Random Forest Classification Project

Breast Cancer Dataset

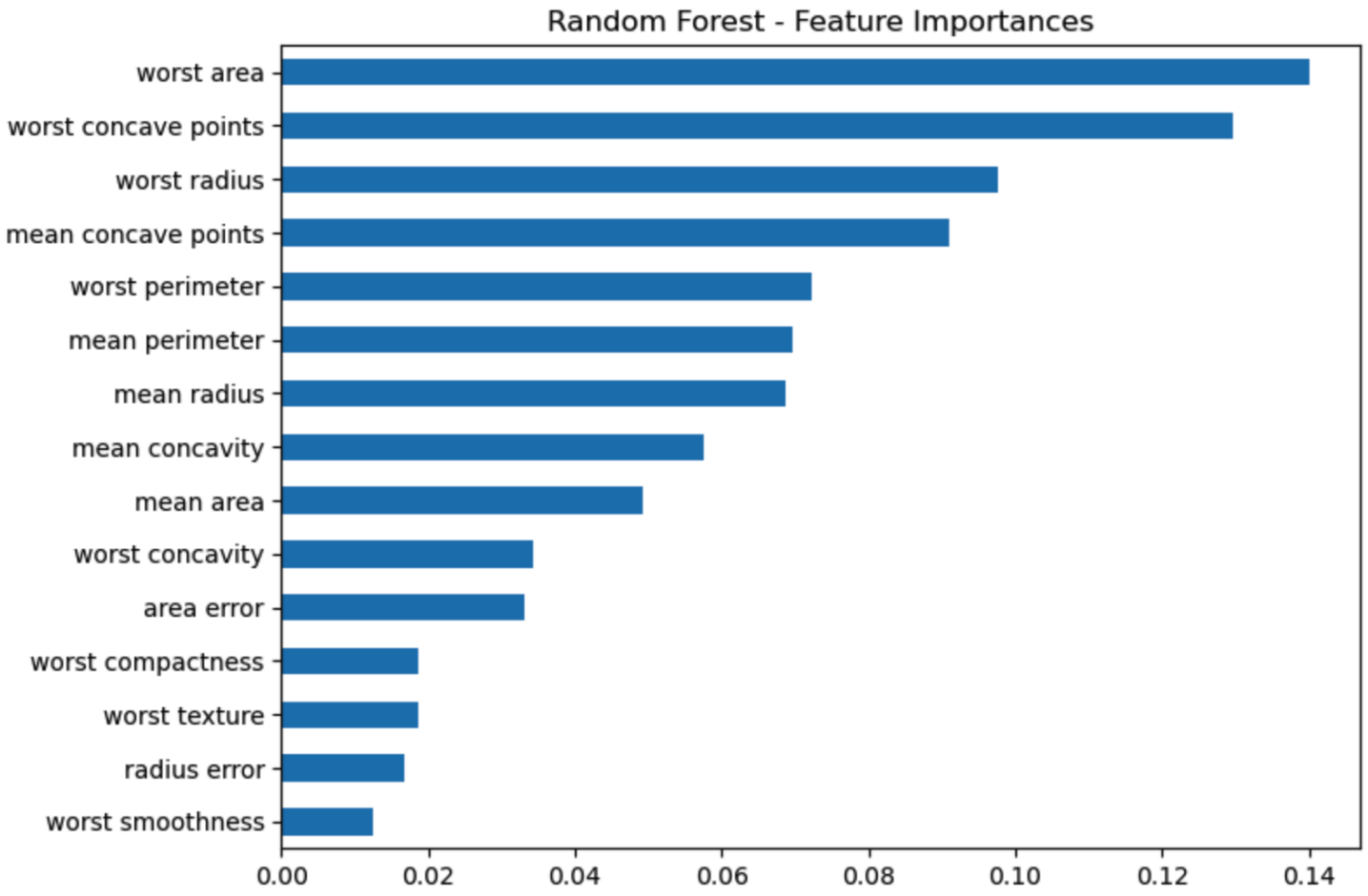
Project Steps

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

Method

- Random Forest is an ensemble learning method.
- It builds many decision trees and combines them.
- Good for capturing non-linear patterns and reducing overfitting.

Feature Insights



Feature Insights

- This chart shows which features were most important in the Random Forest model.
- Features like 'worst perimeter' and 'concave points' had the highest impact.
- Random Forest helps us understand which tumor characteristics drive predictions.

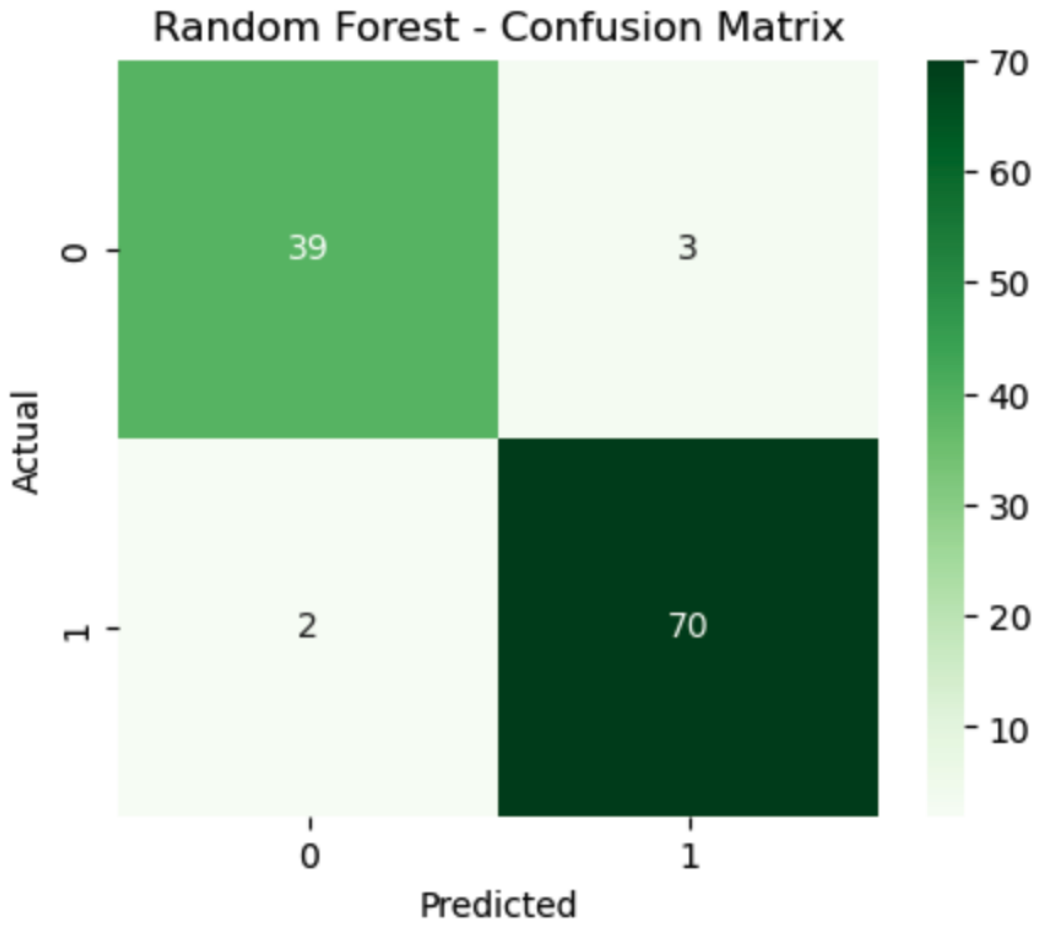
Results

Accuracy: 97.37%

Classification Report:

- Precision ~0.98 (class 0), 0.97 (class 1)
- Recall ~0.95 (class 0), 0.99 (class 1)
- F1-score ~0.96–0.98

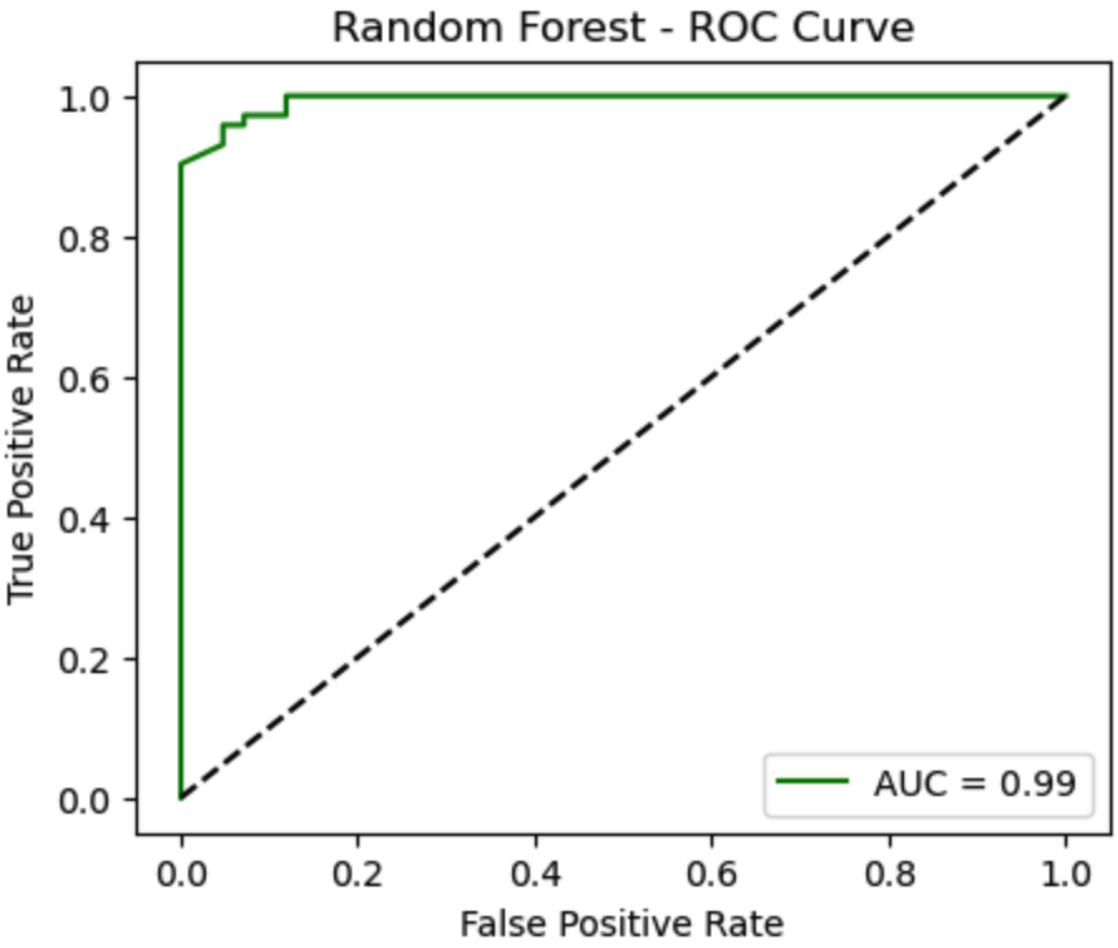
Confusion Matrix



Confusion Matrix

- The confusion matrix shows 5 errors out of 114 samples.
- Class 0 had 3 false positives (benign predicted as malignant),
  and Class 1 had 2 false negatives (malignant predicted as benign).
- Despite these few mistakes, the model is still very accurate and reliable.

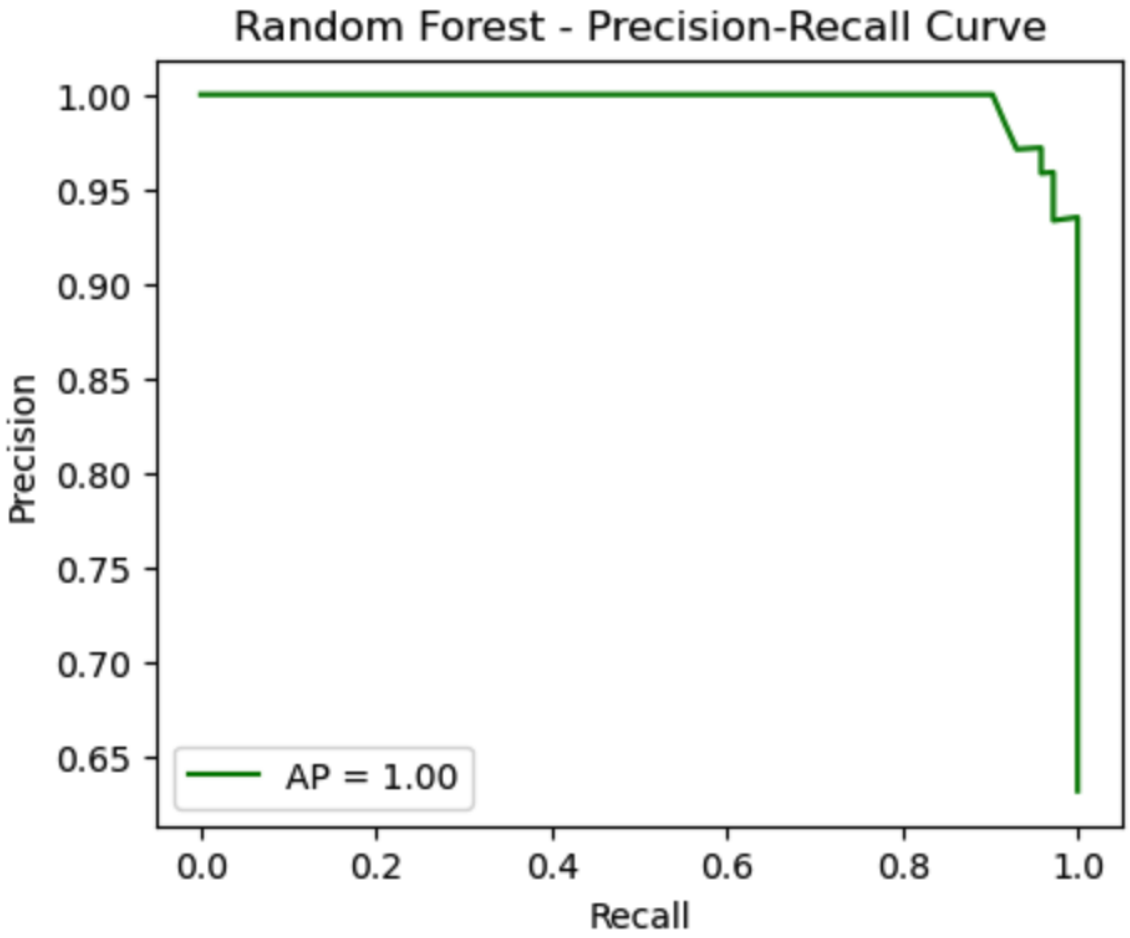
ROC Curve


Random Forest - ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.99

ROC Curve

- The ROC curve is close to the top-left corner.
- AUC is nearly 1, meaning the model is excellent at distinguishing
- between malignant and benign tumors.

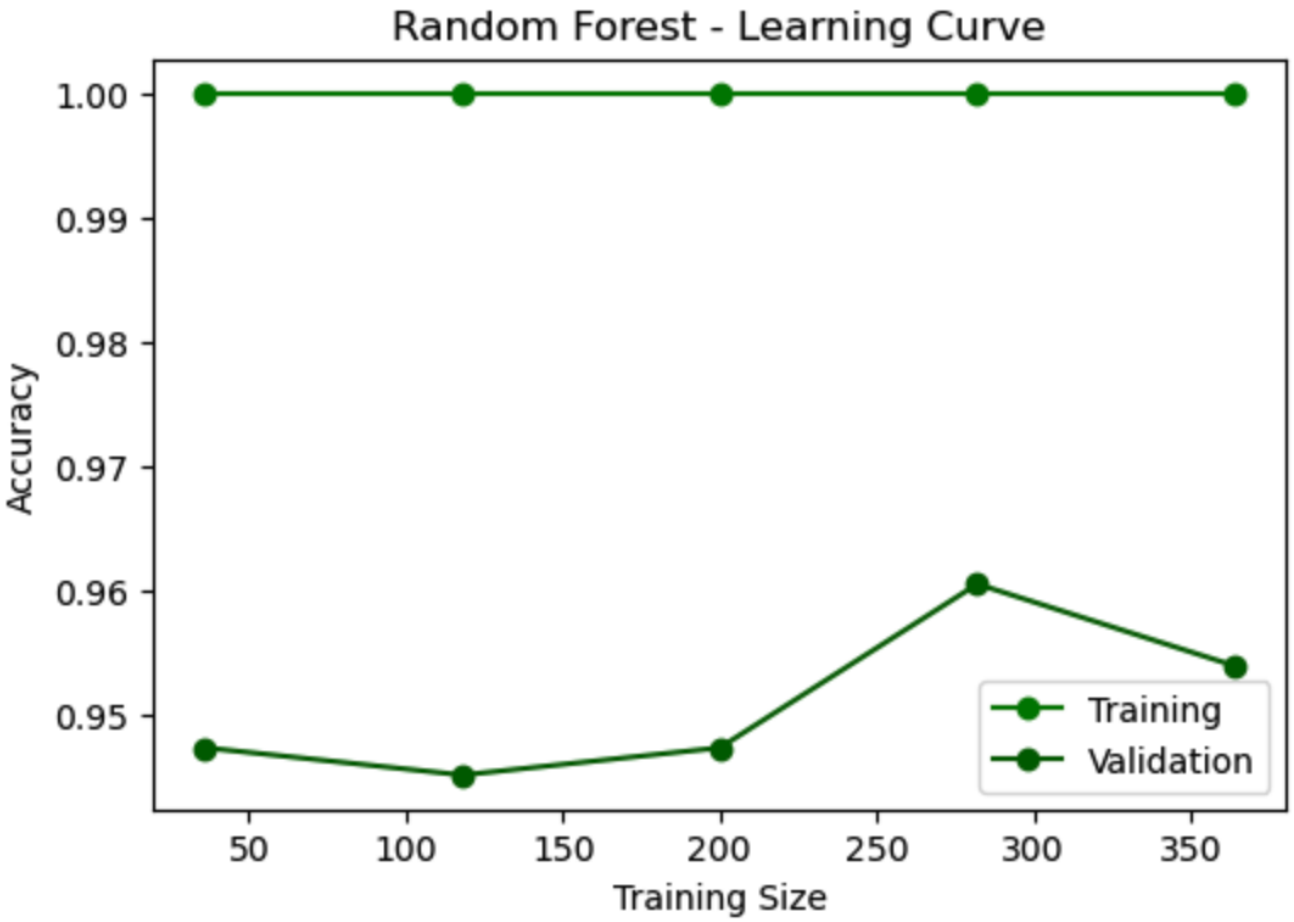
Precision-Recall Curve



Precision-Recall Curve

- Both precision and recall remain very high across thresholds.
- This means the model performs well even if we change the probability cut-off.

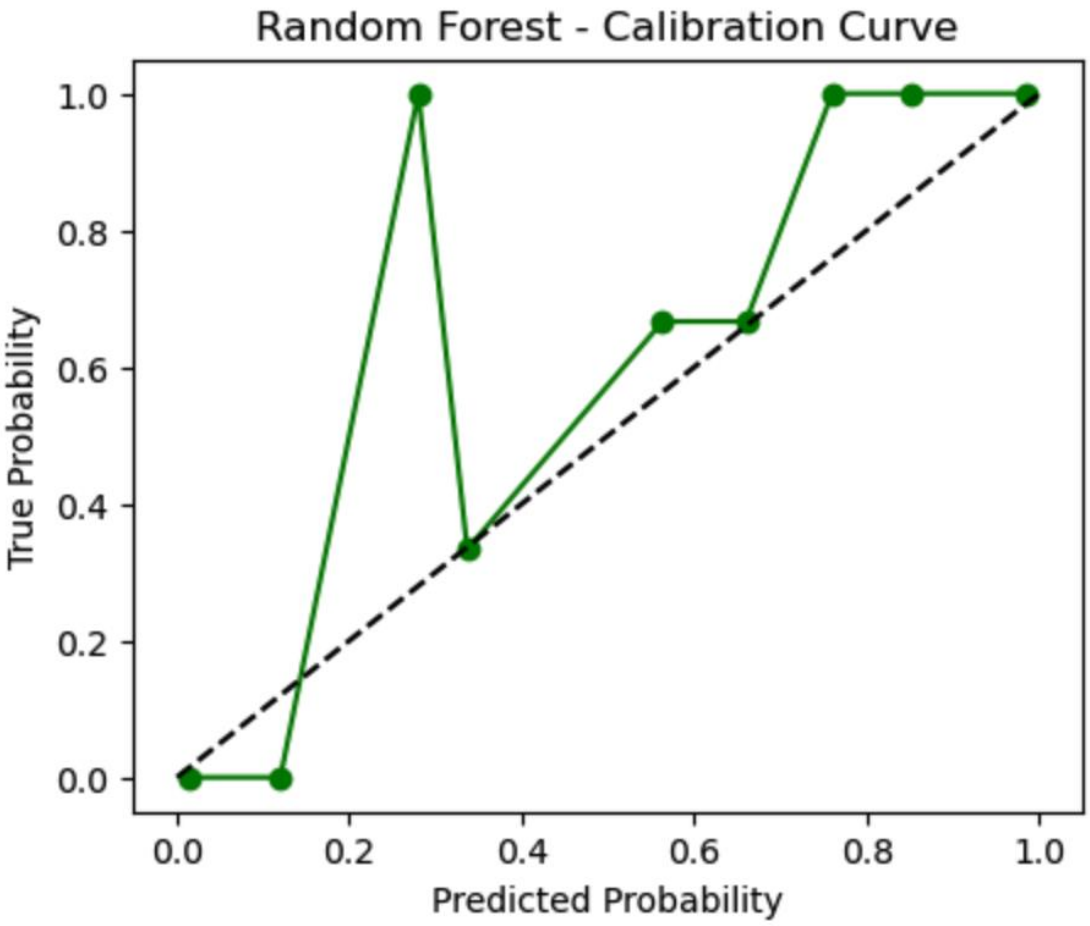
Learning Curve



Learning Curve

- The learning curve shows training accuracy near 1.0.
- Validation accuracy stabilizes around 97%.
- The small gap is normal for Random Forest and shows
- slight variance but strong generalization.

Calibration Curve



Calibration Curve

- The calibration curve shows predicted probabilities
- are close to actual outcomes.
- The model is slightly optimistic at higher probabilities
- but still well-calibrated overall.

Insights

- High accuracy and balanced performance across classes
- AUC close to 1.0, excellent class separation
- Random Forest captures complex, non-linear patterns
- Feature importance gives interpretability

Conclusion

- Random Forest performed extremely well on this dataset.
- It achieved strong accuracy with only a few errors.
- The model is powerful, robust, and interpretable.
- It is a great choice when accuracy and feature insights are important.