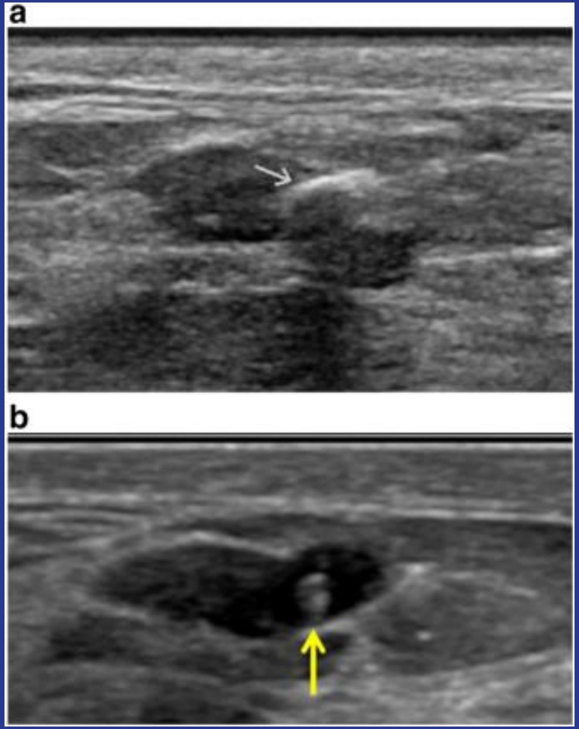
Breast Cancer Tumor Classification

Adam Millington, Nathan Tyler, Bab Jan, Ryan Mosher, Jake Lee



Introduction

Dataset & Description

- Breast Cancer Dataset by Ms. Nancy Al Aswad on Kaggle
- “Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image.”

Goal

Train a supervised ML model on the data to learn which images contained a malignant tumor and which contained a benign tumor.

Process

- Clean & Prep Data
- Upsample using SMOTE
- Test PCA
- Test Multiple ML Models
- Hyperparameter Tuning of Top Model (RandomForest)

Real-World Application:

NHS AI test spots tiny cancers missed by doctors

Snippets:

- “The tool, called Mia, was piloted alongside NHS clinicians and analysed the **mammograms of over 10,000 women.** Most of them were cancer-free, but it **successfully flagged all of those with symptoms, as well as an extra 11** the doctors did not identify.”

- **“Mia isn't perfect**. It had no access to any patient history so, for example, it would flag cysts which had already been identified by previous scans and designated harmless.”

Models Used & Purpose

Model 1

Linear Regression

(77.2% Accuracy)

Simple Model.

Setting Baseline Performance.

Model 2

Support Vector Classifier (SVC)

(91.6% Accuracy)

Finds the hyperplane that best separates the classes.

Robust against overfitting.

Model 3

KNeighborsClassifier

(92.8% Accuracy)

Useful when decision boundary is not necessarily linear.

Can capture complex patterns without underlying model assumptions.

Models Used & Purpose

Model 4

Logistic Regression

(97.2% Accuracy)

Go-To Model for Binary Classification.

Good balance between complexity & interpretability. Good Benchmark.

Model 5

Decision Tree Classifier

(98.4% Accuracy)

Can map complex relationships.

Easy to follow branches to pinpoint most valuable features.

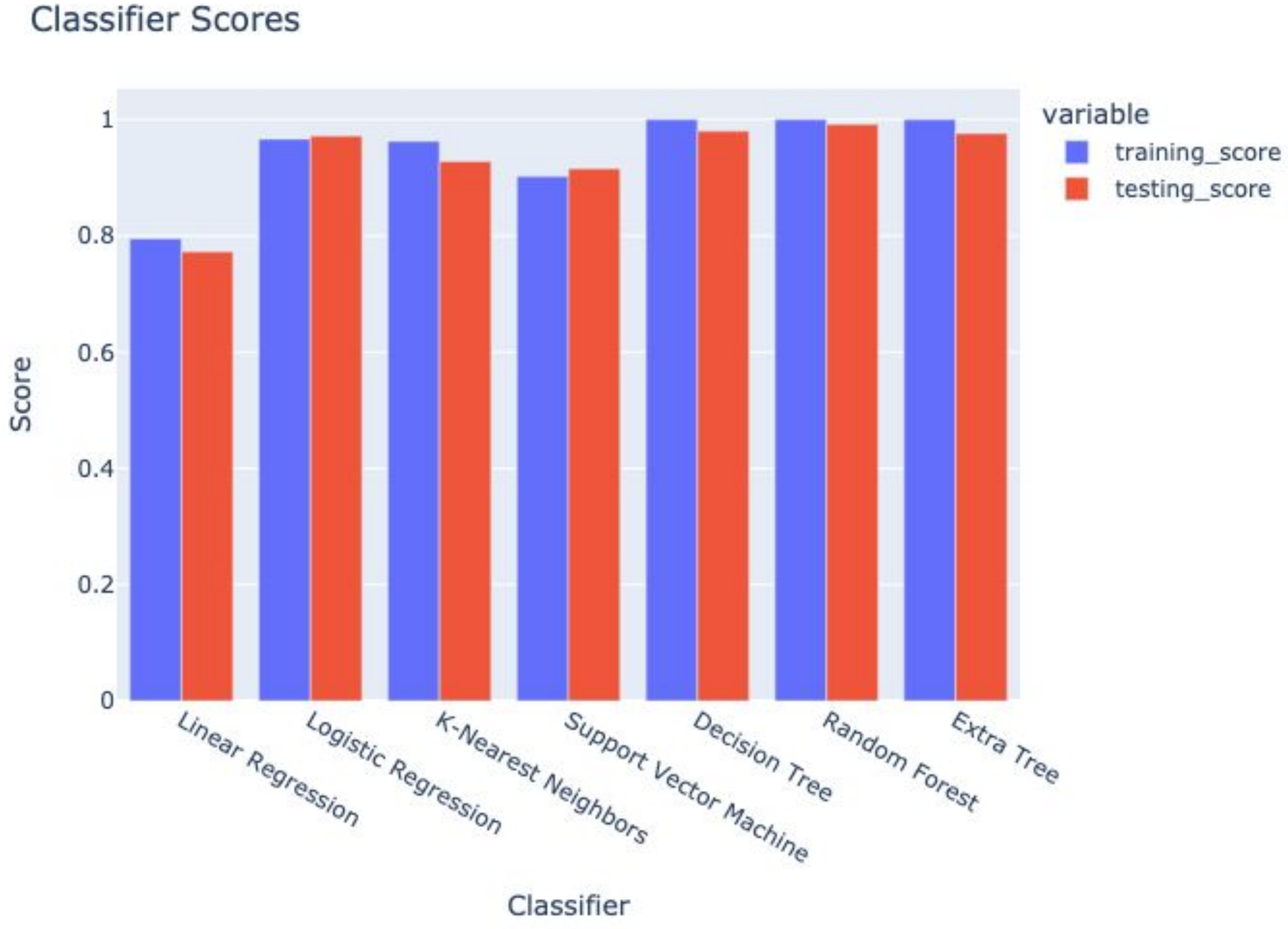
Model 6

Random Forest Classifier

(99.2% Accuracy)

Aggregates predictions of multiple decision trees.

Better measure of feature importance.


Classifier Scores
variable
🟦 training_score
🟥 testing_score
Score
1
0.8
0.6
0.4
0.2
0
Linear Regression
Logistic Regression
K-Nearest Neighbors
Support Vector Machine
Decision Tree
Random Forest
Extra Tree
Classifier

Perfecting Perfection

Top Model:
Random Forest

Random Forest Classifier had an initial model.score of 0.992.

After tuning multiple hyperparameters (n_estimators, max_depth, random state, etc) we were able to consistently increase the score to 0.996.

Perfecting Perfection

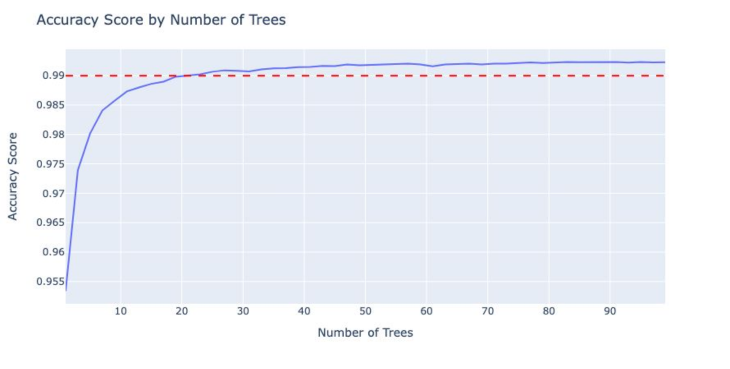
Avoid overfitting by limiting number of trees and max depth

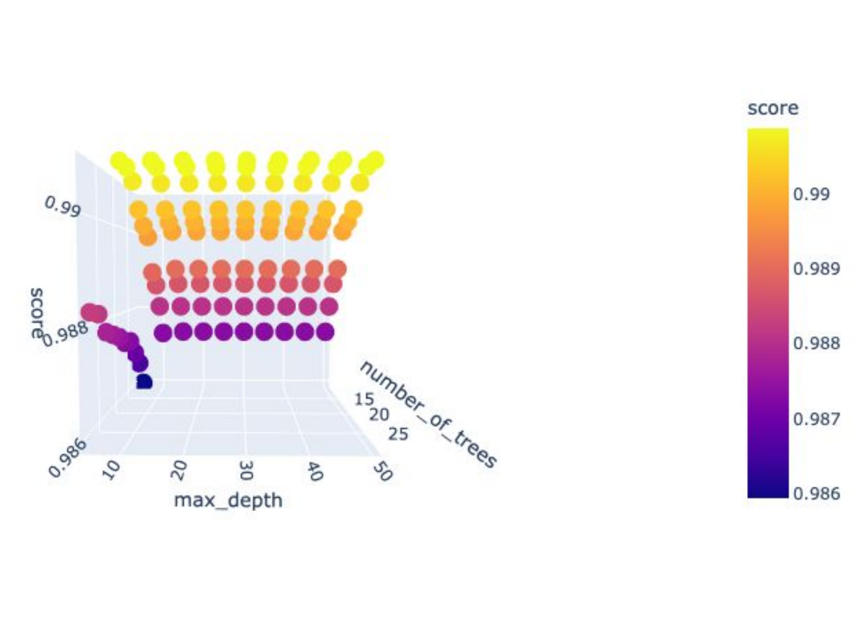
Looking for point of diminishing returns

Accuracy

Optimal # of Trees: 98.0%

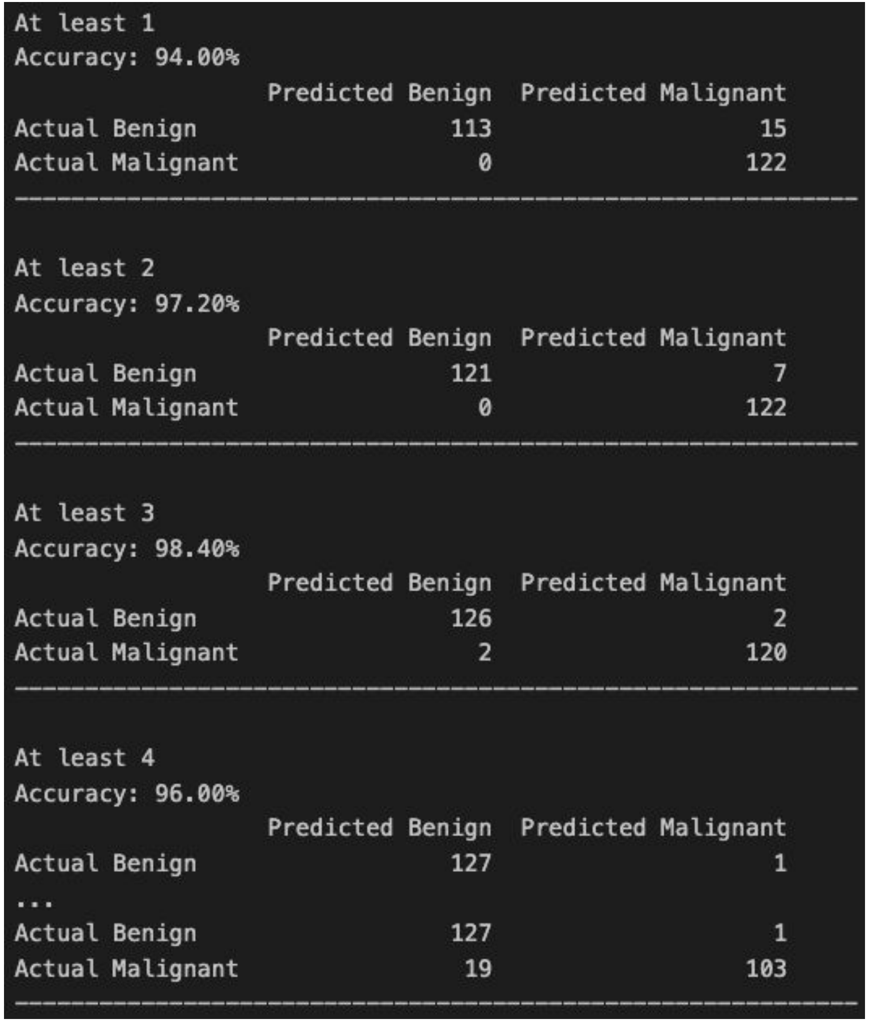
Optimal Max Depth: 98.8%





Biasing Against False Negatives

- In breast cancer diagnosis, there is generally a preference to minimize false negatives while also keeping false positives at an acceptable level.
- The goal is to achieve high sensitivity (few false negatives) to ensure that as many cases of breast cancer as possible are detected early and accurately. This helps in providing timely treatment and improving patient outcomes.
- At the same time, efforts are made to maintain specificity (few false positives) to avoid subjecting patients to unnecessary interventions and the associated physical and emotional burden.
- By Ensembling models we can drive the balance of FP to FN

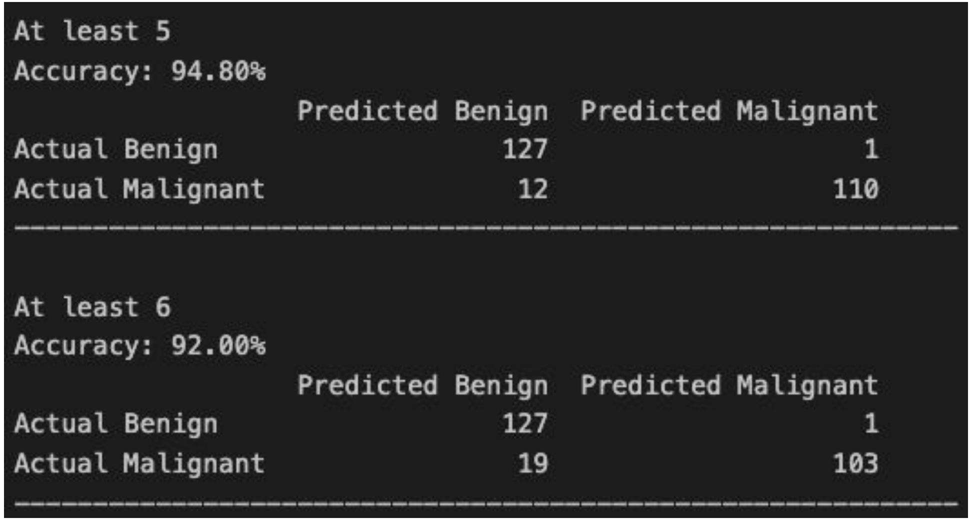


```
At least 1
Accuracy: 94.00%
                  Predicted Benign  Predicted Malignant
Actual Benign                  113                   15
Actual Malignant                 0                  122
-------------------------------------------------------------

At least 2
Accuracy: 97.20%
                  Predicted Benign  Predicted Malignant
Actual Benign                  121                    7
Actual Malignant                 0                  122
-------------------------------------------------------------

At least 3
Accuracy: 98.40%
                  Predicted Benign  Predicted Malignant
Actual Benign                  126                    2
Actual Malignant                 2                  120
-------------------------------------------------------------

At least 4
Accuracy: 96.00%
                  Predicted Benign  Predicted Malignant
Actual Benign                  127                    1
...
Actual Benign                  127                    1
Actual Malignant                19                  103
-------------------------------------------------------------
```

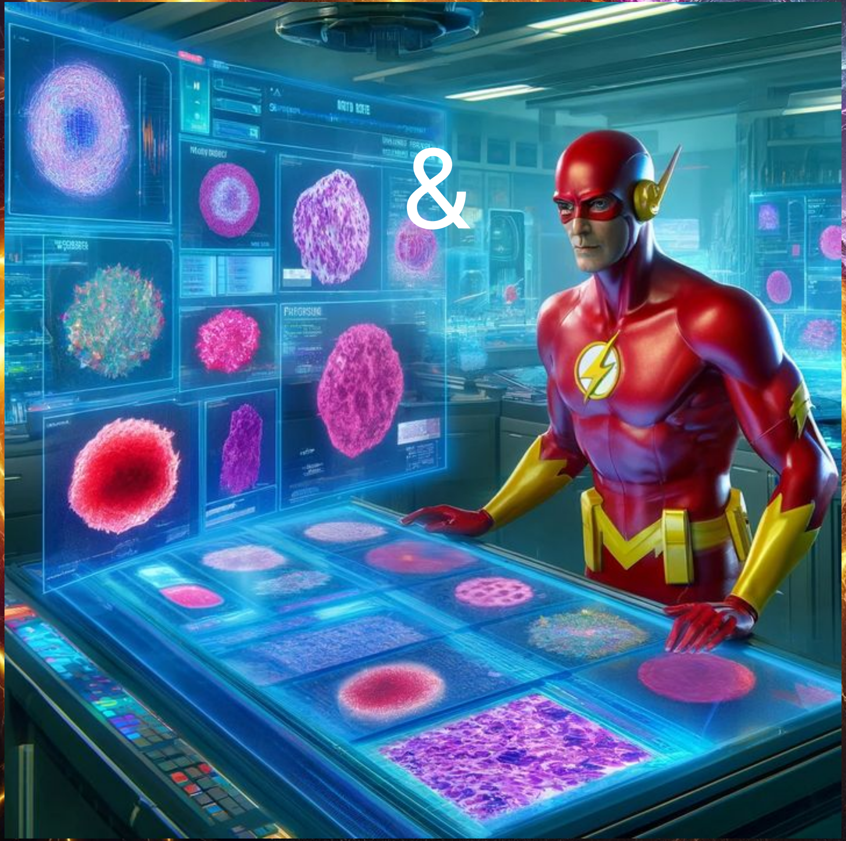


```
At least 5
Accuracy: 94.80%
                  Predicted Benign  Predicted Malignant
Actual Benign                  127                    1
Actual Malignant                12                  110
-------------------------------------------------------------

At least 6
Accuracy: 92.00%
                  Predicted Benign  Predicted Malignant
Actual Benign                  127                    1
Actual Malignant                19                  103
-------------------------------------------------------------
```

Model Interpretation

Feature Importance & Model Confidence



Feature Importance (Explained)

Gini Importance

Measures how much a particular feature contributes to the decision-making process of a model, specifically within tree-based models like Random Forest. It calculates this by looking at how often a feature is used to split data points in a tree and how much it improves the purity of the node (i.e., how well it separates the labels in each node).

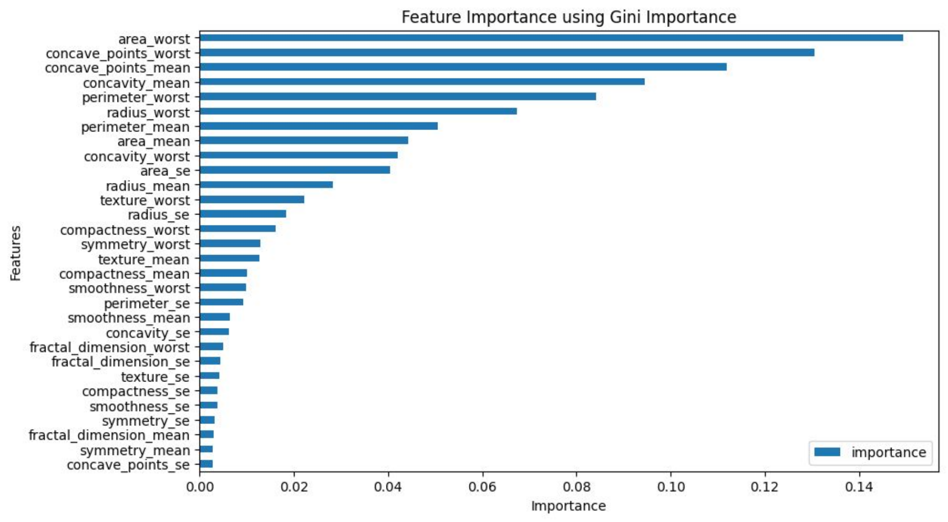
Permutation Importance:

Assesses the importance of a feature by measuring the decrease in the model's performance when the values of that feature are randomly shuffled. This shuffling breaks the relationship between the feature and the outcome, and if the model's performance drops significantly, it suggests the feature is important.

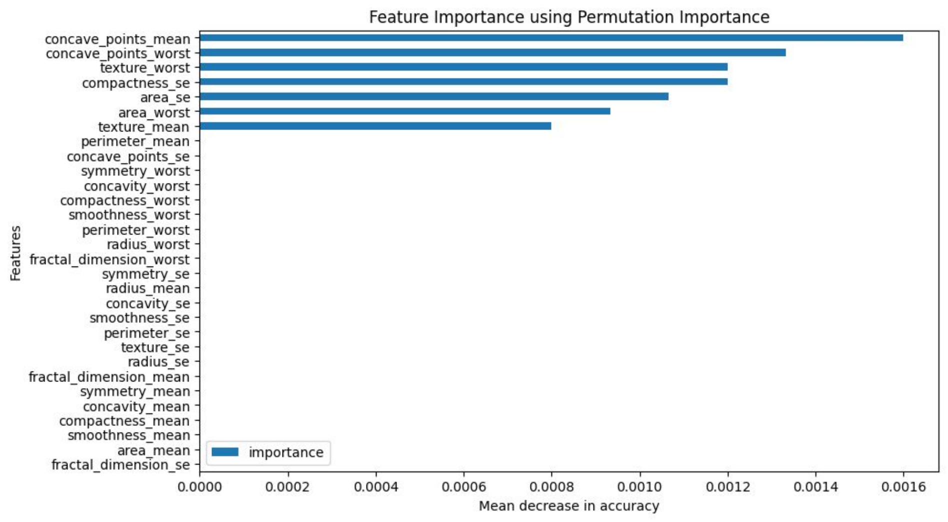
Comparison in Layman's Terms:

- Gini importance is like measuring a player's value based on their stats and direct contributions within the game's standard rules.
- Permutation importance is like evaluating a player's value by seeing how the team's performance suffers when the player is not playing their usual role.

Feature Importance (Gini vs Permutation)

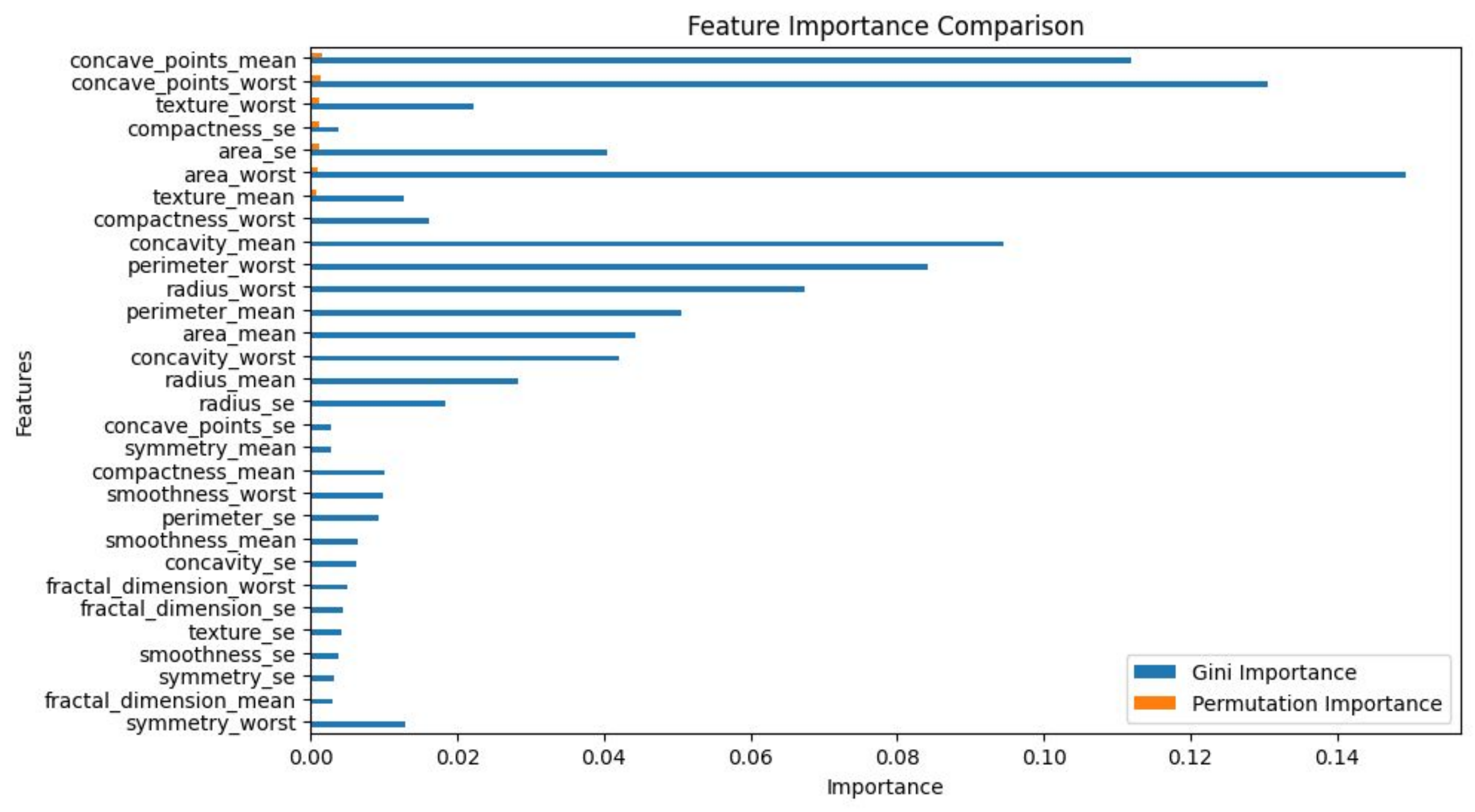


Feature Importance using Gini Importance



Feature Importance using Permutation Importance

Feature Importance (Combined)

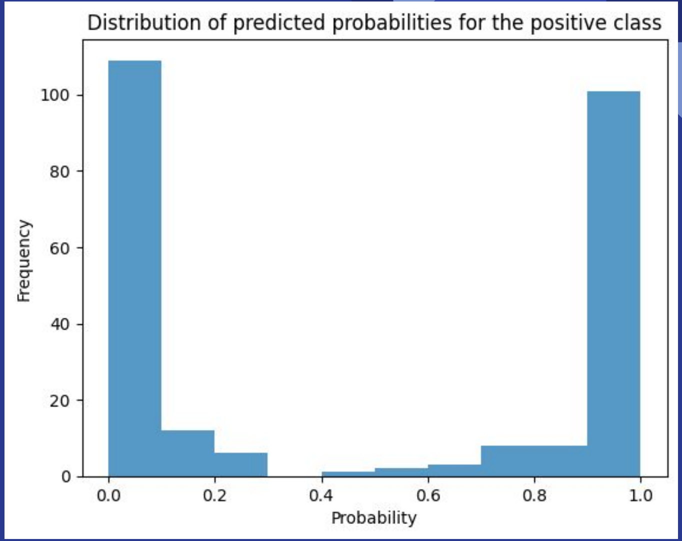


Model Confidence

Brier Score = 0.011385

A Brier score provides a measure of how well a probabilistic forecast matches the actual outcomes.

0.0 = Perfect Model
0.5 = Random
1.0 = Perfectly Inaccurate Model



Two peaks = high confidence in predictions.

Few predictions in the middle range (between 0.2 and 0.8) implies that the model is rarely uncertain about its predictions, which often indicates good calibration. This aligns with low Brier score.

Conclusion

Summary:

- Classified Breast Cancer Tumors using Random Forest Model & Achieved Testing-Set Accuracy Score of 99.6%

Significance:

- Extreme accuracy in early-detection of cancerous tumors can save millions of lives every year.

Model Strength & Limitations:

- Random Forest Model demonstrated excellent predictive capabilities, but may pose challenges for real-world integration. Results must be cautiously evaluated against diverse/larger datasets to ensure generalizability.

Future Work/ Areas for Further Research

Expand the Dataset

There were less number of frequency data available to train the model and more were required which may create future possible scope of the project in partnership or working opportunity with the industry research leaders like Stanford, Harvard and St. Jude etc.

Future Models

Future iterations could include neural networks and deep learning models for image recognition, and natural language processing techniques to combine reports and other information with the datasets.

Doc-CoPilot

(medical co-pilot)

Market Size = $19B USD

Human Error Rate > 25%

Achieved Testing-Set Accuracy Score of 99.6%, Contributing in Solving a **Global Problem** with a significant ROI.

Questions?