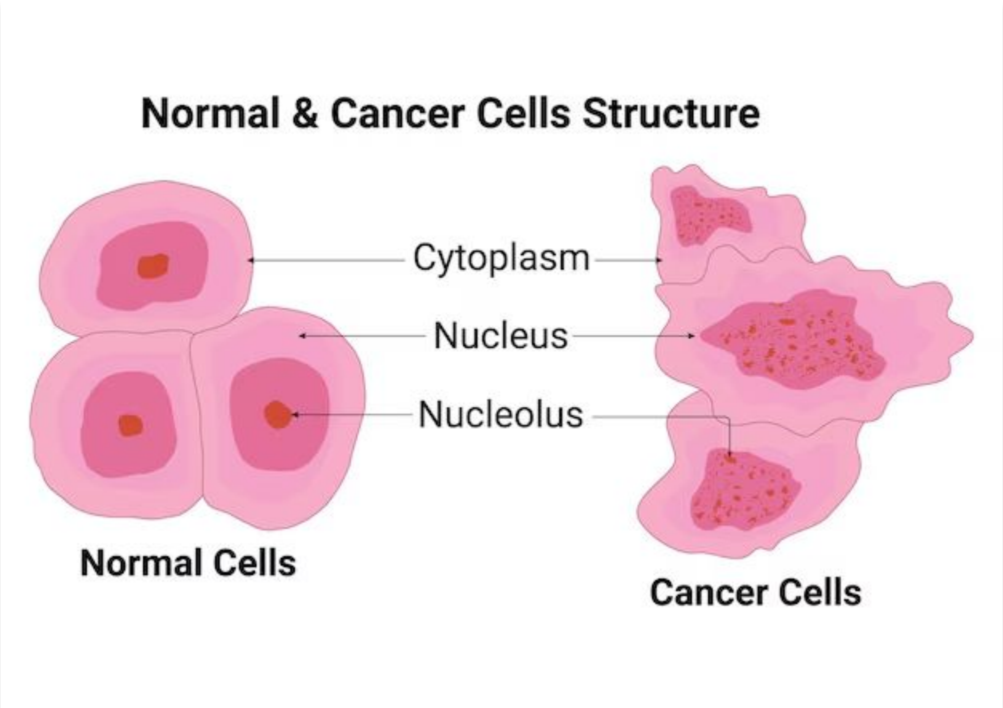
Tumor Classification Based on Nuclei Morphological Features

By Katherine Beaty, Niki Singh, May Al Khalifa, Riley Krisch

The problem



Currently, cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020.

Cancer mortality is reduced when cases are detected and treated early

Goal

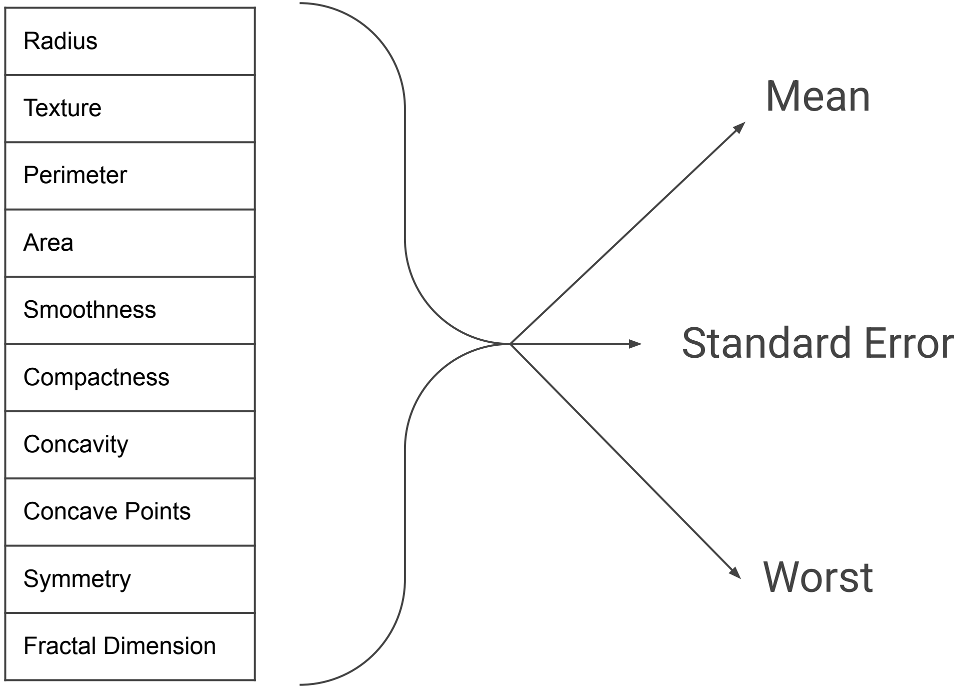
Identify a model that performs the best in recall as a use for early detection.

The Dataset

30
Features

569
Patients

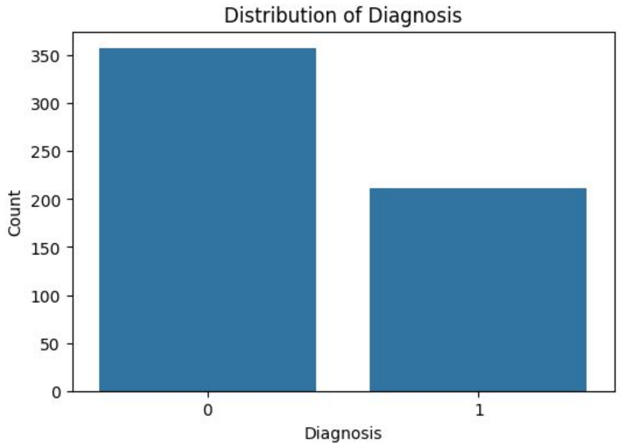
0
Missing Values



The Dataset

~37%

Malignant



~63%

Benign

Somewhat statistically unbalanced

Data Preprocessing

Cleaning

- Dropped irrelevant columns
- Removed all standard error columns
- Replaced spaces with underscores
- Converted diagnosis values from M/B to 1/0

Preparation

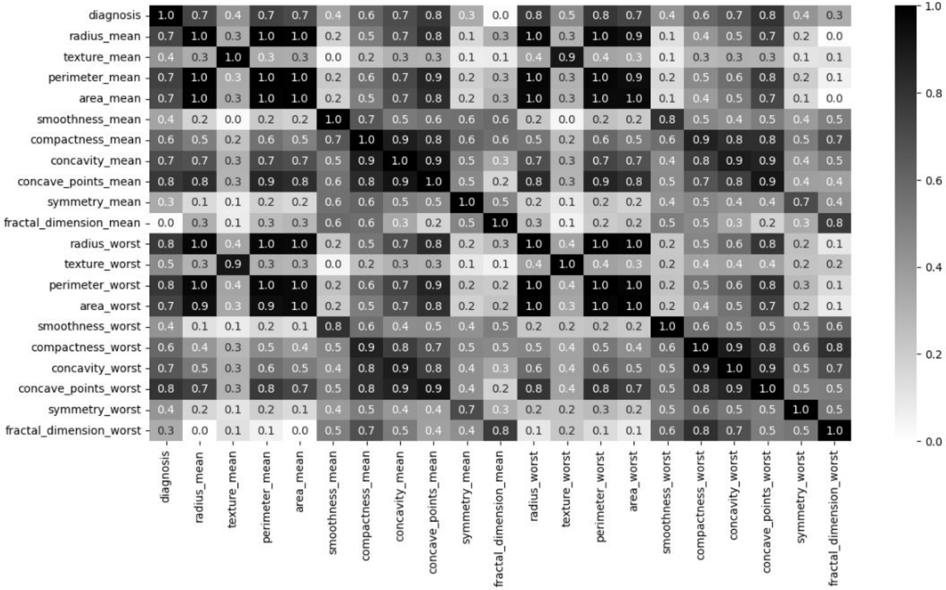
- Used 70/30 train-test split with fixed random state
- Standardized features

Two versions: With PCA and Without PCA

PCA

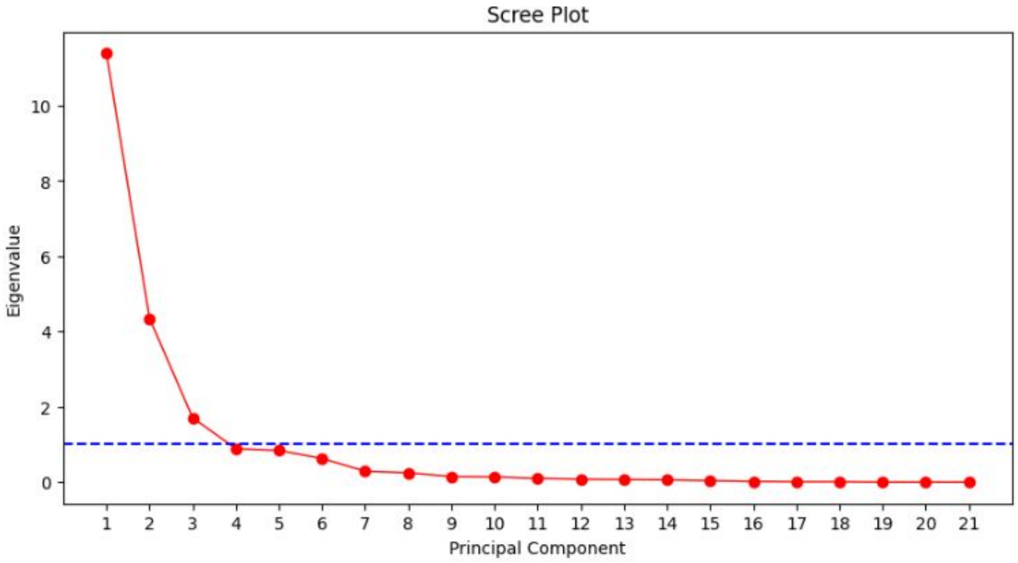
Pre-PCA

- Strongly Correlated Feature Pairs (Multicollinearity)
  - Increase risk of overfitting
  - Cause redundancy



Choosing Components

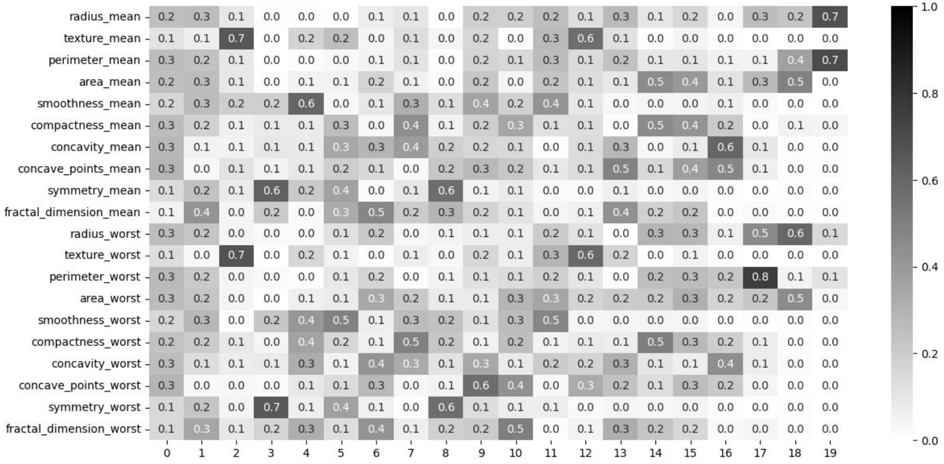
Decided on 4 PCA components to ensure that close to 90% of the original data information is retained



	% of variance explained	Cumulative % explained
0	0.532	0.532
1	0.219	0.751
2	0.086	0.837
3	0.045	0.882
4	0.040	0.922
5	0.030	0.952
6	0.013	0.966
7	0.008	0.973
8	0.007	0.980
9	0.005	0.985
10	0.004	0.989
11	0.004	0.993
12	0.003	0.995
13	0.002	0.997
14	0.001	0.999
15	0.001	0.999
16	0.001	1.000
17	0.000	1.000
18	0.000	1.000
19	0.000	1.000

Post-PCA

- Explains the majority of variance in the dataset
- Strong correlations with meaningful features (e.g. size, texture)
- Balance between interpretability and dimensionality reduction



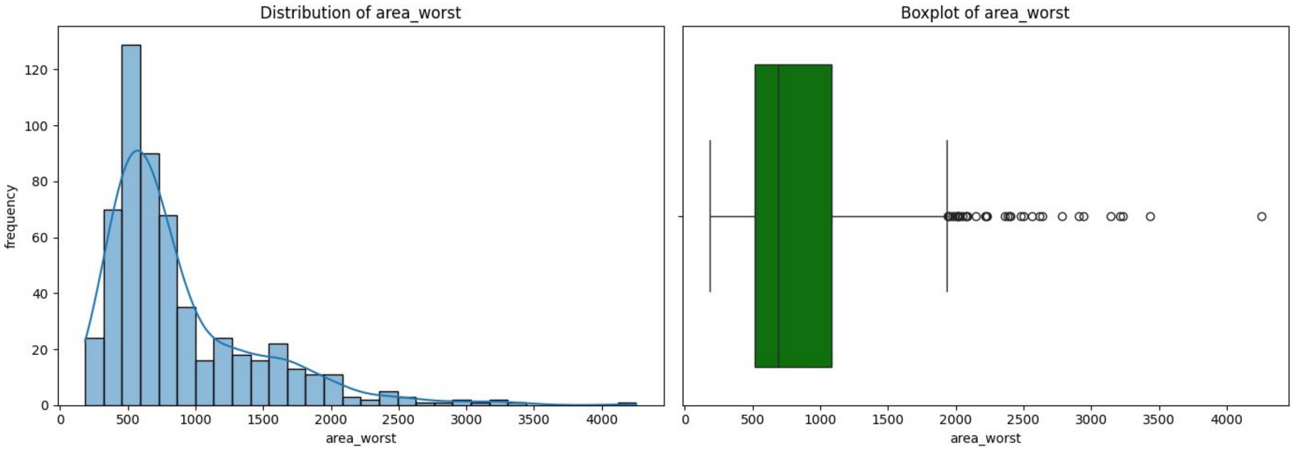
Baseline Model: Majority Rule

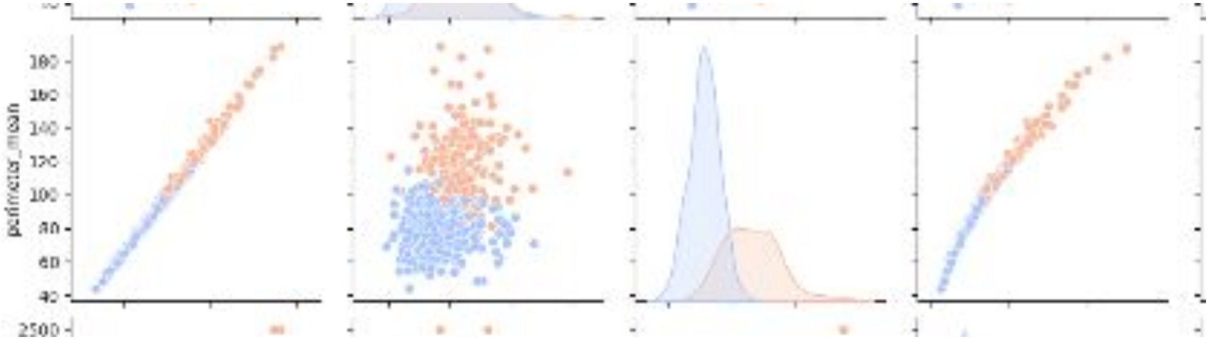
Predicting the majority class of 0 (Benign) diagnosis results in a baseline F1 score of 0%

0%

We aim to improve predictions from this baseline

Exploratory Data Visualization





Example snippet of pairplots

Modeling

Decision Tree (No PCA)

Full

Features	Nodes	Leaves	Maximum Depth
20	41	21	8

Precision	Recall	F1 Score
0.884058	0.968254	0.924242

More intuitive, but risk of overfitting

Pruned

Features	Nodes	Leaves	Maximum Depth
20	9	5	3

Precision	Recall	F1 Score
0.937500	0.952381	0.944882

Improved scores and better generalization on test data

Decision Tree (PCA)

Full

Features	Nodes	Leaves	Maximum Depth
4	43	22	7

Precision	Recall	F1 Score
0.861538	0.888889	0.875000

Better generalization, worse interpretability

Pruned

Features	Nodes	Leaves	Maximum Depth
4	7	4	3

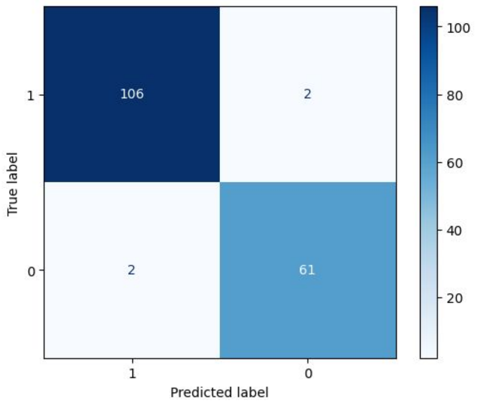
Precision	Recall	F1 Score
0.965517	0.888889	0.925620

Better generalization on test data, performed better on precision, increasing F1 score

Random Forest (n = 10,000)

No PCA

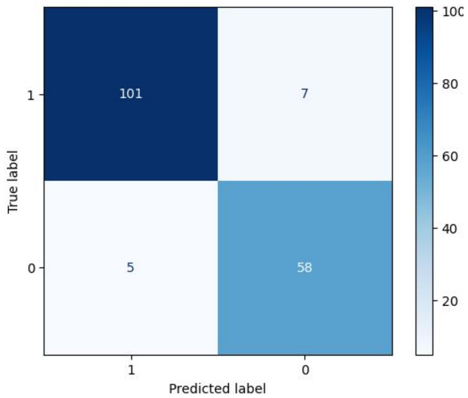
Precision	Recall	F1 Score
0.968254	0.968254	0.968254



Improved scores

PCA

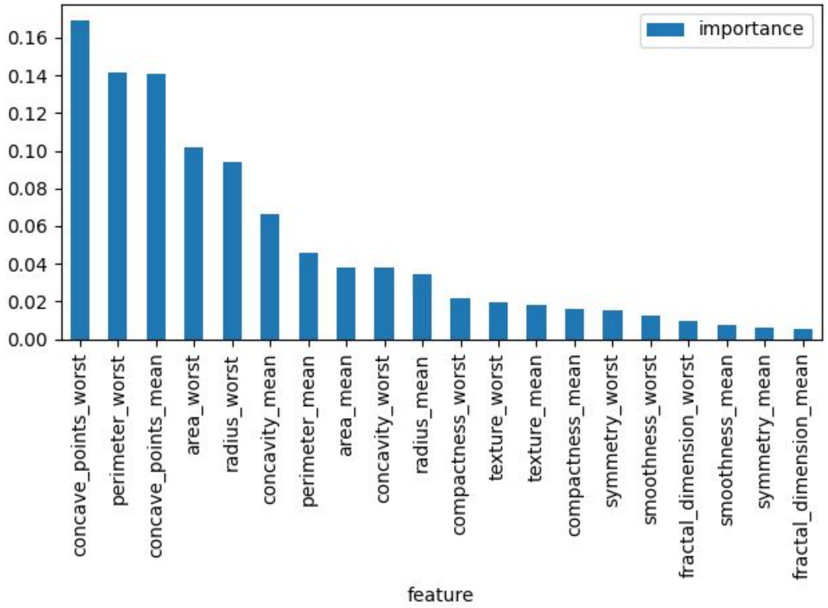
Precision	Recall	F1 Score
0.892308	0.920635	0.906250



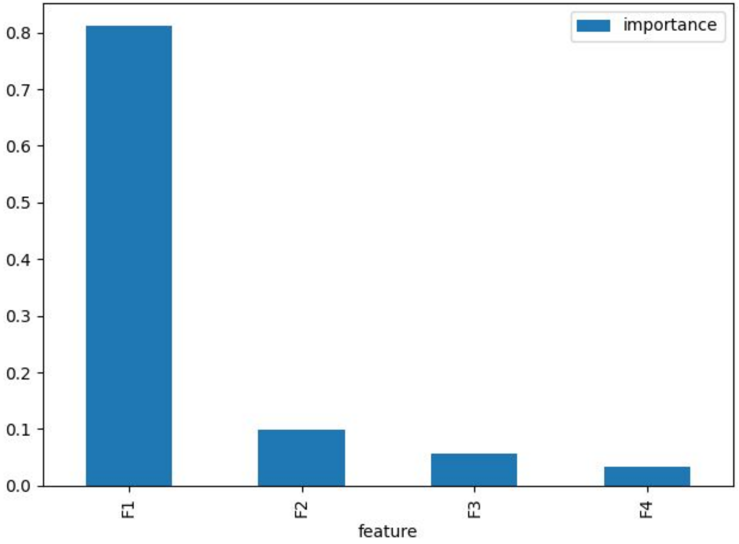
Performed worse than non-pca RF and pruned w/ PCA

Feature Importance

No PCA



PCA

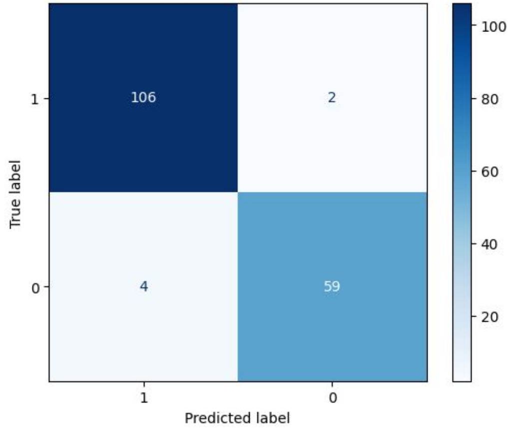


K-Nearest Neighbours

No PCA

best k value = 13

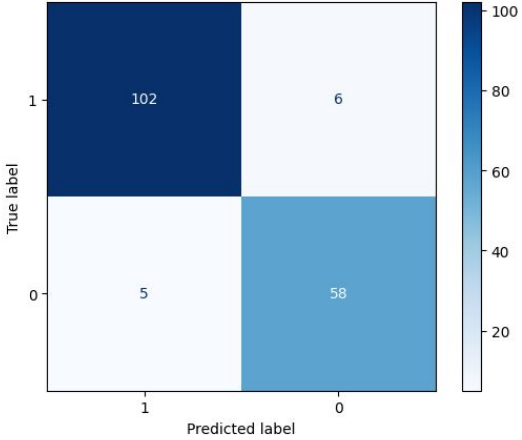
Precision	Recall	F1 Score
0.967213	0.936508	0.951613



PCA

best k value = 13

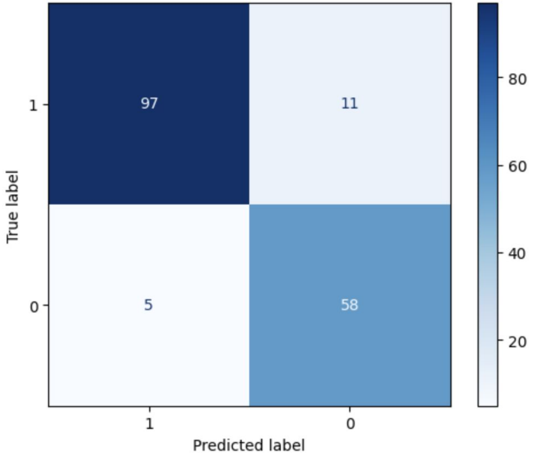
Precision	Recall	F1 Score
0.906250	0.920635	0.913386



Gaussian Naive Bayes

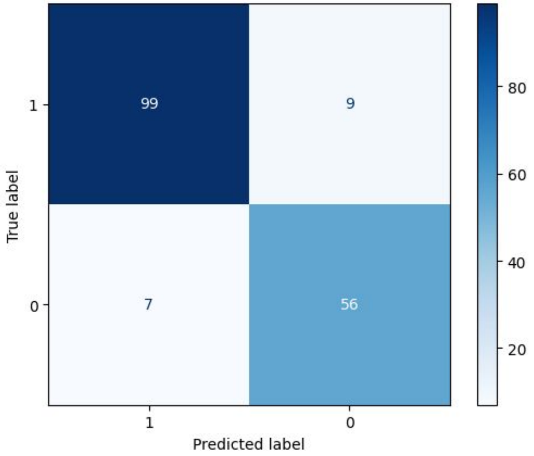
No PCA

Precision	Recall	F1 Score
0.840580	0.920635	0.878788



PCA

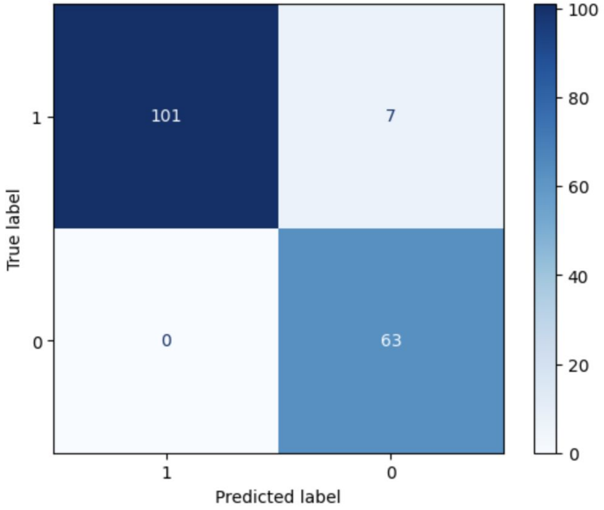
Precision	Recall	F1 Score
0.861538	0.888889	0.875000



Logistic Regression

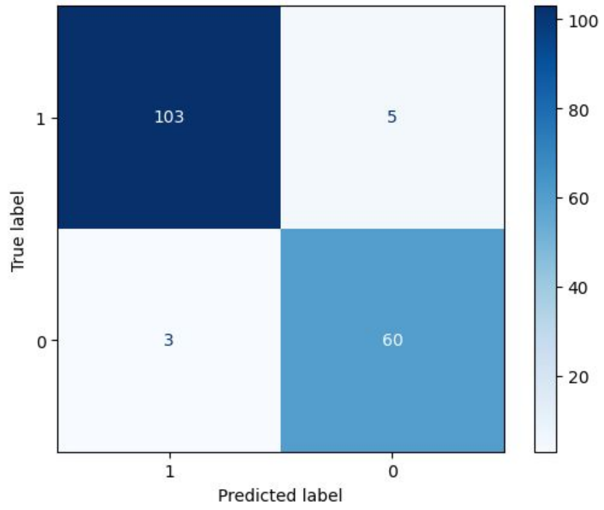
No PCA

Precision	Recall	F1 Score
0.950000	0.967593	0.956938



PCA

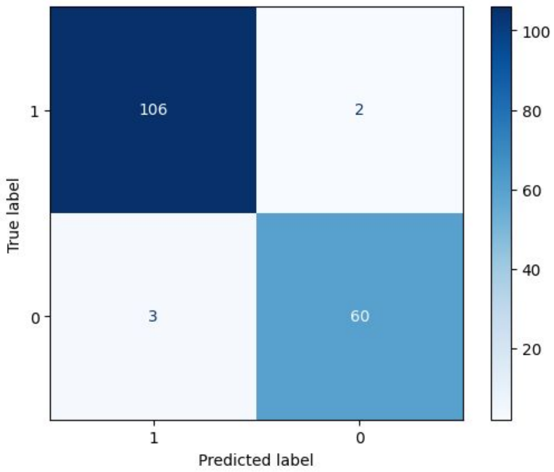
Precision	Recall	F1 Score
0.947388	0.953042	0.950058



Neural Network

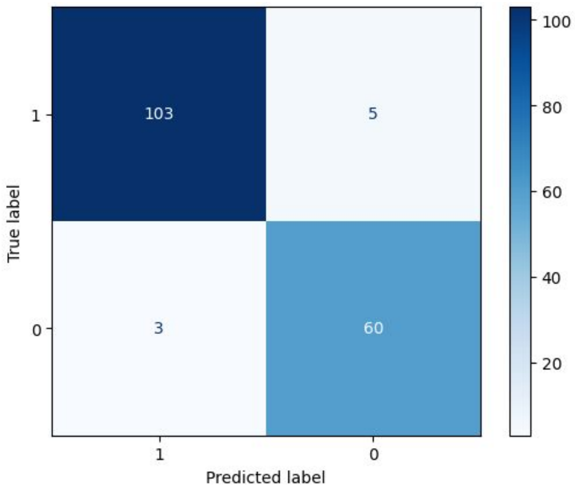
No PCA

Precision	Recall	F1 Score
0.967742	0.952381	0.960000



PCA

Precision	Recall	F1 Score
0.907692	0.936508	0.921875



Cost-Benefit Analysis

Chosen Cost-Benefit Matrix: [[0, 1], [5, 0]]

Model	Full decision tree	Pruned Decision Tree	Random Forest	Neural Network
Threshold	0.01	0.05	0.35	0.01
Min cost	18	19	7	12

Models with PCA	Full decision tree	Pruned Decision Tree	Random Forest	KNN	Naive Bayes	Logistic Regression	Neural Network
Threshold	0.01	0.43	0.23	0.24	0.12	0.36	0.4
Min cost	44	37	18	12	19	7	8

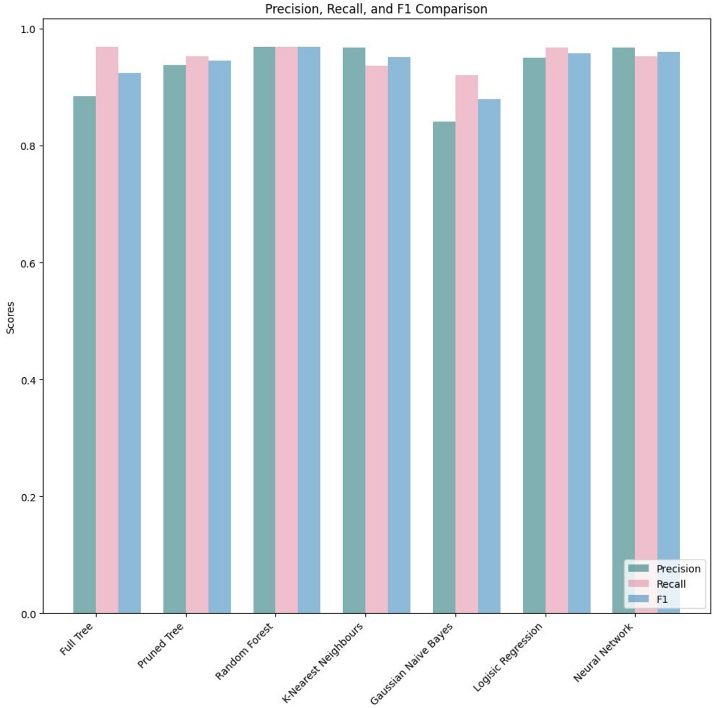
Best Model: Random forest and Logistic regression which both minimized costs to 7 at a more sensitive threshold of 0.35 and 0.36 respectively.

Challenges

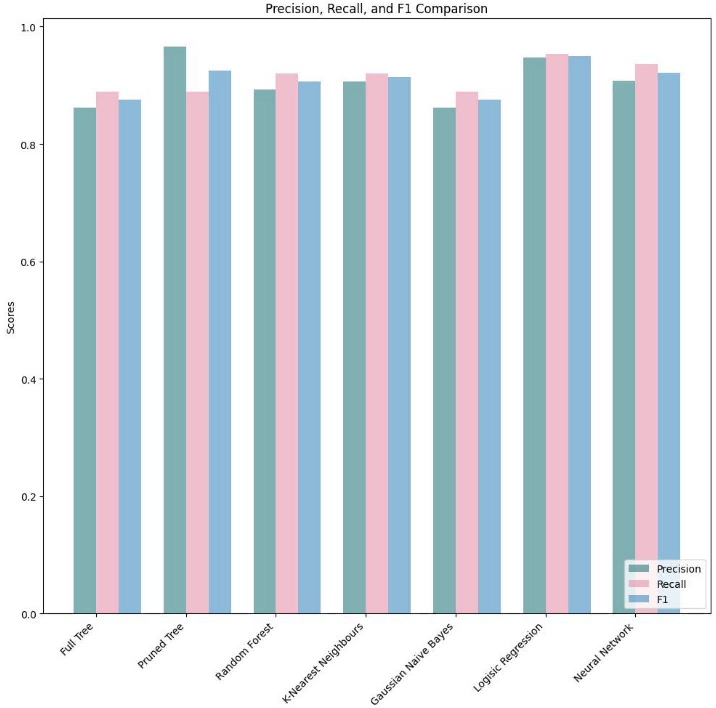
1. Choosing PCA vs. Non-PCA for final model
2. Transition from accuracy to F1 score
3. Training with PCA

Conclusions

No PCA



PCA



Solution

Due to this information being used in a clinical setting, we prescribed it best to use a *Non-PCA Random Forest* as the ideal model for classifying Benign or Malignant Tumors for Breast Cancer. This retains interpretability.