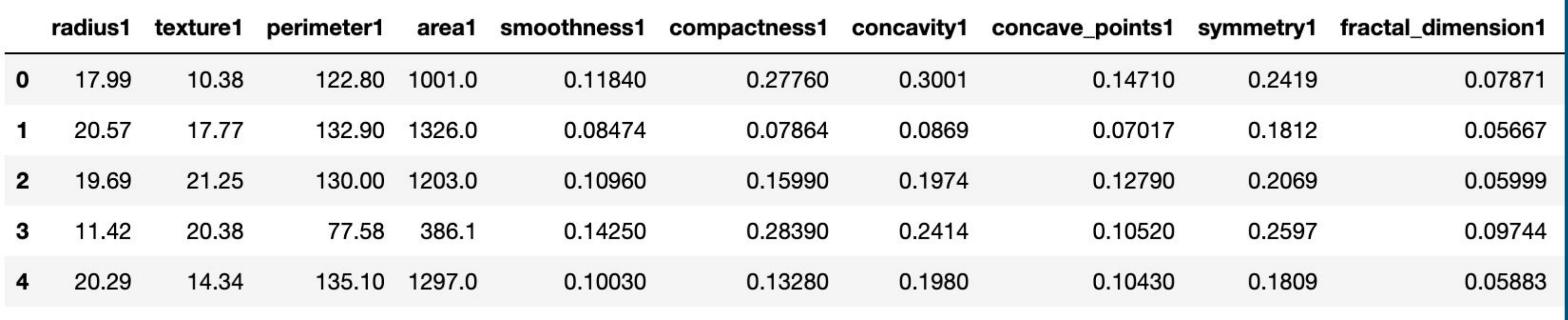
Wisconsin Breast Cancer Data

Josh Cochrane, Anna Mintz, Mackenzie Proper, Thanasis Pittas & John Douglas

Introduction to the Dataset

Comprises 569 samples, each with a diagnosis label (malignant or benign) and 30 real-valued features, detailing the characteristics of cell nuclei within images of breast masses

For each sample of cells, the 1 labels represent mean values, 2 labels represent the largest (more malignant) values, and 3 represents standard errors of a sample



	radius1	texture1	perimeter1	area1	smoothness1	compactness1	concavity1	concave_points1	symmetry1	fractal_dimension1
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883

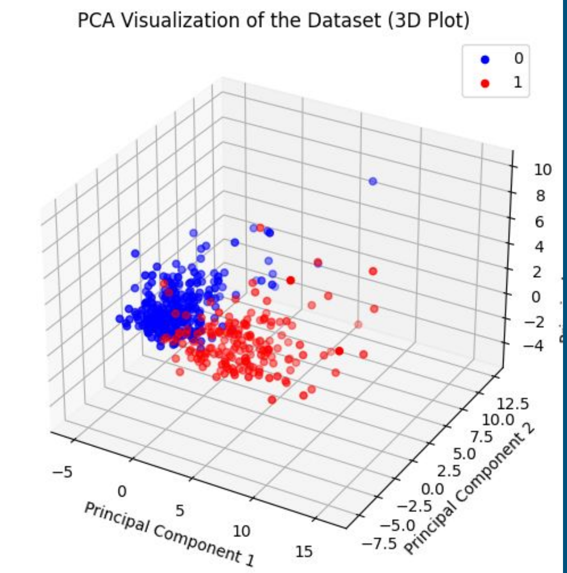
Methodology

We seek to gain deeper insight into the biological qualities of breast cancer

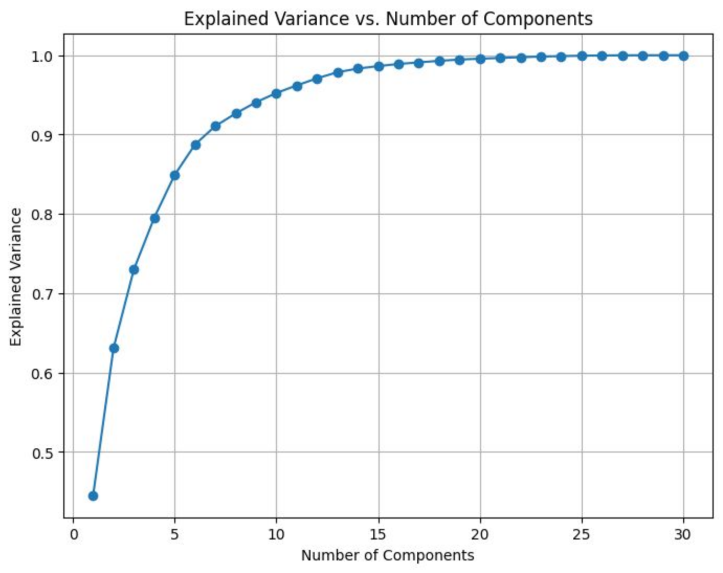
- Train models that accurately classify new samples as benign or malignant
- Used techniques such as feature selection, classification, and unsupervised learning algorithms
- Prioritize minimizing false negatives over false positives

PCA

Dataset looks reasonably separable



Few feature combinations explain most of the variance



⇒ We can hope for high accuracy & successful feature selection.

Feature Selection & Engineering

Permutation Important features:

- Mean Area
- SE Perimeter
- SE Area
- SE Texture

Robust with minimal influence from outliers.

Lasso:

- Largest Area
- SE Perimeter
- SE Area

Models based on this subset are extremely influenced by perimeter outliers, which could simply be an error in data collection.

Logistic Regression

$$\min_{w} \left\{ \frac{1}{2}\|w\|_1 + C\left(-\sum_i y_i \ln f_w(x_i) + (1-y_i)\ln(1-f_w(x_i)) \right) \right\}$$

1) Training without regularization (C=+∞):

Accuracy: 0.9825,
Precision: 1.0000,
Recall: 0.9524

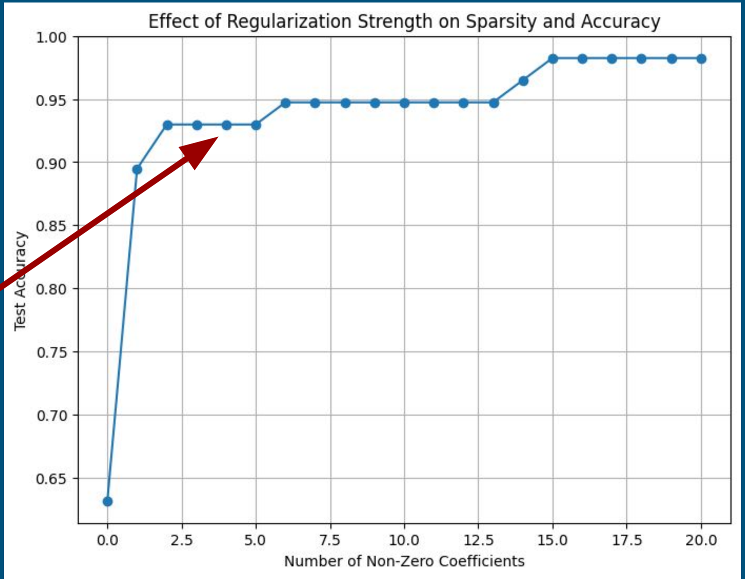
2) Training with L1 regularization (C ∈ [0.005,4]):

C=0.01

Non-zero coefficients:

'area1', 'perimeter3', 'area3', 'texture3'

Accuracy: 0.94
Precision: 0.86
Recall: 1.0

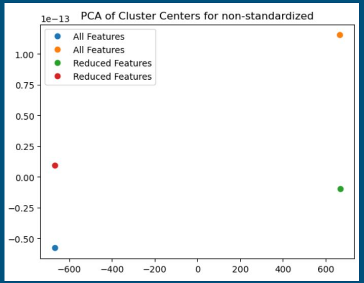


Support Vector Classifier Model

- Radial Basis Function (RBF) Support Vector Classifier
  - Ideal for complex datasets with nonlinear relationships between features and classes
- Using GridSearchCV to determine the best parameters
- Data subset: area1, perimeter3, area3, texture3
- Results:
  - Best Hyperparameters
    - C = 40
    - Gamma = 0.1
  - Accuracy = 93.86%
  - Precision = 96.23%
  - Recall = 91.07%

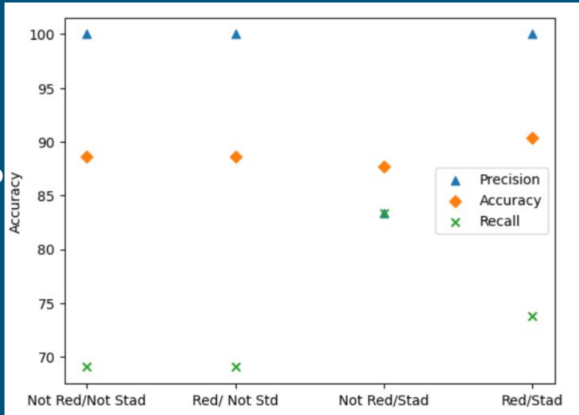
Random Forest Classifier Model

- An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes
  - Effectively avoids overfitting, can handle large datasets well
- Using GridSearchCV to determine the best parameters
- Data subset: area1, perimeter3, area3, texture3
- Results:
  - Best Parameters
    - Max Depth = None
    - Number of Estimators = 200
  - Accuracy = 95.61%
  - Precision = 96.63%
  - Recall = 94.64%



K-Means Clustering Model

- Unsupervised Learning model to classify the data based on clustering
  - Useful for determining natural groupings of data to combine like samples
- KMeans Clustering Hyperparameters
  - Know two clustered needed (Benign and Malignant) so n_cluster = 2
  - Used default n_iter = 10
- Scoring (Non-Standardized)
  - Accuracy = 88.60%; Precision = 100%; Recall = 69.05%
  - Same for reduced
- Scoring (Standardized)
  - Non- Reduced
    - Accuracy = 87.72%; Precision = 83.33%; Recall = 83.33%
  - Reduced
    - Accuracy = 90.35%; Precision = 100%; Recall = 73.81%



Conclusion

- The healthcare industry is extremely overworked
  - Because of this we selected a subset of important features to decrease data collection time and increase interpretability.
- We tested each of these models on all features - achieving a highest accuracy of 98% - and the selected subset of features.
  - This lead to a small, expected loss of 4% accuracy
- For each of the models, we tested accuracy, precision, and recall
  - We want to maximize the amount of true positives, even if that results in a decrease in accuracy
- Overall, we found that for stacking models, standardization doesn’t improve the model
- Logistic Regression, SVC, and Random Forest both were highly effective, even with reduced features for this dataset
- Our final model is Random Forest with hyperparameters, num_estimators = 200, resulting in a model requiring only 4 variables to achieve a diagnostic accuracy of 95.61%, and precision 96.23%.

Sources

William H. Wolberg, W. Nick Street, Olvi L. Mangasarian: Breast Cancer Wisconsin (Prognostic). UCI Machine Learning Repository, 1995. https://doi.org/10.24432/C5DW2B.