Analyses of Diagnostic Breast Cancer Data Using Machine Learning and Statistical Methods

Thilini Herath

Contents

Abstract

This study identifies the key features for classifying the two breast cancer types: malignant and benign. The dataset consists of 30 predictor variables, and the most important two features were selected using a combination of techniques, including t-tests and random forest feature importance. This approach yields a logistic regression model with excellent interpretability. Given that malignant is the more severe type, the final model was optimized to maximize the recall for malignant cases, achieving a recall of 97% and an overall accuracy of 95% on the testing data.

1 Introduction

This report presents an analysis of Breast Cancer Wisconsin (Diagnostic) Data Set using statistical and machine learning techniques. The primary objectives are to identify key variables classifying the two cancer types and make an interpretable model to differentiate them. The dataset used in this analysis is sourced from the UCI Repository. The analysis includes exploratory data analysis, supervised learning using Random Forest Classifier, Decision Tree Classifier and logistic Regression tasks.

2 Datasets and Research Questions

The Breast Cancer Wisconsin (Diagnostic) dataset provided by the University of Wisconsin-Madison is to classify whether a tumor is malignant (cancerous) or benign (non-cancerous) based on features derived from cell nuclei. Malignant is the more severe type, and one of the goals of this analysis is to correctly identify malignant cases, as they are severe. In the classification task, malignant cases are treated as the positive class. The dataset contains 569 samples and 30 numerical features describing tumor characteristics, along with an identifier column and the diagnosis label. The features are derived from digital images of fine needle aspirates (FNA) of breast masses. Each feature is calculated from the cell nuclei in the image and includes:

Feature Name	**Description**
Radius	Mean of distances from the center to points on the perimeter.
Texture	Standard deviation of gray-scale values.
Perimeter	Length of the tumor boundary.
Area	Size of the tumor.
Smoothness	Variations in the smoothness of the boundary.
Compactness	$\text{perimeter}^2/\text{area} - 1.0$
Concavity	Severity of concave portions of the contour.
Concave points	Number of concave portions of the contour.
Symmetry	Symmetry of the tumor.
Fractal dimension	Coastline approximation.

Table 1: Feature descriptions for tumor characteristics.

2.1 Research Questions

The research will gather detailed information on the following questions:

- Which features among the 30 numerical attributes contribute most significantly to the classification of tumors as malignant or benign and can we identify the top-performing features to build a simplified and interpretable model?
- How can feature importance scores (e.g., from Random Forest or Logistic Regression) help interpret the biological relevance of tumor characteristics like radius, texture, and concavity in determining malignancy?
- How do different machine learning algorithms (e.g., Random Forest, Decision Tree, Logistic Regression) compare in terms of performance metrics such as accuracy, precision, recall, F1-score, and ROC?
- What is the optimal decision threshold for classification models to minimize false negatives (malignant tumors classified as benign) while maintaining a reasonable false positive rate?

3 Exploratory Data Analysis (EDA)

3.1 Class Distribution Analysis

The histogram plots in Figure 1 organize the 30 columns into groups of related feature types (e.g., radius, texture, perimeter). Each group contains 3 similar feature names. Visualizing the distribution of each feature for the two diagnosis categories (M = Malignant, B = Benign) helps identify patterns or differences in feature distributions between the two diagnosis types.
Most features show clear distinctions in their distributions between the two categories. Malignant cases (orange) tend to have higher values for many features such as radius, perimeter, area, compactness, concavity, and concave points.
For features like radius, texture, and area, the distributions for Malignant tumors are shifted towards larger values compared to Benign tumors. Features like smoothness, symmetry, and fractal dimension exhibit overlap in the distributions, indicating these may be less effective in separating the two categories.
This visualization indicates which features may be more discriminative for classification models, with features like radius, area, and concave points standing out as strong predictors for malignancy. Next, we explore this further using two-sample t-tests.

3.2 t-test on Predictor's Distribution

Table 2 shows the results of two-sample t-tests comparing the means of features between the two groups (M = Malignant and B = Benign) in the dataset. The table includes only the first five and last five features of the ordered test results based on the t-statistic.
t-statistic: A measure of the difference between the means of the two groups relative to the variability in the data.
Larger absolute values indicate stronger evidence of a difference.
p-value: A small p-value (typically < 0.05) indicates that the difference in means is statistically significant.
CI Lower & CI Upper: The 95% confidence interval for the difference in means. If the interval does not include 0, the difference in means is considered statistically significant. Features like concave-points3, perimeter3, radius3 and concave-points1 have extremely low p-values, indicating strong evidence of differences between M and B groups. Their confidence intervals do not include 0, further confirming the significance.

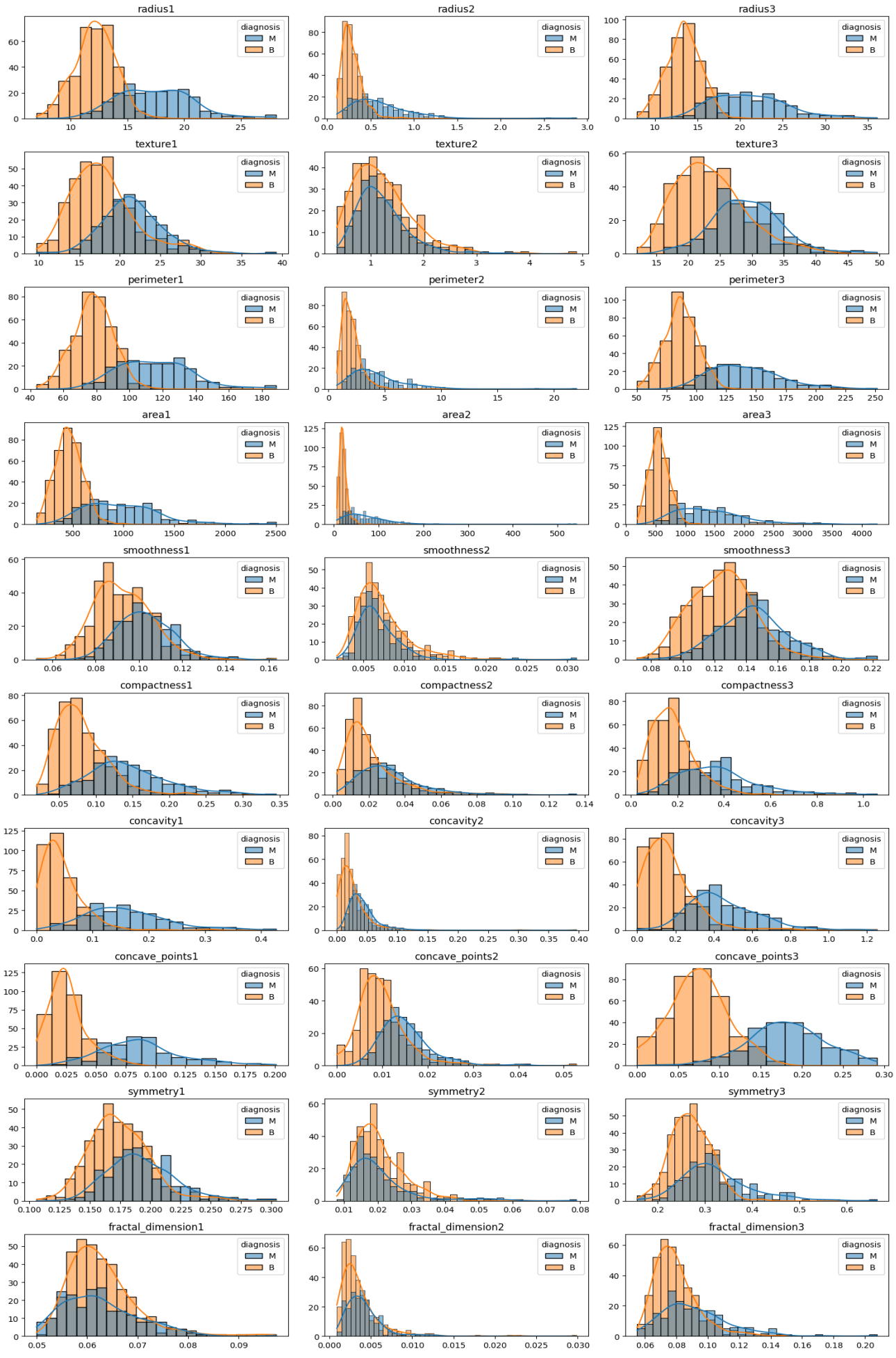

Figure 1: Class Distribution of Non-Fraudulent and Fraudulent Transactions

Column	T-statistic	P-value	CI Lower	CI Upper
concave_points3	29.117659	1.06×10^{-96}	0.100513	0.115073
perimeter3	25.332210	1.03×10^{-72}	50.138875	58.589909
concave_points1	24.844810	3.13×10^{-71}	0.057338	0.067208
radius3	24.829745	3.56×10^{-71}	7.140056	8.369964
perimeter1	22.935314	1.02×10^{-66}	34.089743	40.490200
fractal_dimension2	2.036236	0.042202	0.000015	0.000838
symmetry2	-0.142055	0.887122	-0.001654	0.001431
texture2	-0.207865	0.835417	-0.098927	0.079996
fractal_dimension1	-0.296866	0.766722	-0.001428	0.001053
smoothness2	-1.622869	0.105297	-0.000919	0.000088

Table 2: T-test Analysis: Top 5 and Bottom 5 features

Features such as concave-points, radius, perimeter, and area show the strongest statistical significance and large effect sizes, making them the most promising predictors for classification. In contrast, features related to symmetry and fractal dimension appear less useful for distinguishing between the two categories.

We further evaluate feature importance using a Random Forest classifier, which determines feature importance based on the predictive power of its decision trees.

4 Random Forest Classifier

Table 3 presents the results of a Random Forest Classifier applied to the dataset to classify diagnoses (B for benign and M for malignant tumors). The model was trained using 70% of the data and tested on the remaining 30%, while preserving the class imbalance.
Model accuracy score is 0.9649. This computes the proportion of correctly predicted diagnoses out of all predictions. The model achieves 96.49% accuracy on the test set, meaning that 96.49% of the predictions are correct.

Metric	Benign (B)	Malignant (M)	Accuracy	Macro Avg	Weighted Avg
Precision	0.95	0.98	-	0.97	0.97
Recall	0.99	0.92	-	0.96	0.96
F1-Score	0.97	0.95	0.96	0.96	0.96
Support	107	64	171	171	171

Table 3: Classification Report Table

Below, we provide detailed explanations of each the metrics in the Table 3:

- **Precision:** The proportion of true positive predictions out of all positive predictions. 95% of the samples predicted as benign are actually benign. 98% of the samples predicted as malignant are actually malignant.
- **Recall:** The proportion of actual positive samples correctly identified. 99% of the actual benign samples are correctly identified. 92% of the actual malignant samples are correctly identified.
- **F1-Score:** The harmonic mean of precision and recall. A high F1-score indicates excellent performance for these classes.

- **Support:** The number of actual occurrences for each class (B and M). There are 107 benign samples in the test set. There are 64 malignant samples in the test set.
- **Accuracy:** 96% of all predictions are correct.
- **Macro Avg:** Average performance across both classes, treating each class equally.
- **Weighted Avg:** Average performance weighted by the number of samples in each class.

Conclusion
The Random Forest Classifier performs exceptionally well, achieving high accuracy and strong precision, recall, and F1-scores for both classes. The slight imbalance in recall (0.99 for B vs. 0.92 for M) suggests that the model might be slightly biased toward the majority class (B). The overall results demonstrate that the model is effective at distinguishing between benign and malignant tumors.

4.1 Confusion Matrix

The image in Figure 2 shows the confusion-matrix for the Random Forest Classifier. It contains counts of:

- **True Positives (TP):** Correct predictions for Malignant (M). The model correctly predicted 59 malignant (M) tumors as malignant.
- **True Negatives (TN):** Correct predictions for Benign (B). The model correctly predicted 106 benign (B) tumors as benign.
- **False Positives (FP):** Benign predicted as Malignant. The model incorrectly predicted 1 benign (B) tumor as malignant (M).
- **False Negatives (FN):** Malignant predicted as Benign. The model incorrectly predicted 5 malignant (M) tumors as benign (B).

Most predictions are correct: 106+59=165 out of 171 total samples. Only 6 misclassifications (1+5) occurred.

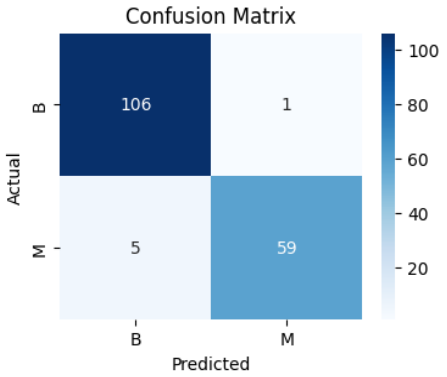

Figure 2: Confusion-Matrix for Random Forest Classifier

4.2 Feature Importance

The image in Figure 3 evaluate the feature importance of variables in the dataset using a Random Forest Classifier. It identifies which features contribute the most to predicting whether a tumor is benign (B) or malignant (M).
The five most important features are:

Feature Name	Importance
perimeter3	0.147270
area3	0.144130
concave-points1	0.135878
radius3	0.131583
concave-points3	0.107682

Table 4: Feature importance values from the Random Forest model.

Notably, four of these features, namely, perimeter3, concave-points1, concave-points3, and radius3 also rank among the top five features identified by the t-test. This alignment suggests that the Random Forest feature importance is somewhat consistent with the results of the t-tests. Since the t-test evaluates mean differences between cancer types, it can be argued that the Random Forest model captures this significance through the geometric properties of the features.

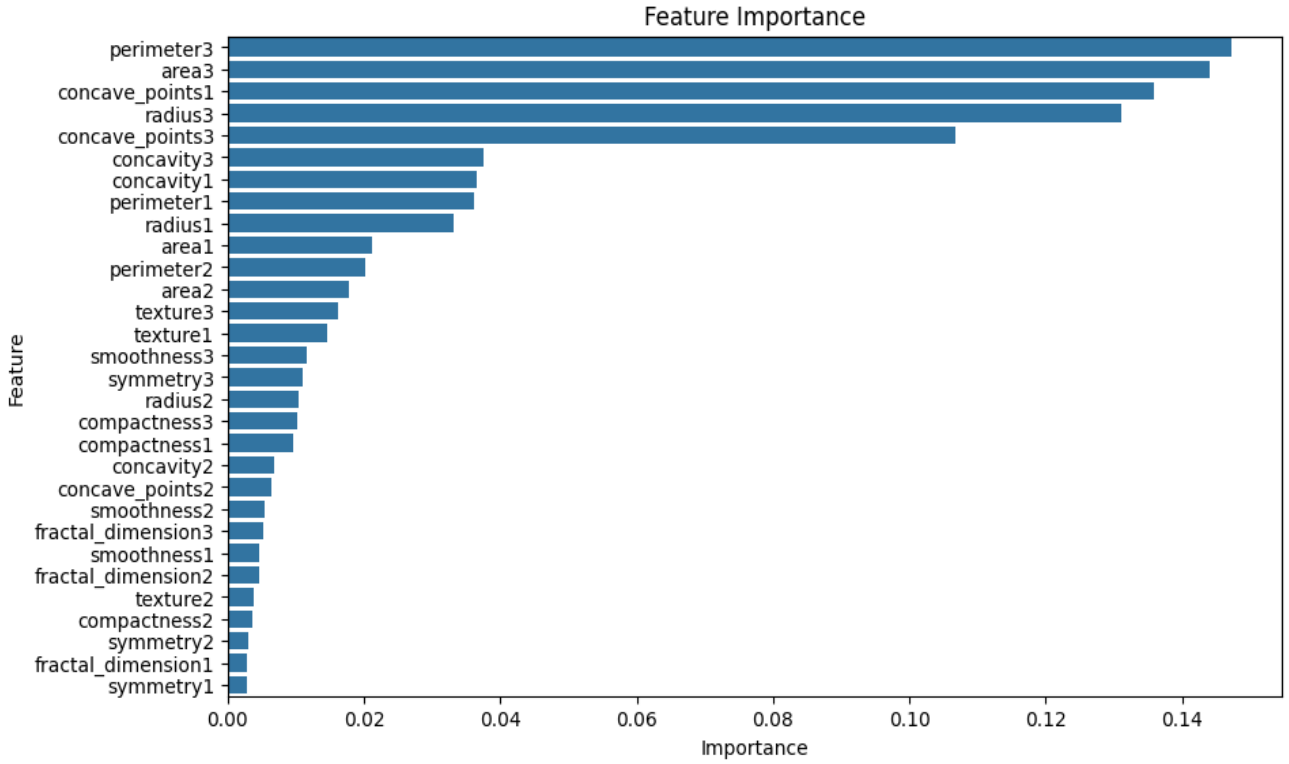

Figure 3: Feature Importance

5 Decision Tree Classifier

The image in Figure 4 shows the Decision Tree Classifier. Since Random Forest creates many decision trees based on bootstrap samples, next we create a single decision tree with the selected five variables (based on variable selection) with a tree-depth of 2 levels.
The first decision split in the tree is based on the variable perimeter3, using a cut-off point of 113.15, which is identified as the most important variable in the Random Forest model (see Table 4) and the second most significant variable from t-tests. At this split, 243 benign observations have a perimeter3 value less than or equal to 113.15, while only 7 observations have a perimeter3 value greater than 113.15. This split is critical, as it places 97.2% of the benign observations on one side. Consequently, the decision tree has a structure where the majority of the left side is dominated by benign observations, while the right side is predominantly malignant. Further splits in the tree continue to classify observations into more distinct groups, eventually leading to a terminal node that contains only malignant observations (further to the right in the tree).

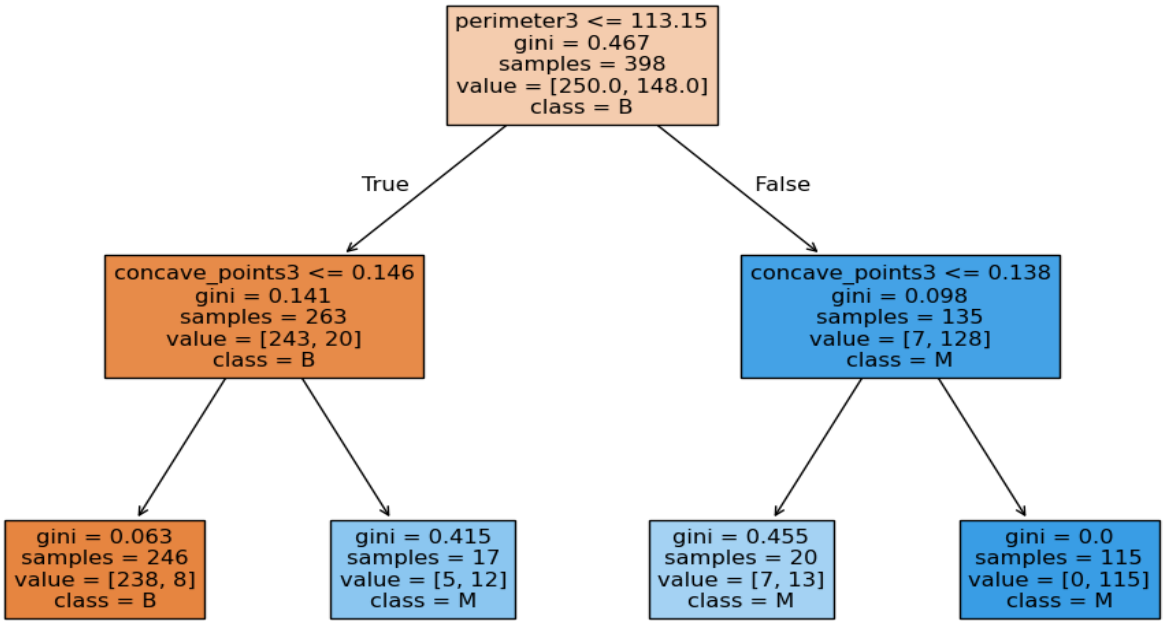

Figure 4: Decision Tree Classifier

Table 5 presents the results of a Decision Tree Classification Report applied to the dataset to classify diagnoses (B for benign and M for malignant tumors).

Metric	Benign (B)	Malignant (M)	Accuracy	Macro Avg	Weighted Avg
Precision	0.94	0.88	-	0.91	0.92
Recall	0.93	0.91	-	0.92	0.92
F1-Score	0.93	0.89	0.92	0.91	0.92
Support	107	64	171	171	171

Table 5: Decision Tree Classification Report Table

Conclusion
The model was trained using the same training and testing data as the Random Forest model in Section 4 to ensure a fair comparison of performance between the models. Despite using only two features, this decision tree classifier achieves an overall accuracy of 91.81%, correctly classifying 93% of benign observations and 91% of malignant observations. Given its higher interpretability compared to the Random Forest model, and with only a 4.68% reduction in overall accuracy, this decision tree model is a strong candidate for studying feature contributions in cancer treatment.

5.1 Confusion Matrix

Most predictions are correct: 99+58=157 out of 171 total samples. Only 14 misclassifications (6+8) occurred.

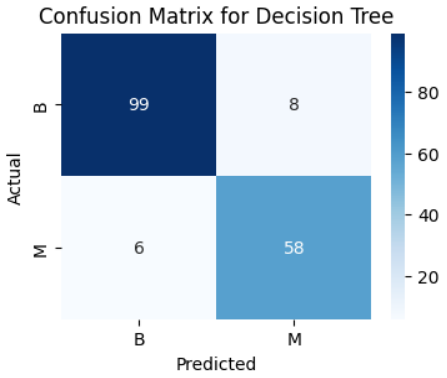

Figure 5: Confusion-Matrix for Decision Tree Classification

6 Logistic Regression

6.1 logistic regression - using 5 selected features

The logistic regression model is to classify tumor diagnoses as benign (B) or malignant (M) using selected features (perimeter3, area3, concave-points1, radius3, concave-points3). The model was trained using the same training and testing data as in Sections 4 and 5 to ensure a fair comparison of performance between the models.

Key Metrics:

- **Dependent Variable:** The target (diagnosis), mapped to 0 and 1 (0 = Benign, 1 = Malignant).
- **Pseudo R-squared:** 0.8344, a measure of how well the model explains variability in the response variable.
- **Log-Likelihood:** Indicates model fit (higher is better).

Coefficients Table:

	coef	std err	z	$P > \|z\|$	[0.025, 0.975]
const	-1.6546	8.060	-0.205	0.837	[-17.452, 14.142]
perimeter3	-0.0614	0.090	-0.681	0.496	[-0.238, 0.115]
area3	0.0278	0.011	2.488	0.013	[0.006, 0.050]
concave_points1	10.4043	21.676	0.480	0.631	[-32.080, 52.889]
radius3	-1.3450	1.156	-1.163	0.245	[-3.611, 0.921]
concave_points3	50.9876	13.311	3.831	0.000	[24.899, 77.076]

- **Coefficients:** Represent the effect of a 1-unit increase in each feature on the log-odds of the target being Malignant (M).
- **Statistically Significant Features:** Features with $p < 0.05$, such as area3 and concave-points3, are statistically significant predictors of the target.

Metric	Benign (B)	Malignant (M)	Accuracy	Macro Avg	Weighted Avg
Precision	0.95	0.97	-	0.96	0.95
Recall	0.98	0.91	-	0.94	0.95
F1-Score	0.96	0.94	0.95	0.95	0.95
Support	107	64	171	171	171

Table 6: Random Forest Classification Report Table

According to the logistic regression model, the only statistically significant features are area3 and concave-points3. Notably, the concave-points3 feature is also identified by the Decision Tree classifier as a key predictor of the response class. Additionally, the area3 feature is ranked as the second most important feature in the Random Forest model's feature selection. This suggests that the logistic regression model relies on a slightly different combination of features compared to the Decision Tree. Next, we delve deeper into this combination to understand how these two variables capture and predict the response variable.

The model achieves an overall accuracy of 95%, higher than the accuracy of the decision tree model. High precision, recall, and F1-scores for both classes suggest the model is effective. Slightly lower recall for malignant tumors (91%). This indicates some malignant cases are misclassified.

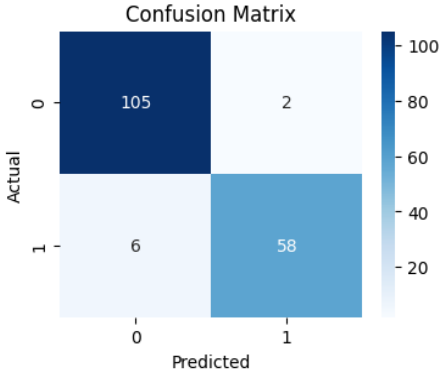

Figure 6: Confusion-Matrix for Logit Regression

6.2 logistic regression - using 2 selected features

Next we develop a logistic regression model for predicting tumor diagnosis (benign (B) or malignant (M)) only using area3 and concave-points3 features.

Key Metrics:

- **Dependent Variable:** Diagnosis, encoded as 0 (Benign) or 1 (Malignant).
- **Pseudo R-squared:** 0.8288, indicating the model explains approximately 82.88% of the variability in the data.
- **Log-Likelihood:** A measure of model fit (higher is better).

Coefficients Table:

Variable	Coef	Std Err	z	P> |z|	[0.025]	[0.975]
const	-15.1198	2.206	-6.855	0.000	-19.443	-10.797
area3	0.0104	0.002	5.331	0.000	0.007	0.014
concave_points3	47.5758	8.753	5.435	0.000	30.420	64.731

Both area3 and concave-points3 are statistically significant predictors ($P < 0.05$).
The model achieves 94.15% accuracy, meaning 94.15% of predictions are correct. High precision, recall, and F1-scores for both classes suggest the model is effective. Misclassifications are relatively few (6 benign → malignant and 2 malignant → benign).

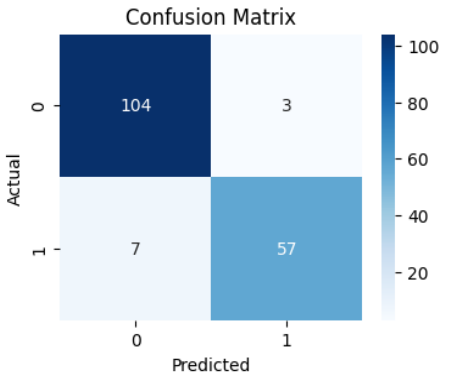

Figure 7: Confusion-Matrix for Logit Regression with a threshold of 0.50

Metric	Benign (B)	Malignant (M)	Accuracy	Macro Avg	Weighted Avg
Precision	0.94	0.95	-	0.94	0.94
Recall	0.97	0.89	-	0.93	0.94
F1-Score	0.95	0.92	0.94	0.94	0.94
Support	107	64	171	171	171

This model correctly identifies 89% of malignant observations, which is crucial given the greater severity of malignant cases compared to benign ones. Notably, this level of accuracy is achieved using only two features. The model significantly enhances interpretability compromising only 1% of overall accuracy.
The model equation is

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -15.1198 + 0.0104 \cdot \text{area3} + 47.5758 \cdot \text{concave\_points3}$$

- $P(Y=1)$: Probability of the outcome being 1 (malignant).
- $P(Y=0)$: Probability of the outcome being 0 (benign).

Here, we use a probability threshold of 0.5 for classification. This means that if the predicted probability for an observation is greater than or equal to 0.5, it is classified as malignant; otherwise, it is classified as benign.
Next, we examine the effect of the threshold on model performance. Our objective is to improve the model's ability to recognize the malignant class. To achieve this, we utilize a Receiver Operating Characteristic (ROC) curve.

7 Receiver Operating Characteristic (ROC) curve

We analyze how varying the threshold impacts the model's performance, focusing on enhancing its effectiveness in identifying the malignant class. To address this, we use a Receiver Operating Characteristic (ROC) curve and the F1-score to identify the optimal threshold.
Key Concepts

- **ROC Curve:** A graphical representation of the trade-off between True Positive Rate (TPR) (sensitivity/recall) and False Positive Rate (FPR) at various classification thresholds.
- **AUC (Area Under the Curve):** Measures the model's ability to distinguish between classes. A value closer to 1.0 indicates excellent classification performance.
- **F1-Score:** The harmonic mean of precision and recall. Useful for assessing the model's balance between false positives and false negatives.

Based on the ROC curve shown in Figure 8, we derive the following insights

- An AUC of 0.98 suggests excellent discriminatory ability between the classes.
- Thresholds around 0.32–0.50 yield the best F1-scores (0.92–0.93), effectively balancing precision and recall.
- Lower thresholds (e.g., 0.30) improve recall (sensitivity) but reduce precision, leading to lower F1-scores.
- Higher thresholds (e.g., 0.70) increase precision but reduce recall.

Threshold	0.82	0.81	0.70	0.62	0.51	0.50	0.48	0.47	0.46	0.38	0.32	0.10	0.09
F1-score	0.88	0.89	0.91	0.92	0.92	0.93	0.93	0.93	0.92	0.93	0.93	0.91	0.90

Table 7: Thresholds and corresponding F1-scores.

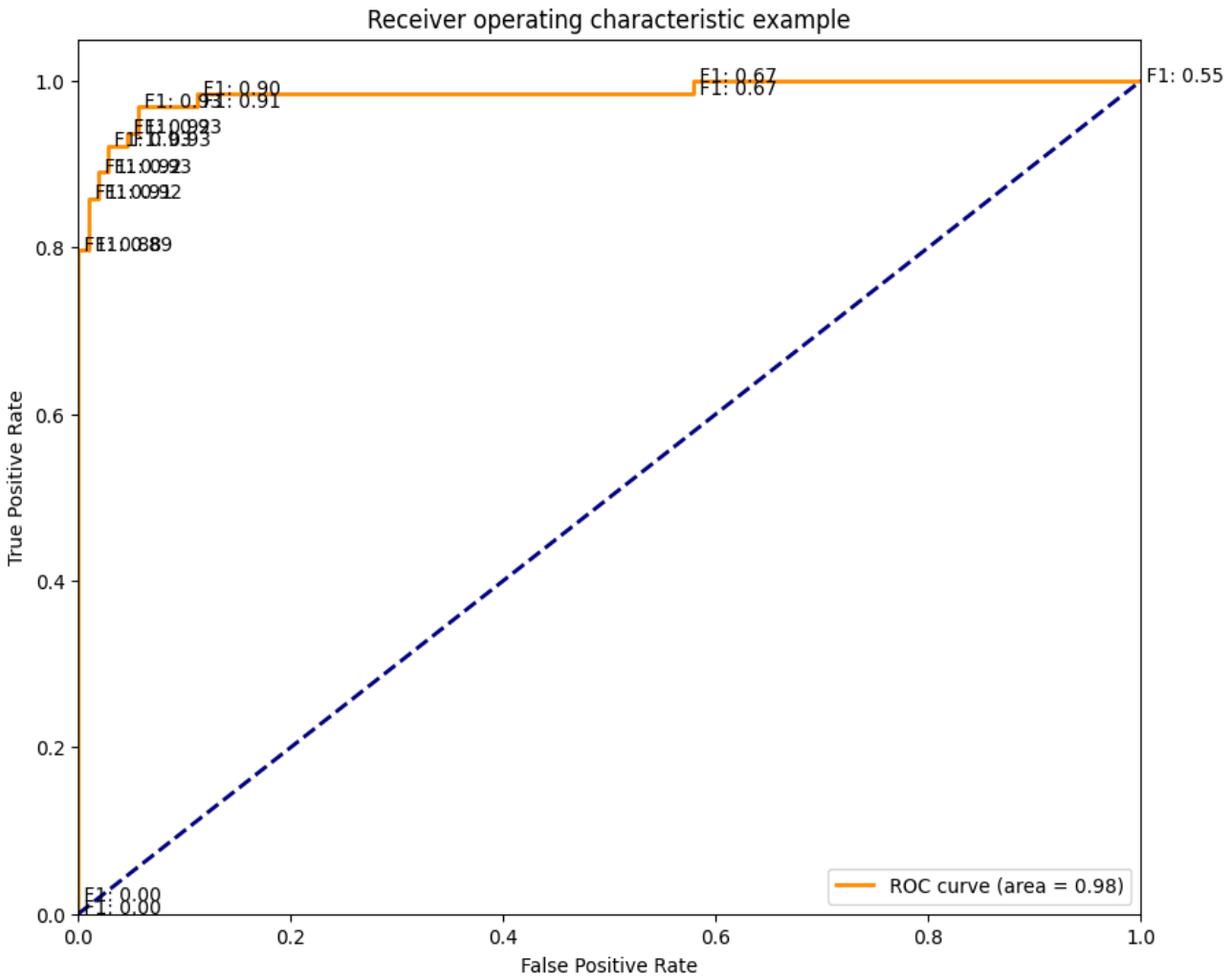

Figure 8: Receiver operating characteristic example

We choose a threshold of 0.32, allowing the model to classify more observations into the malignant class while still safely maintaining accurate classification of the benign class. We obtain the following improved classification report and confusion matrix for a threshold of 0.32.

Metric	**Benign (B)**	**Malignant (M)**	**Accuracy**	**Macro Avg**	**Weighted Avg**
Precision	0.98	0.91	-	0.95	0.95
Recall	0.94	0.97	-	0.96	0.95
F1-Score	0.96	0.94	0.95	0.95	0.95
Support	107	64	171	171	171

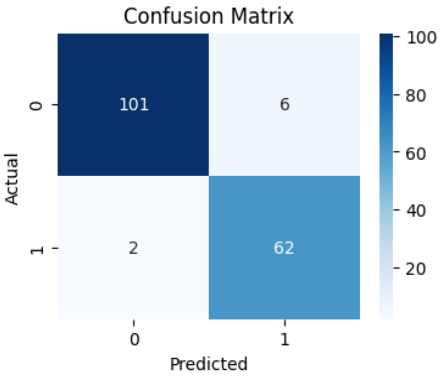

Figure 9: Confusion-Matrix for Logit Regression with a threshold of 0.32

The model classifies 97% of malignant cases, an 8% improvement achieved through a threshold of 0.32 compared to the standard threshold of 0.5. The overall accuracy increases to 95%. In conclusion, this approach provides a model capable of distinguishing malignant from benign cases with excellent interpretability.