MSBD566 – Predictive Modeling and Analytics

Midterm Project: Report

Tiffany North-Reid, MPH, PhD Y2 Student
Meharry Medical College
October 22, 2025

Project Description

The purpose of this project is to classify breast tumor samples as malignant or benign based on diagnostic features derived from fine-needle aspirate (FNA) images. This task is clinically important because accurate and automated classification can assist oncologists in early detection, thereby improving patient outcomes and reducing unnecessary biopsies. The project applies a Logistic Regression classifier to perform binary classification. The model outputs the probability that a tumor sample is malignant, enabling threshold-based decision making for medical risk assessment.

Data Description

The *Breast Cancer Wisconsin (Diagnostic)* dataset accessible via Kaggle link is a well-established medical dataset commonly used for machine learning classification research. It contains diagnostic measurements collected from digitized images of fine-needle aspirate (FNA) samples of breast masses. Each record represents one patient and includes a set of numeric features that describe the visual and geometric characteristics of cell nuclei. The dataset contains a total of 569 observations, of which 212 correspond to malignant tumors and 357 to benign ones. There are 30 numeric predictor variables, one target variable (`diagnosis`), and one identification number for each sample. The dataset is de-identified and freely available for educational and research purposes. It contains no protected health information (PHI). From a technical standpoint, the dataset is clean and ready for analysis, with no missing values and all features normalized to comparable scales.

Method and Analysis

(i) Method Used

The analytical approach for this project was based on **Logistic Regression**, a probabilistic classification method that models the relationship between a binary response

variable and multiple predictor variables. The model estimates the probability that a tumor sample is malignant by fitting a linear combination of input features through the log-odds transformation:

$$\log \frac{p}{1-p} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j,$$

where p represents the probability of malignancy, x_j denotes each diagnostic feature, and β_j are the coefficients learned during training. The implementation used the `LogisticRegression` function from the `scikit-learn` library with parameters `solver='liblinear'`, `penalty='l2'`, and `max_iter=1000`. Prior to training, all numeric features were standardized using the `StandardScaler` method to ensure that each variable contributed equally to the model and that coefficient magnitudes were directly comparable. The data was divided into 80% training and 20% testing subsets, with stratification applied to preserve the class balance of malignant and benign samples.

(ii) Why This Method Was Chosen

Logistic Regression was chosen because it provides a balance of interpretability, computational efficiency, and predictive reliability. In contrast to more opaque machine learning algorithms, Logistic Regression offers transparent insight into the relative influence of each variable on the probability of malignancy. The sign and magnitude of each coefficient indicate whether an increase in a given feature raises or lowers the odds of a malignant diagnosis. This interpretability is particularly valuable in a biomedical context, where understanding which features contribute to a diagnosis is as important as the accuracy of the prediction itself. The method is computationally efficient, performs well on tabular data, and produces probabilistic outputs that can be thresholded to suit specific clinical objectives, such as maximizing sensitivity for detecting cancer.

(iii) Variables or Features Used

All thirty diagnostic measurements in the dataset were used as input features. These measurements represent morphological and textural characteristics of cell nuclei derived from digitized fine-needle aspirate (FNA) images of breast masses. The attributes include metrics such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each summarized through their mean, standard error, and worst (maximum) values. The target variable, `diagnosis`, was encoded numerically as 1 for malignant and 0 for benign. Before model fitting, all features were standardized to have zero mean and unit variance.

(iv) Results, Tables, and Visualizations

The Logistic Regression classifier achieved outstanding predictive performance on the test dataset. The overall accuracy was approximately 97.4%, indicating that nearly all samples were correctly classified. The precision score reached 0.976, the recall was 0.952, and the resulting F1-score was 0.964. The area under the Receiver Operating Characteristic (ROC) curve (AUC) was 0.996, demonstrating an almost perfect ability to distinguish between benign and malignant cases. These results are summarized in Table 1.

Table 1: Test-set performance for Logistic Regression.

Model	Accuracy	Precision	Recall	F1	ROC AUC
Logistic Regression	**0.9737**	0.9756	0.9524	0.9639	**0.9960**

To provide a more comprehensive view of the model's behavior and diagnostic capability, several visualizations were produced and are presented in Figures 1–5. Figure 1 shows the ROC curve, which demonstrates near-perfect separation between malignant and benign samples. The confusion matrix in Figure 2 reinforces this result by showing that the model made very few misclassifications, with only a small number of false negatives observed among the malignant cases.

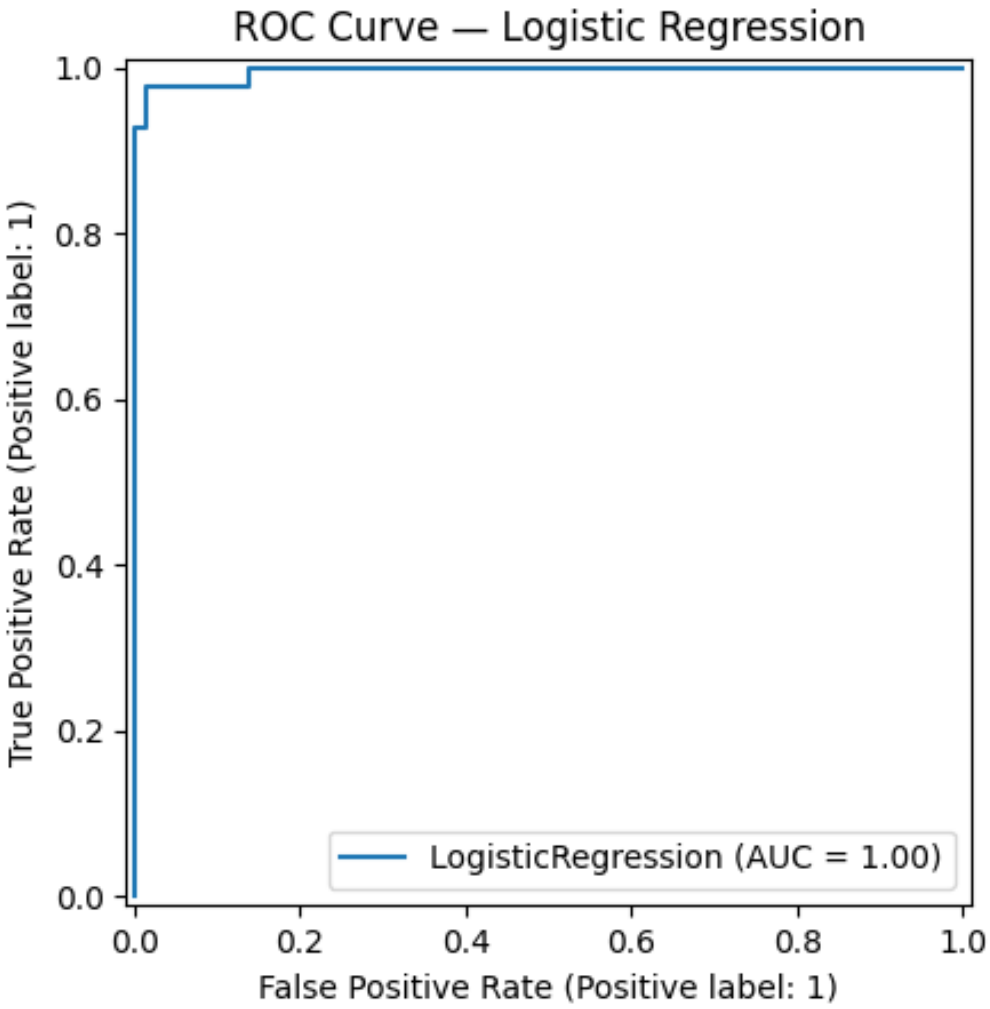

Figure 1: ROC Curve — Logistic Regression

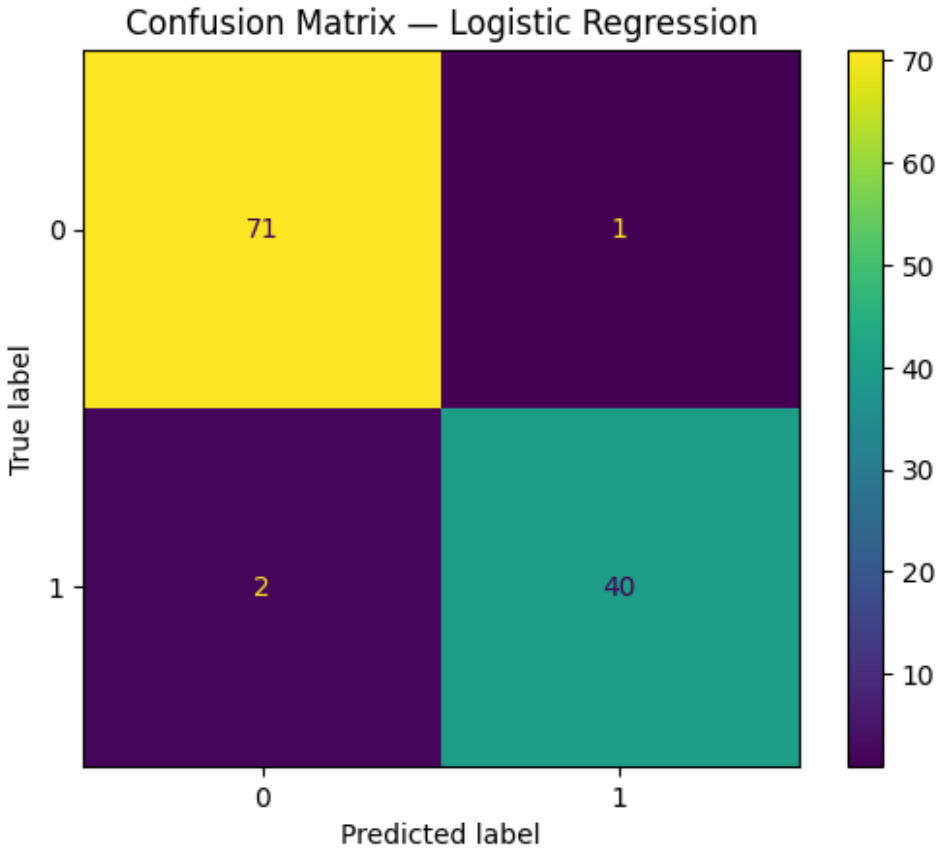

Figure 2: Confusion Matrix — Logistic Regression

Feature interpretability was examined through the absolute magnitudes of the model coefficients, shown in Figure 3. This plot highlights that diagnostic features describing tumor size and shape irregularity—particularly `worst_radius`, `worst_perimeter`, and `worst_concave_points`—had the strongest positive association with malignancy. This finding aligns with established medical understanding that malignant tumors tend to exhibit greater nuclear irregularity and larger overall size.

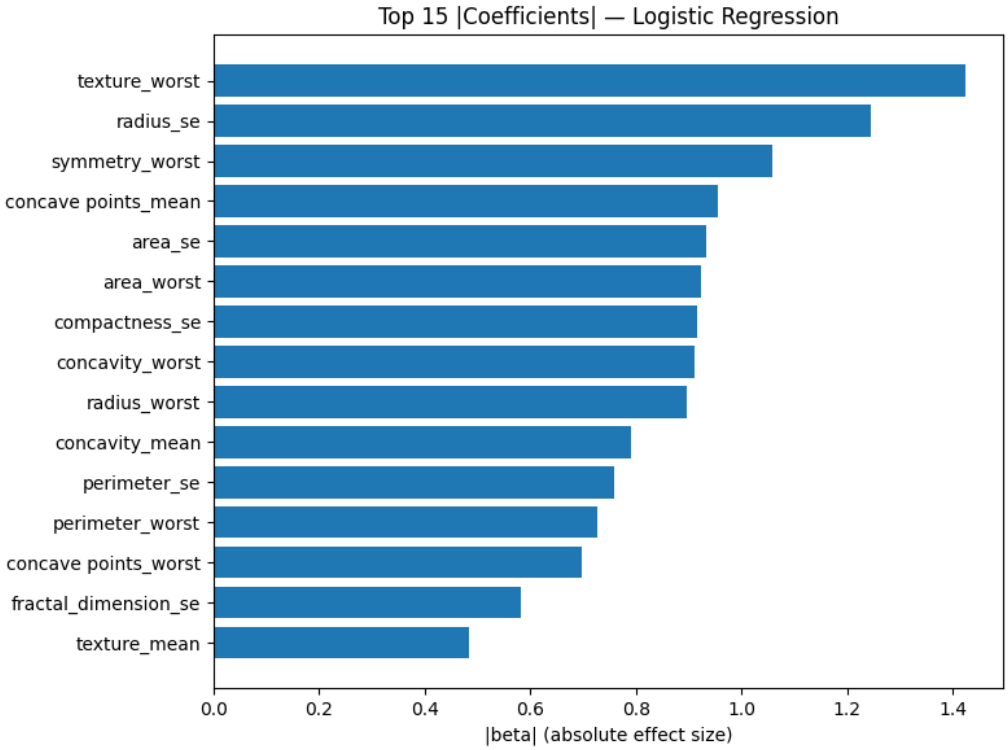

Figure 3: Top 15 $|\beta|$ Coefficients — Feature importance for interpretability

Figures 4 and 5 illustrate how classification thresholds affect model precision and recall. The precision curve (Figure 4) shows how increasing the decision threshold enhances confidence in malignant predictions, while the recall curve (Figure 5) demonstrates how lowering the threshold improves the model's sensitivity.

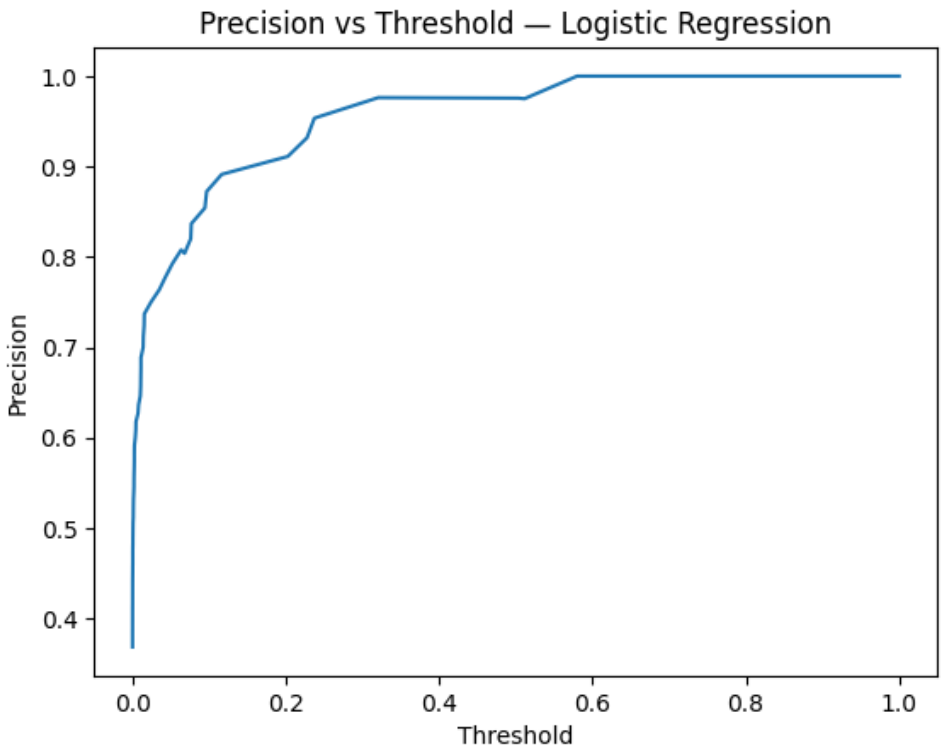

Figure 4: Precision vs Threshold — Trade-off between threshold and precision

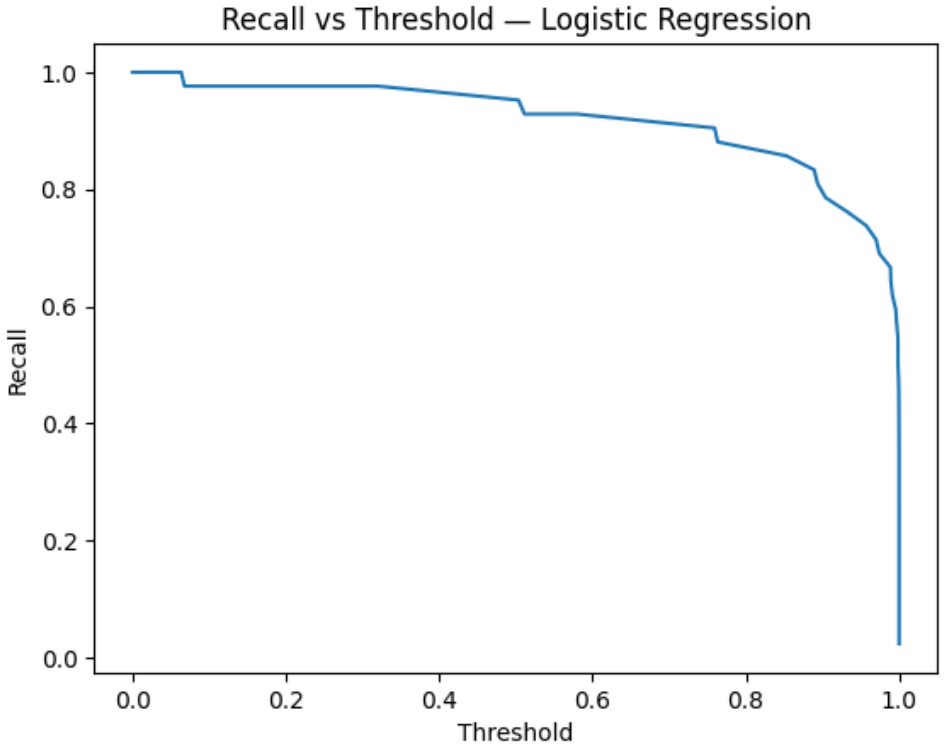

Figure 5: Recall vs Threshold — Increasing sensitivity by lowering threshold

Feature Pruning for Interpretability. To assess model efficiency, the classifier was

retrained using only the top $|\beta|$ features. As shown in Figure 6, performance plateaued after approximately 8–10 predictors: accuracy remained above 96% and ROC AUC near 0.996. This finding indicates that a compact, interpretable subset of predictors captures nearly all the model's predictive power.

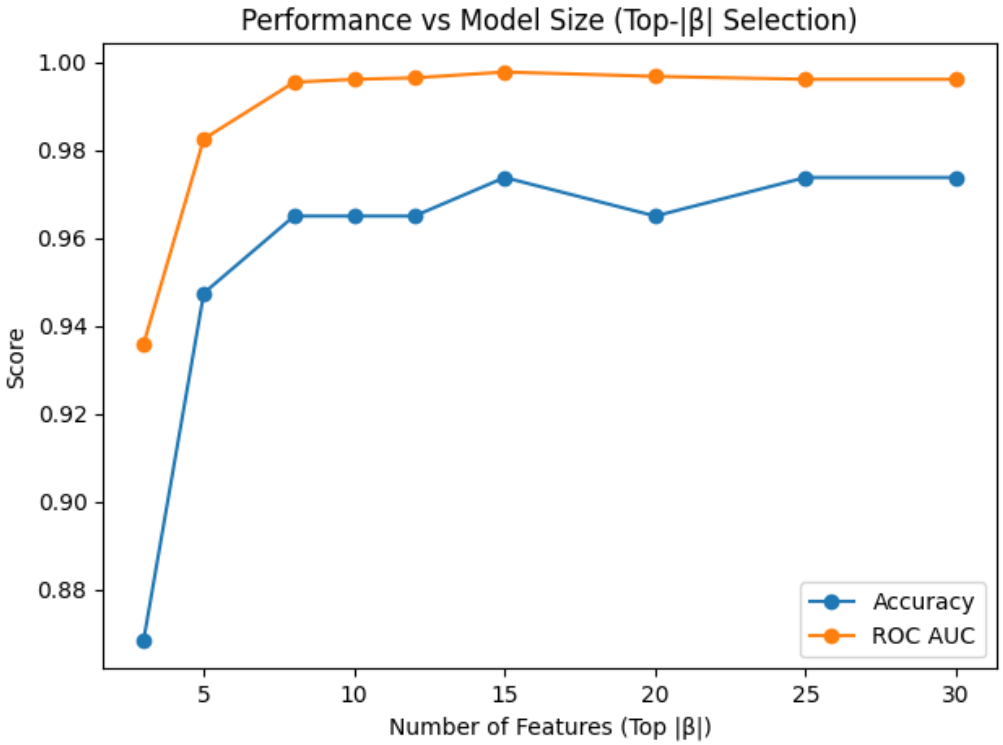

Figure 6: Performance vs Model Size (Top-$|\beta|$ feature selection). Accuracy and AUC plateau after about 8–10 features.

(v) Interpretation and Discussion

The model demonstrated excellent discriminative ability and interpretability. The near-perfect ROC AUC confirms that the classifier can reliably differentiate between benign and malignant tumors. Feature pruning confirmed that the top 8–10 predictors account for nearly all predictive performance, suggesting that a smaller, more interpretable model can perform equivalently to the full feature set. Although a few malignant samples were misclassified as benign, the recall remained above 95%, and the model can easily be tuned to increase sensitivity if required. Overall, Logistic Regression proved to be a robust, transparent, and computationally efficient method for breast cancer diagnosis classification, achieving both predictive accuracy and interpretability in alignment with the course objectives.

Simple Evaluation

(i) Performance Assessment

The Logistic Regression model achieved consistently high performance across all evaluation metrics. An accuracy of approximately 97.4% confirms that the classifier correctly labeled nearly all tumor samples in the test set. The precision of 0.976 indicates that most predictions of malignancy were correct, while the recall of 0.952 demonstrates that the model successfully detected almost all malignant cases. The F1-score of 0.964 provides a balanced measure of performance between precision and recall. The ROC AUC value of 0.996 suggests that the model distinguishes extremely well between malignant and benign classes across all probability thresholds.

The confusion matrix further validates these results, showing very few misclassifications. Most errors corresponded to false negatives—malignant tumors incorrectly predicted as benign—which, although rare, are the most critical type of mistake in a medical setting. The feature-pruning analysis further showed that comparable performance was maintained using as few as ten features, highlighting the model's robustness and efficiency.

(ii) Evaluation Discussion

Overall, the evaluation confirms that Logistic Regression is a suitable and highly effective model for this classification problem. Its strong ROC performance and balanced precision–recall trade-off indicate that the model generalizes well to unseen data. Beyond predictive accuracy, its interpretability provides meaningful insights into which biological features most strongly influence the outcome. The high coefficients for size- and shape-related variables are consistent with medical literature, reinforcing the validity of the results.

From a practical perspective, the simplicity and transparency of Logistic Regression make it an attractive choice for diagnostic decision support. Future work could explore class-weighted versions of the model to further minimize false negatives, apply cross-validation to enhance generalizability, or compare performance with ensemble models such as Random Forests. In conclusion, the evaluation demonstrates that Logistic Regression offers both analytical rigor and clinical relevance, serving as a strong baseline and an interpretable diagnostic tool.