

Logistic Regression Classification Project

Breast Cancer Dataset

Project Steps

- Load Breast Cancer dataset
- Preprocess: train/test split + scaling
- Train Logistic Regression model
- Make predictions
- Evaluate performance
- Visualize results

Method

- Logistic Regression is a simple linear classification model.
- It estimates the probability of a sample belonging to a class.
- Well-suited for binary problems like cancer detection.

Feature Insights



Feature Insights

- This chart shows the most important features driving the Logistic Regression model.
- Features with **positive coefficients** (on the right) increase the likelihood of predicting malignant tumors.
- Features with **negative coefficients** (on the left) push the model towards predicting benign tumors.

Results

Accuracy: 98.25%

Confusion Matrix:

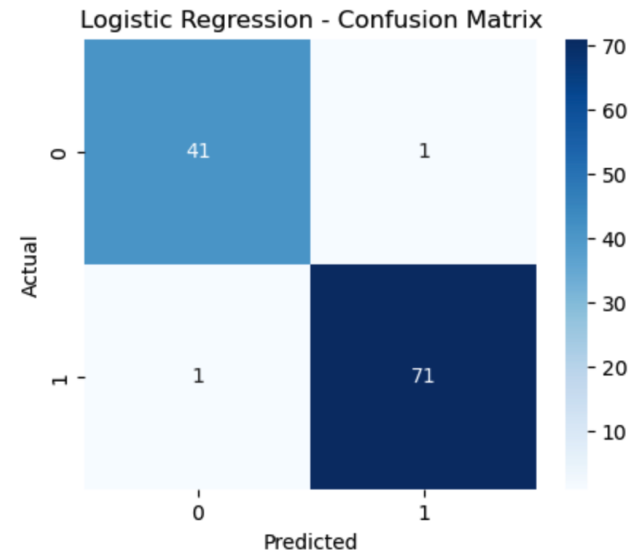
- $\begin{bmatrix} 41 & 1 \\ 1 & 71 \end{bmatrix}$

Classification Report:

- Precision ~ 0.98 – 0.99
- Recall ~ 0.98 – 0.99
- F1-score ~ 0.98 – 0.99

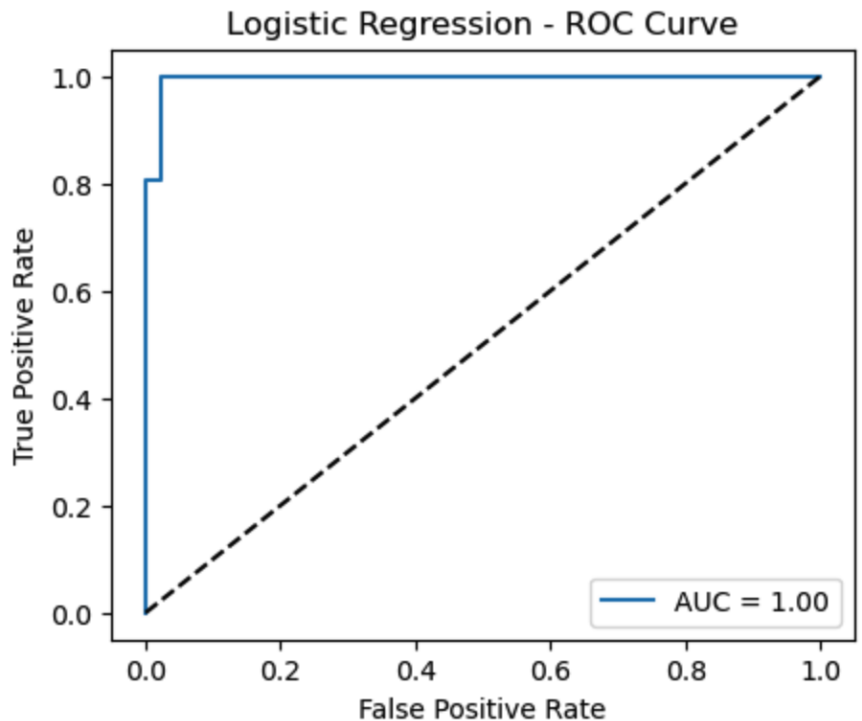
Confusion Matrix

The confusion matrix shows how well the model classified the test data. Out of 114 samples, it only misclassified 2 cases — one false positive and one false negative. This confirms the model is highly accurate and reliable.



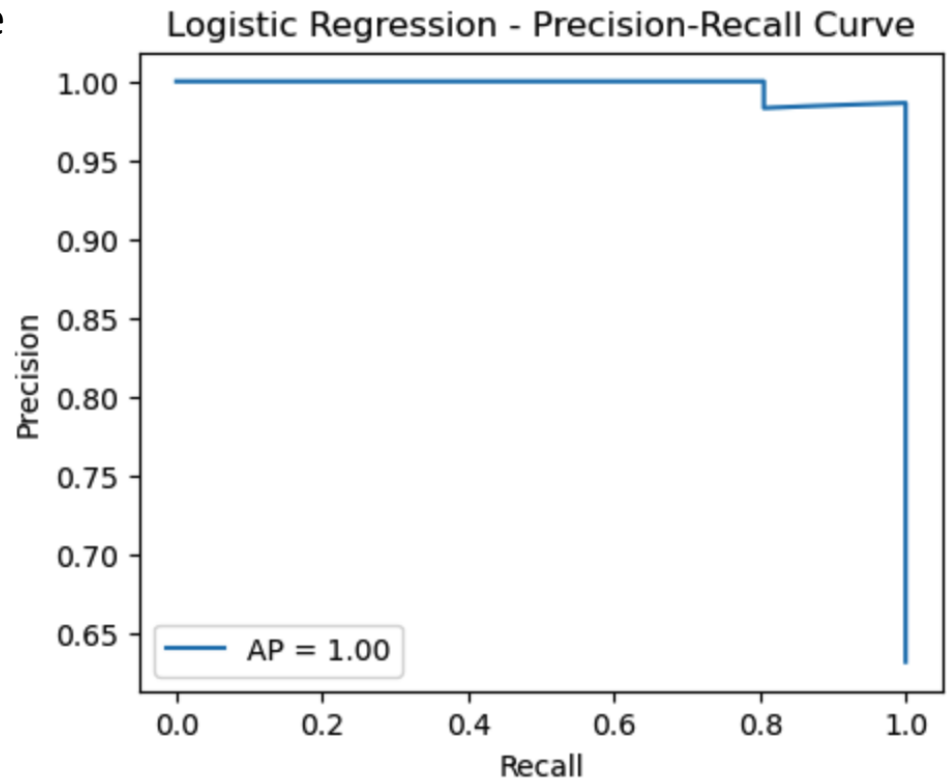
ROC Curve

The ROC curve measures the trade-off between sensitivity and specificity. The curve is close to the top-left corner, and the AUC is nearly 1.0. This means the model is excellent at distinguishing between malignant and benign cases.



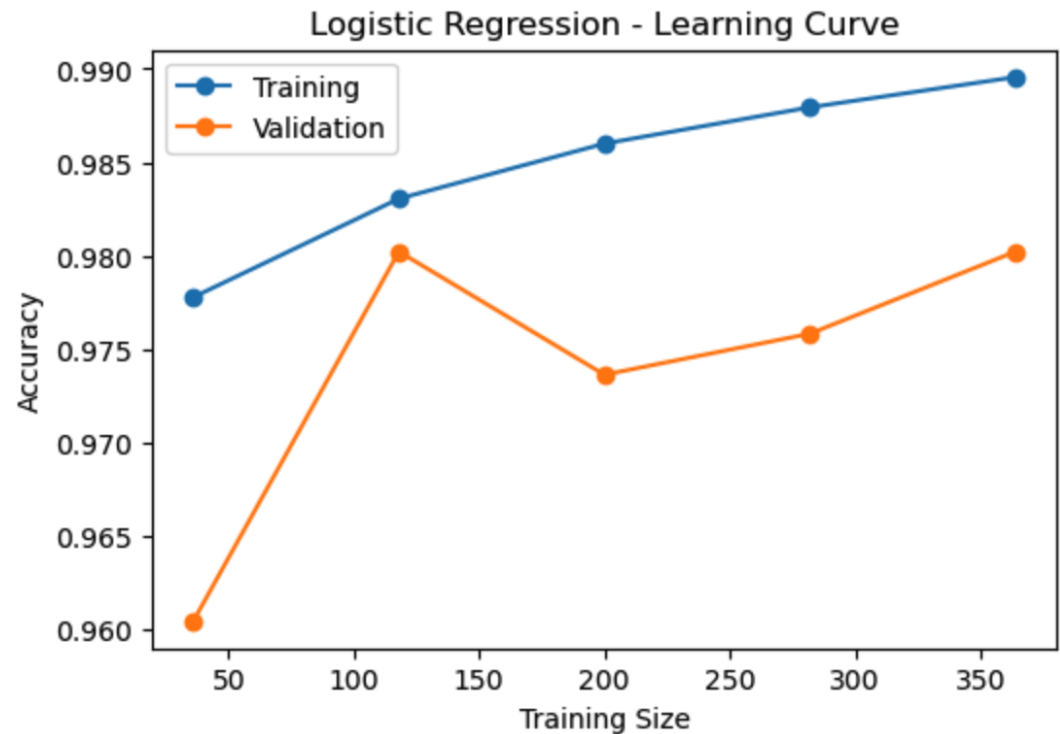
Precision Recall Curve

The Precision–Recall curve shows that both precision and recall remain high across thresholds. This tells us the model maintains strong performance even if we adjust the probability cut-off.



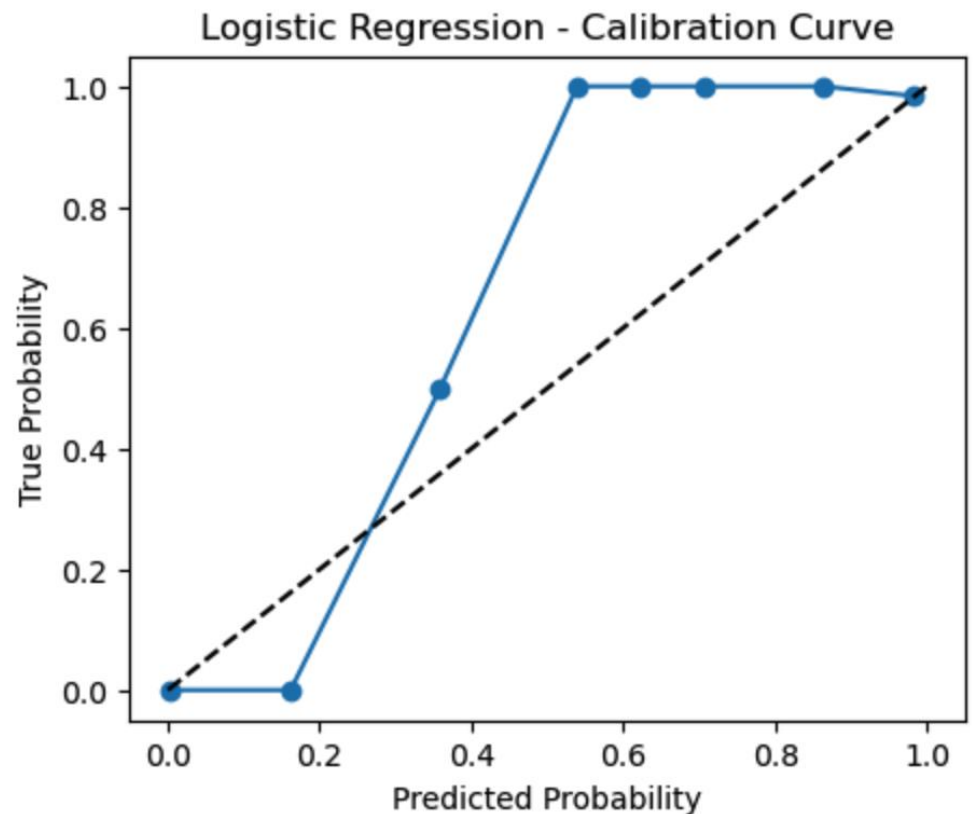
Learning Curve

The learning curve shows training and validation accuracy converging near 98%. This means the model is not overfitting and can generalize well to new, unseen data.



Calibration Curve

The calibration curve compares predicted probabilities with actual outcomes. Since the line is close to the diagonal, it means the model's probability estimates are well-calibrated and trustworthy.



Insights

- Very few misclassifications (only 2 errors).
- High accuracy, precision, and recall.
- ROC curve AUC close to 1 (excellent discrimination).
- Learning curve shows good generalization.

Conclusion

- Logistic Regression worked extremely well for this dataset.
- It achieved almost perfect accuracy with strong balance
- between precision and recall.
- The model is simple, interpretable, and effective
- for medical classification tasks like this one.