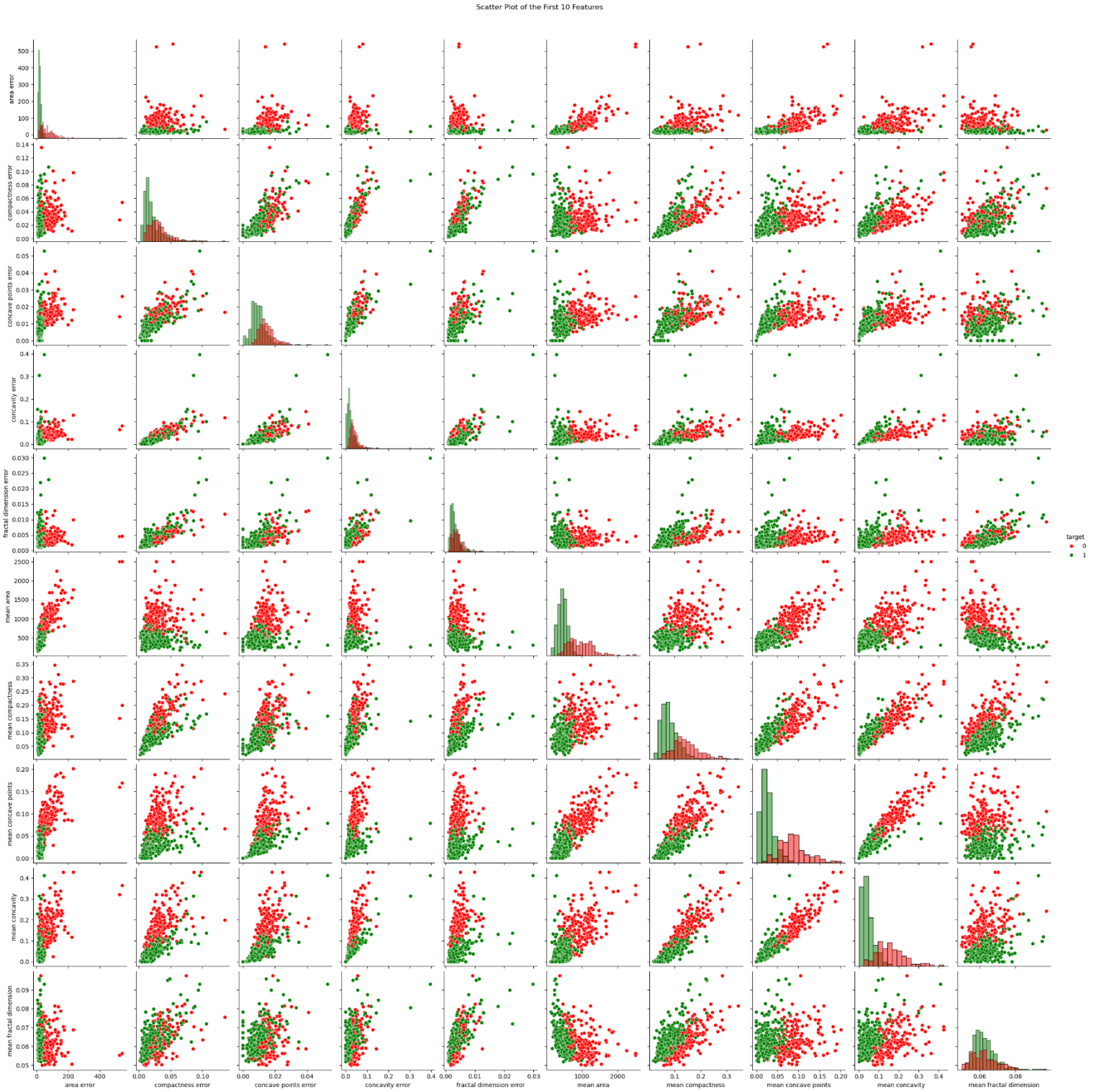
1 First Inspections on the Dataset and Preprocessing

D1 [2 marks] Scatter plot of the first 10 features in the dataset



D2 [2 marks] Comment on scatter plot

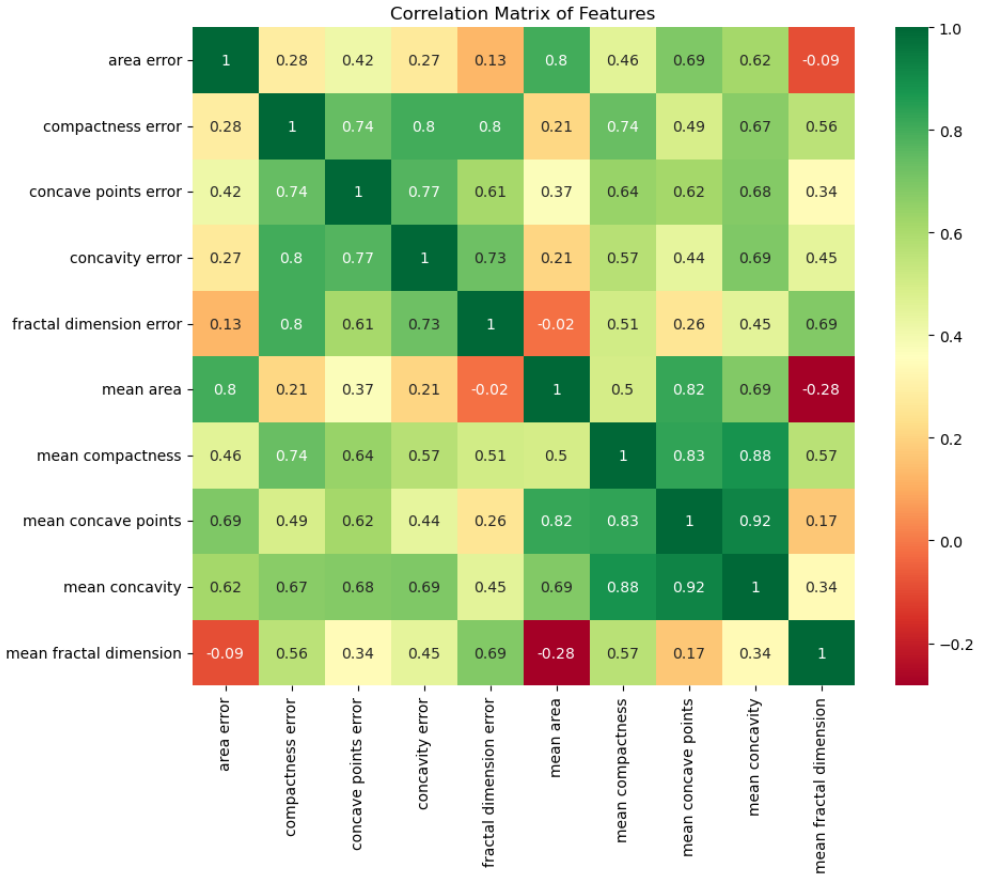
When the dots cluster in a straight line, there is a linear relationship between the two features such as *mean concavity* against *mean concave points* and *mean concavity* against *mean compactness*. When the dots are more scattered, there is no linear relationship between the features such as *mean area* against *mean fractal dimension*.

If the clusters of malignant and benign cases are well-separated, it means that the features are effective in classifying between the two targets. But if the clusters overlap, it suggests some similarity in the feature values between malignant and benign cases, meaning the features are not as effective in classifying between the two targets. Based on the scatter plot, it is observed that there is a significant overlap between the two classes, indicating that this classification task is not linearly separable and requires a more complex approach. A linear model would not be able to perform well in this task.

There are instances with the feature values that are significantly higher or lower than the majority of cases, indicating outliers. For example, there are 2 outliers for malignant cases (red dots) in the scatter plot of *area error* against the other nine features.

If two features are highly correlated and exhibit a strong linear relationship, one of the features could be removed to reduce redundancy. If there are non-linear relationships, it may be important to retain these features even if they are highly correlated with other features. This is because non-linear relationships can capture important patterns in the data that would be lost if the features were removed.

D3 [1 mark] Correlation matrix Heatmap



D4 [1 mark] Do the correlation coefficients support your previous observations?

Yes, the correlation coefficients support my previous observations in D2. *Mean area* versus *mean fractal dimension* observed in D2 to have a non-linear relationship is shown to have a negative correlation. *Mean concavity* versus *mean concave points* and *mean concavity* versus *mean compactness* observed in D2 to have a linear relationship is shown to have a high correlation coefficient value, meaning that these features are highly correlated. Overall, most of the correlation coefficients are not near 1.0, meaning that most features are not highly correlated. Hence, a complex model, and not a linear model, is required for this classification task.

2 Fitting a Decision Tree model with default hyperparameters

D6 [3 marks] Decision Tree model performance results

Training Set Scores
- Accuracy : 1.00
- Precision : 1.00
- Recall : 1.00

Test Set Scores
- Accuracy : 0.96
- Precision : 0.97
- Recall : 0.97

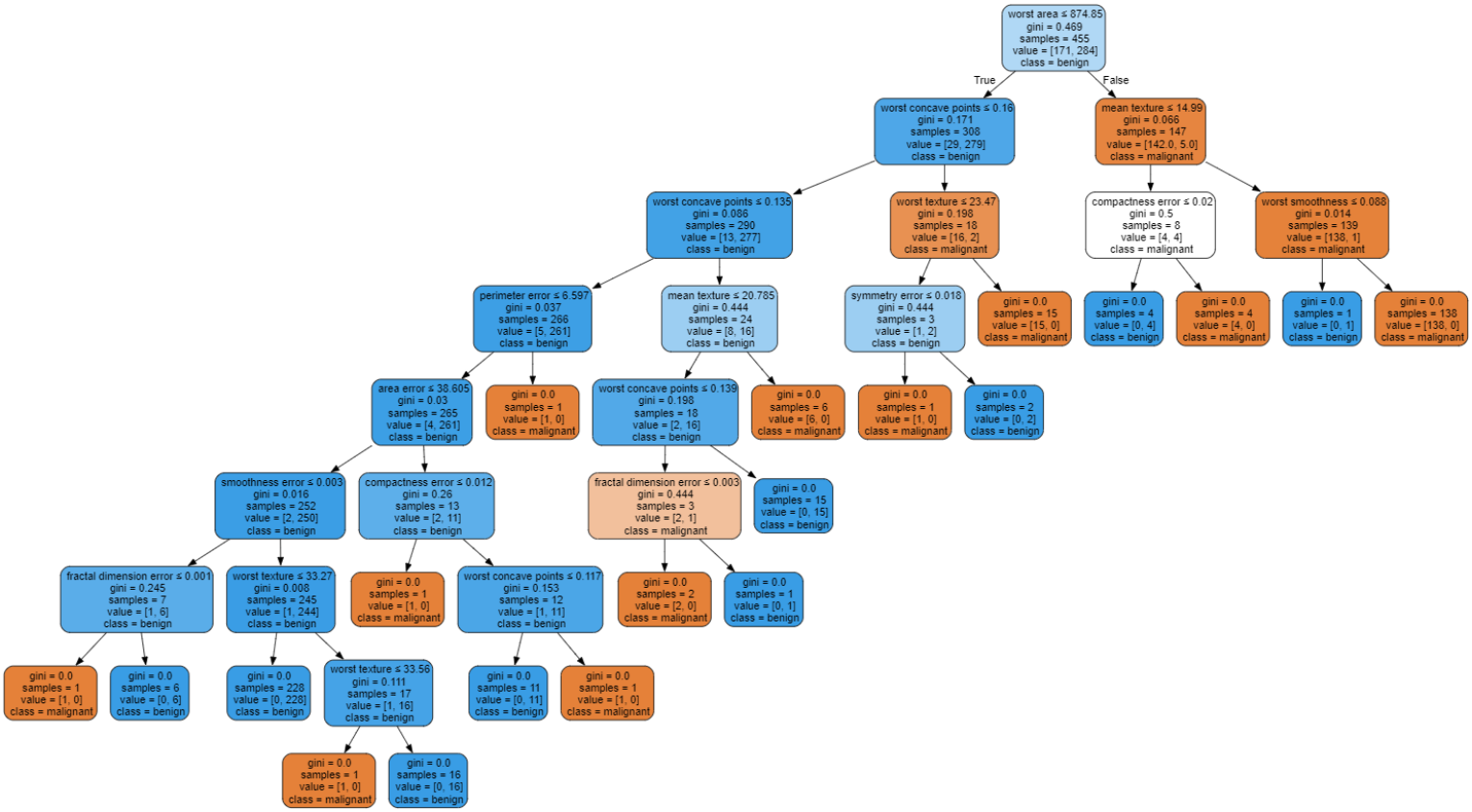
Test Set Confusion Matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	71	2
	Negative	2	39

D7 [2 marks] Comment on above results

The model's performance on the training set with accuracy, precision, and recall scores of 1.0 demonstrates that the model has perfectly predicted all instances in the training set. The perfect accuracy, precision and scores on the training set compared to slightly lower scores on the test set suggest some degree of overfitting. The classifier may have memorised the training data, leading to high performance on the training set but slightly lower performance on the test set. The classifier's performance on the test set is slightly lower but still high. This means that the model is still generalising well to unseen data. If the test set performance were significantly worse than the training set performance, it would indicate strong evidence of overfitting.

D8 [2 marks] Decision Tree diagram



D9 [2 marks] Comment on Decision Tree diagram

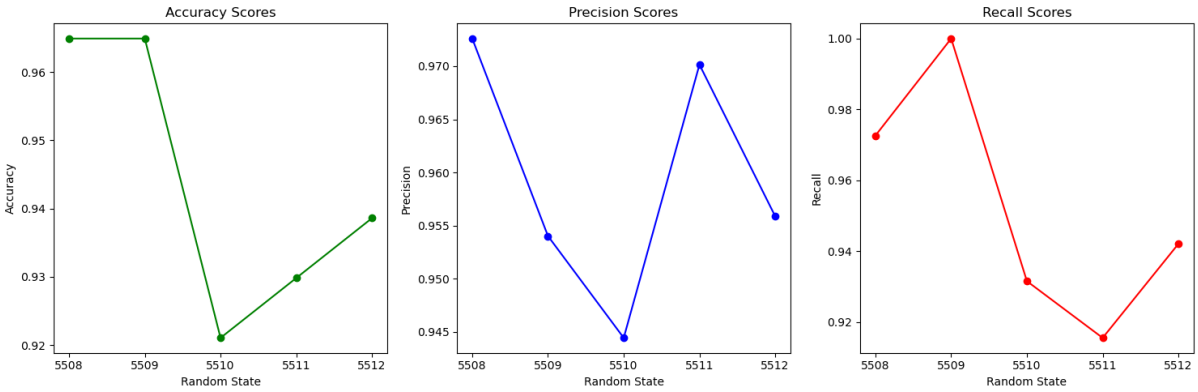
There are 8 levels in the tree diagram.

The diagram shows that the model is overfitting. This is because the tree produced is dense and complex, which means that it is more likely to memorise the training data instead of learning general patterns. It also shows that the model performs perfectly on the training set with Gini values of 0.0.

Each leaf shows the number of samples that fall into that leaf node and the distribution of classes within those samples. The diagram shows that most leaves have values ranging from 1 to 17, suggesting that the model is classifying the dataset into many subsets. This can mean that the model is capturing noise in the training data, causing the model to overfit.

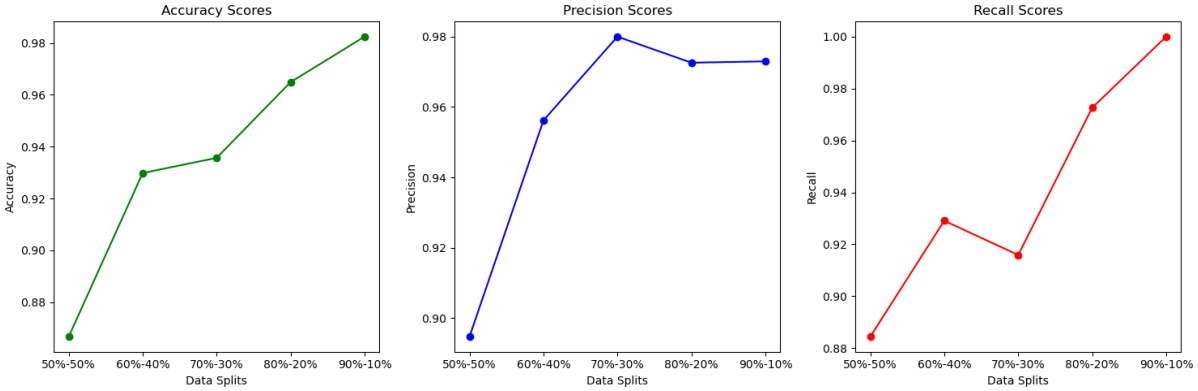
This tree diagram is interpretable and shows the decision-making process clearly from the root node to a lead to understand how the model arrived at the prediction. It also shows the features and thresholds used for decision-making at each node.

D10 [3 marks] Five different data split random states & Decision Tree model scores



The inconsistency in the results of the accuracy, precision, and recall scores on the test set for each split with different random states is expected when using decision trees, which are high variance models. This means that small changes to the data can lead to very different models. Despite this variability, the scores on the test set for models trained on different dataset with random states are relatively high, all above 0.90, indicating that the model still generalises well to unseen data.

D11 [3 marks] Five different data split sizes and their Decision Tree model scores



As the data splits go from 50%-50% to 90%-10%, the accuracy, prevision and recall scores increase. This may be because a larger proportion of data allocated to the training set allows the model to learn from more examples, which is more likely to be representative of the overall dataset. By training on more data, the model is less likely to overfit and is able to generalise better to unseen data in the test set, which results in higher accuracy, precision and recall scores on the test set. However, smaller test sets might not be an accurate indicator of performance as it cannot accurately represent the overall dataset.

3 Fitting a Decision Tree model with optimal hyperparameters

D12 [4 marks] Decision Tree model performance results and optimal parameters

Optimal Hyperparameters
- max_depth : 4
- min_samples_leaf : 5
- min_samples_split : 2

Training Set Scores
- Accuracy : 0.97
- Precision : 0.98
- Recall : 0.98

Test Set Scores
- Accuracy : 0.95
- Precision : 0.97
- Recall : 0.95

Test Set Confusion Matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	69	2
	Negative	4	39

D13 [2 marks] Comment on the impact of fine-tuning hyperparameters

Fine-tuning the hyperparameters of the decision tree resulted in worse performance of the decision tree model compared to using the default hyperparameters. The model with optimal hyperparameters achieved slightly lower but still high performance on the training set and the test set. Fine-tuning did not do what I expected as it did not improve the model's generalisation ability. By restricting the parameters, this may have simplified the model, which can introduce bias that prevented the model from capturing the patterns in the effectively.

D14 [3 marks] Grid search with 10-fold cross-validation with different scoring

Accuracy scoring

Optimal hyperparameters
- max_depth : 4
- min_samples_leaf : 5
- min_samples_split : 2
- Confusion matrix on test set :

		Actual Values	
		Positive	Negative
Predicted Values	Positive	69	2
	Negative	4	39

Precision scoring

Optimal hyperparameters

- max_depth : 5
- min_samples_leaf : 2
- min_samples_split : 5
- Confusion matrix on test set :

		Actual Values	
		Positive	Negative
Predicted Values	Positive	68	2
	Negative	5	39

Recall scoring

Optimal hyperparameters

- max_depth : 3
- min_samples_leaf : 5
- min_samples_split : 2
- Confusion matrix on test set :

		Actual Values	
		Positive	Negative
Predicted Values	Positive	69	3
	Negative	4	38

Fine-tuning the hyperparameters for precision and recall scoring metrics results in different optimal parameter values compared to accuracy scoring. The model with hyperparameters fine-tuned for optimised precision aims to minimise false positives, even if it leads to more false negatives. The model with hyperparameters fine-tuned for optimised recall aims to capture as many positive cases (benign) as possible, even if it leads to more false positives. Considering the problem of diagnosing breast cancer as malignant or benign, optimising the hyperparameters for precision will be better. This is because maximising precision ensures that the model is highly likely to be correct when predicting a positive case (benign), minimising the risk of false positives. In a medical context, it is important to avoid misclassifying a malignant tumour as benign (false positive), as this would have serious consequences for the patient.

4 Fitting a Decision Tree with optimal hyperparameters and a reduced feature set

D15 [1 mark] Feature Importance in descending order

Feature	Importance
worst area	0.781424
worst concave points	0.147249
mean texture	0.036734
mean smoothness	0.018980
area error	0.007651
worst texture	0.005969
mean concavity	0.001992
texture error	0.000000
worst symmetry	0.000000
worst smoothness	0.000000
worst fractal dimension	0.000000

worst concavity	0.000000
concave points error	0.000000
worst compactness	0.000000
concavity error	0.000000
mean area	0.000000
mean compactness	0.000000
smoothness error	0.000000
perimeter error	0.000000
compactness error	0.000000
mean symmetry	0.000000
fractal dimension error	0.000000
mean fractal dimension	0.000000
mean concave points	0.000000
symmetry error	0.000000

D16 [3 marks] Retained and Removed Features

Retained Features:
- mean smoothness
- mean texture
- worst area
- worst concave points

Removed Features:
- area error
- compactness error
- concave points error
- concavity error
- fractal dimension error
- mean area
- mean compactness
- mean concave points
- mean concavity
- mean fractal dimension
- mean symmetry
- perimeter error
- smoothness error
- symmetry error
- texture error
- worst compactness
- worst concavity
- worst fractal dimension
- worst smoothness
- worst symmetry
- worst texture

Total Feature Importance Retained: 0.98

D17 [3 marks] Compare Complete set of features vs Reduced set of features

Decision Tree with Complete Features

Training Set

- Accuracy : 0.97
- Precision : 0.98
- Recall : 0.98

Test Set

- Accuracy : 0.95
- Precision : 0.97
- Recall : 0.95
- Confusion matrix :

		Actual Values	
		Positive	Negative
Predicted Values	Positive	69	2
	Negative	4	39

Decision Tree with Reduced Features

Training Set

- Accuracy : 0.98
- Precision : 0.99
- Recall : 0.98

Test Set

- Accuracy : 0.92
- Precision : 0.97
- Recall : 0.90
- Confusion matrix :

		Actual Values	
		Positive	Negative
Predicted Values	Positive	66	2
	Negative	7	39

D18 [1 mark] Comment on above results

The model trained on the reduced set of features has slightly higher accuracy and precision scores on the training set but lower accuracy and recall scores on the test set, compared to the model trained on the complete set of features. The reduced feature set model predicted slightly more false negatives (7 versus 4) compared to the complete feature set model, meaning it misclassifies more benign tumours as malignant. In the context of diagnosing breast cancer, it is generally safer to misclassify benign tumours as malignant (false positive) than to misclassify malignant tumours as benign (false negative). Therefore, even though the recall score for the reduced feature set model is lower, it is not a significant issue. While the model with the complete set of features performs slightly better overall, the model with the reduced set of features still demonstrates good performance without compromising the patient's health, suggesting that the selected features are informative and can help simplify the model without significantly compromising its effectiveness for this classification task.

5 Fitting a Random Forest

D19 [3 marks] Random Forest model performance results and optimal parameters

Optimal Hyperparameters
- max_depth : 5
- n_estimators : 50

Training Set Scores
- Accuracy : 0.99
- Precision : 0.99
- Recall : 1.00

Test Set Scores
- Accuracy : 0.98
- Precision : 0.99
- Recall : 0.99

Test Set Confusion Matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	72	1
	Negative	1	40

D20 [2 marks] Compare Decision Tree vs Random Forest

The Random Forest model outperforms the Decision Tree model in terms of both training and test set performance, with higher accuracy, precision, and recall scores. The Random Forest model's confusion matrix shows fewer misclassifications (1 false positive and 1 false negative) compared to the Decision Tree model (2 false positives and 4 false negatives). The Random Forest model's performance improvement is expected because it reduces overfitting by averaging the predictions of multiple decision trees trained on different subsets of the data. This ensemble method allows for a more robust model that can capture generalised patterns in the data. It also improves the generalisation ability to unseen data with the random feature selection at each split.

D21 [2 marks] Discussion

Model's ability to generalise well and provide consistent results across different unseen data is crucial for a medical application such as breast cancer diagnosis where misclassifications can have serious consequences. Before these models can be considered reliable for real-world use or before considering the need for a more complex model, they need to be tested on more data to ensure their performance is consistent and reliable across different datasets. Automating the decision process for classifying breast cancer diagnosis with machine learning algorithms can help process large amounts of data quickly with consistent predictions but should still be validated by expert medical knowledge. The dataset should be representative of the target population and include a sufficient number of examples for each class. The dataset used in this assignment had too little data to ensure that the model is trained to handle all cases. The dataset should be balanced in its distribution of classes to avoid bias. Each class (malignant and benign) should be represented in a similar proportion to prevent the model from being skewed towards the majority class.