Breast Cancer Classification

Myisha Chaudhry and
Kayleigh Habib

Agenda

1	Purpose
2	Pre-Processing EDA
3	Methodology
4	Results
5	Conclusion

1. Overview

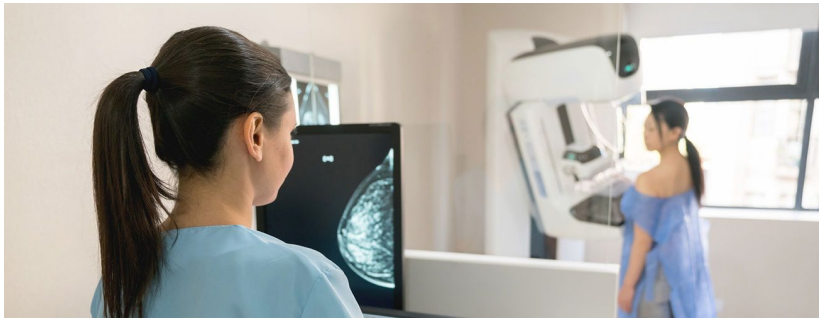
Introduction

Breast Cancer

- 1/8 women will develop breast cancer
- Successfully treated with the help of early detection

Mammograms

- X-ray machines produce images of breast tissue by flattening the breast
- Results assessed by a doctor to determine if further tests are required



Data Description and Cleaning

Original Data:

961 Records and 6 columns

- **BI-RADS assessment**: ordinal, non-predictive
- **Age**: patient's age
- **Shape**: round = 1, oval = 2, lobular = 3, irregular = 4
- **Margin**: circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5
- **Density**: high = 1, iso = 2, low = 3, fat-containing = 4
- **Severity**: benign = 0 or malignant = 1

	BI-RADS assessment	Age	Shape	Margin	Density	Severity
0	5	67	3	5	3	1
1	4	43	1	1	NaN	1
2	5	58	4	5	3	1
3	4	28	1	1	3	0
4	5	74	1	5	NaN	1

Cleaned Data:

836 Records and 5 columns

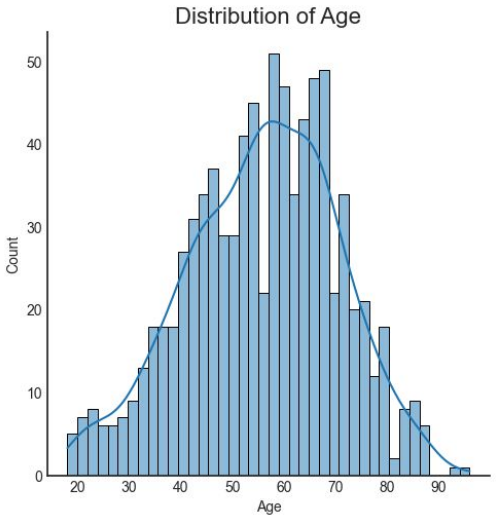
- Missing data
  - Changed to use median for Age
  - Remove remaining missing values
- Remove non-predictor features
  - BI-RADS should not be used as a predictor

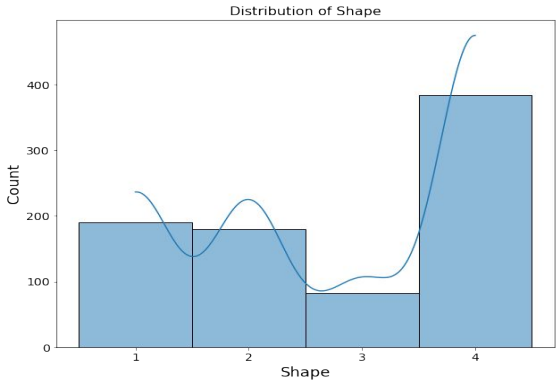
	Age	Shape	Margin	Density	Severity
0	67	3	5	3	1
2	58	4	5	3	1
3	28	1	1	3	0
8	57	1	5	3	1
10	76	1	4	3	1

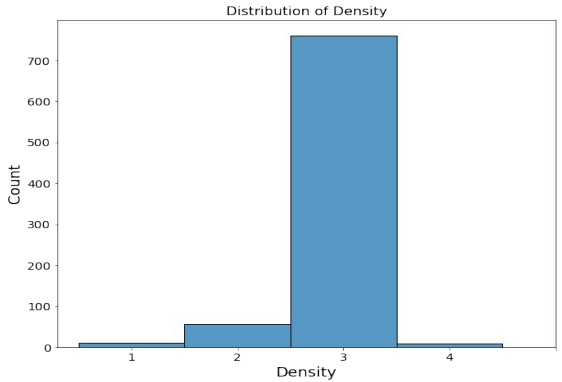
2. Pre-Processing and Exploratory Data Analysis

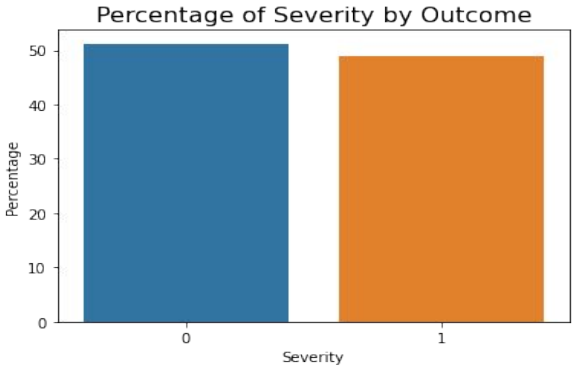
Exploratory Data Analysis

- Shape: Highest category is 4 (irregular)
- Density: Highest category is 3 (low)
- Severity: Almost even split
- Age: Most data falls into ages 60 to 70



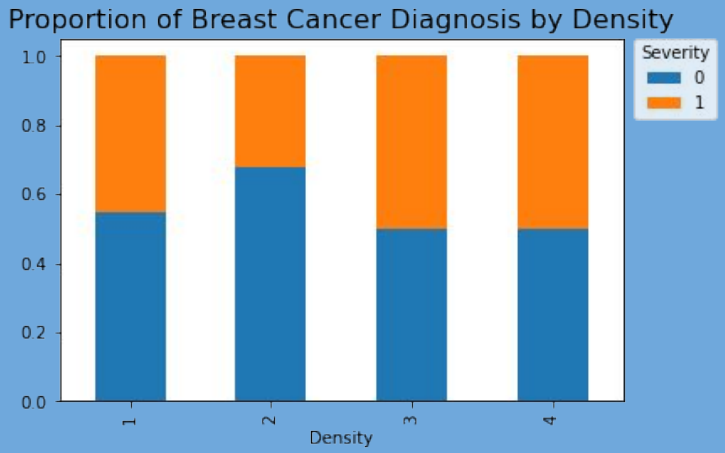


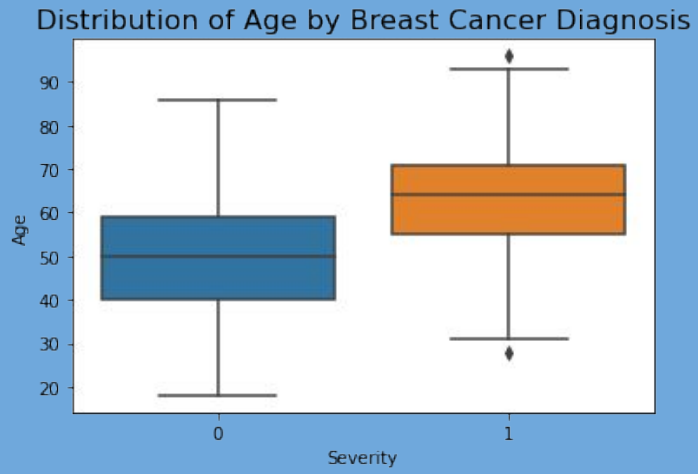




Relationship between Predictors and Response

- Density vs Severity
  - slightly higher proportion of malignant cases in levels 1 and 2
- Distribution of Age
  - Severity 0 is classification for the younger population
  - Severity 1 is classification for older population



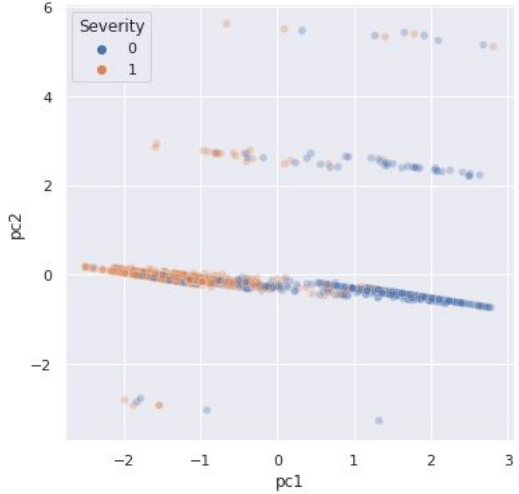


Principal Component Analysis

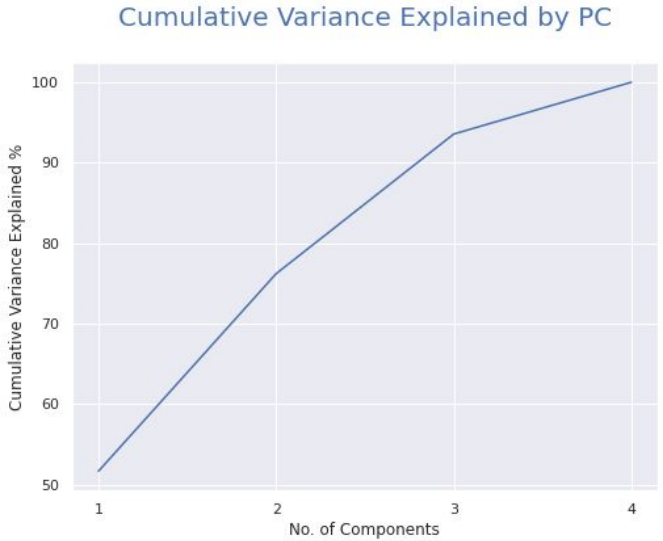
	pc1	pc2	pc3	pc4
Age	-0.469405	-0.609182	-0.623743	-0.139644
Shape	0.127863	0.094379	0.032516	-0.986756
Margin	0.871943	-0.385116	-0.294938	0.066432
Density	-0.055008	-0.686786	0.723117	-0.048988

Identifies key variables or presence of outliers

a. First component Margin had the strongest effect
b. Second and third component Age and Density had the strongest effect
c. Fourth component Shape had the most impact



Shows that the two classes are separable



Shows that the first 3 components explain more than 90% of the variance

One-Hot Encoding

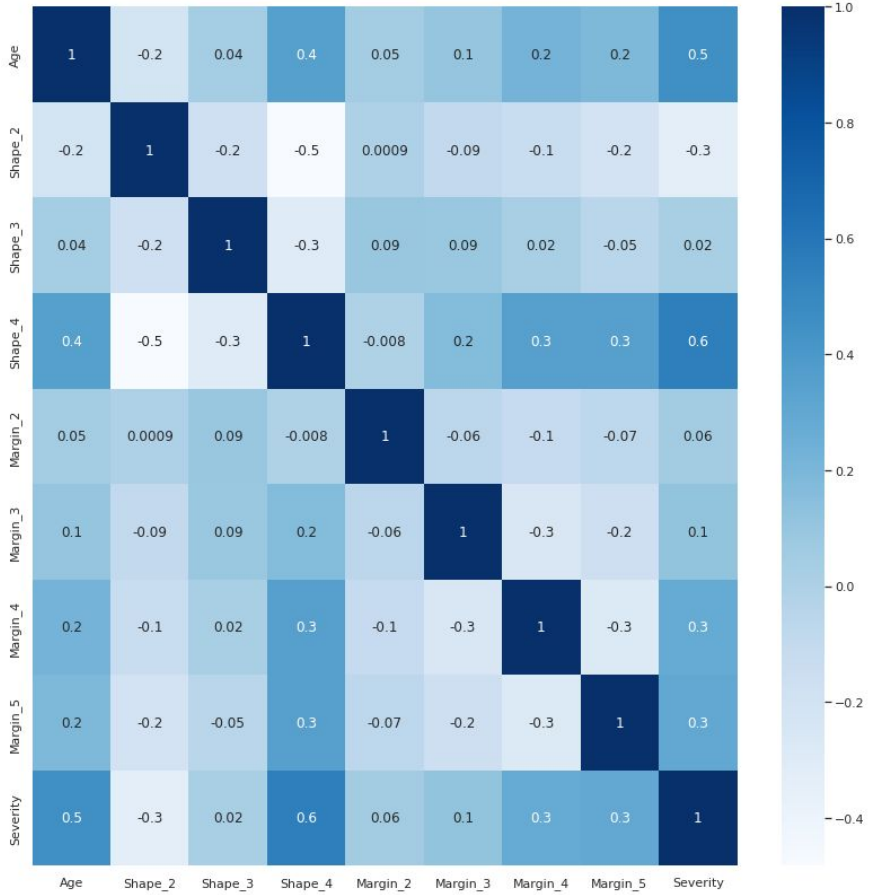
Applied to nominal variables so each level becomes a new column with values 0 and 1 assigned, and one level is dropped to prevent collinearity

This was applied to the Shape and Margin features

The resulting dataset was used for the remainder of the analysis

	Age	Density	Shape_2	Shape_3	Shape_4	Margin_2	Margin_3	Margin_4	Margin_5	Severity
0	67	3	0	1	0	0	0	0	1	1
2	58	3	0	0	1	0	0	0	1	1
3	28	3	0	0	0	0	0	0	0	0
8	57	3	0	0	0	0	0	0	1	1
10	76	3	0	0	0	0	0	1	0	1

Correlation Matrix



- Used to identify strong correlations between features
- Stronger relationship = darker shade
- No significant correlations identified

3. Model Planning and Evaluation

Splitting the Data

- Training Data:
  - 80% of data (669 observations) assigned to the training set
- Test Data:
  - 20% of data (168 observations) assigned to the test set
- Scaling Data:
  - To reduce range variations and improve model performance
- Cross-validation used:
  - Training data was further divided into training and validation data used for optimizing model fit

Model 1: Logistic Regression

What is Logistic Regression?

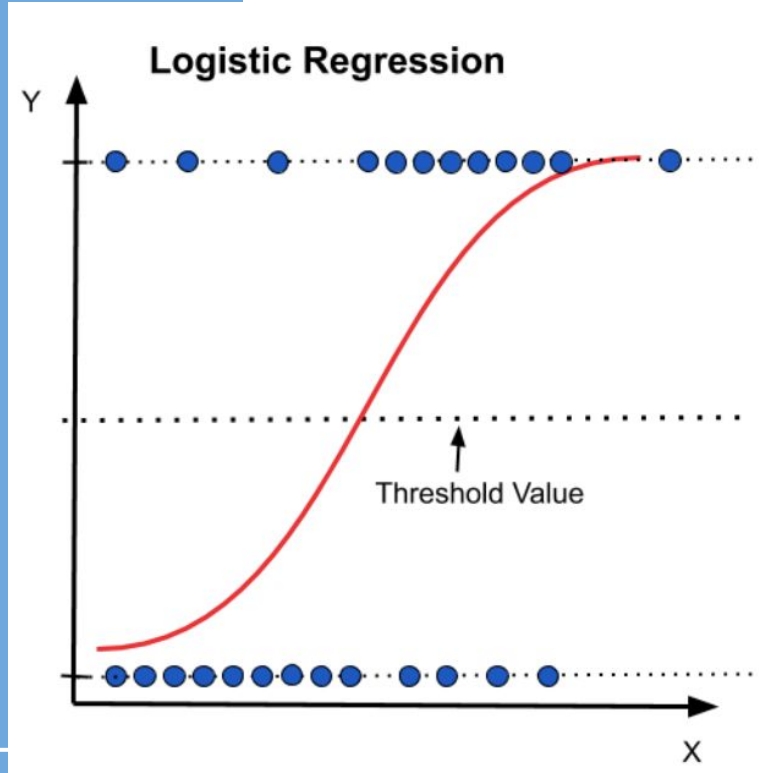
- Determine the probability that a record belongs in one of these categories (benign or malignant)
- Threshold used to reduce occurrence of false positives and false negatives

Why we chose it?

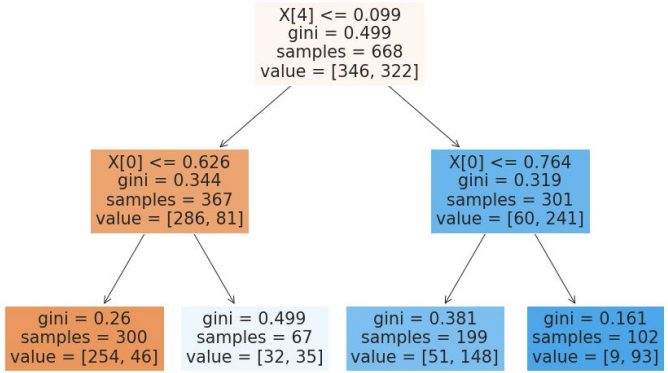
- Easy to implement and interpret
- Makes no assumptions on distributions

Results

- Hyperparameters: Penalty = LASSO, Solver = liblinear, C = 1
  - AUC = 0.87
  - Test Accuracy = 79.8%



Model 2: Decision Tree



What is Decision Tree?

- Recursively divides inputs of data into decisions to create classifications
- Nodes represent features, branches represent decisions, leaves represent outcomes

Why we chose it?

- Easy to interpret
- Although they may result in overfitting, pruning can be done

Results

- Hyperparameters: Max depth = 2, Min sample for split = 2, Min sample leaf = 10, Criterion = gini
  - AUC = 0.79
  - Test Accuracy = 76.2%

Model 3: K- Nearest Neighbours

What is KNN?

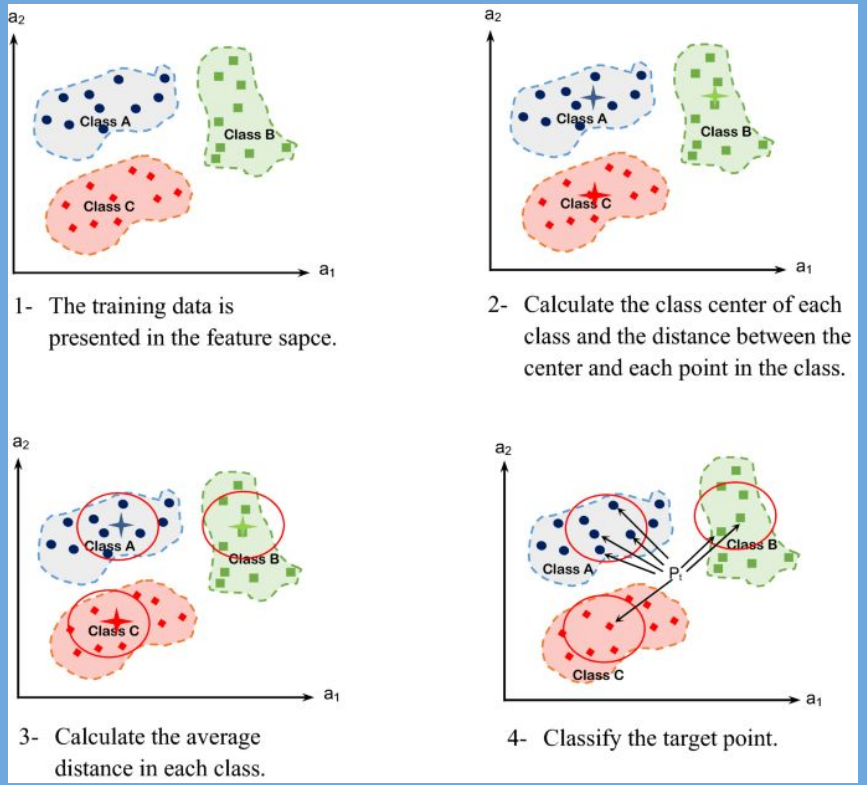
- Groups observations close to each other based on proximity to other similar records

Why we chose it?

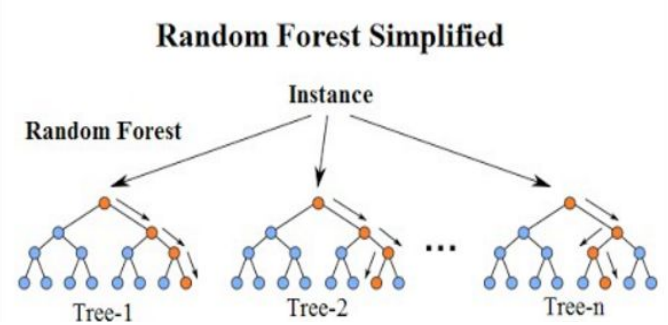
- Easy to implement
- Can evolve with new data

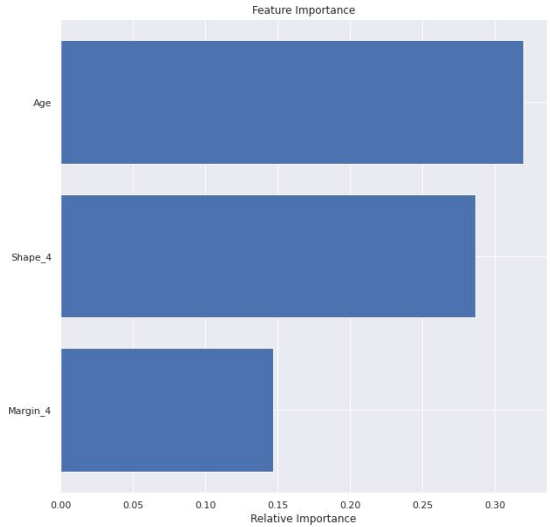
Results

- Hyperparameters: metric = manhattan, number of neighbours = 5, weights = uniform
  - AUC = 0.84
  - Test Accuracy = 81.0%



Model 4: Random Forest





What is Random Forest?

- Combines multiple decision trees
- Each decision tree uses subset of features
- Results of each tree aggregated to produce outcome

Why we chose it?

- Helps to reduce overfitting and variance
  - But can be time-consuming

Results

- Hyperparameters: Number of estimators = 12, Max depth = 3, Criterion = gini
  - AUC = 0.85
  - Test Accuracy = 80.4%

Model 5: Support Vector Machine

What is SVM?

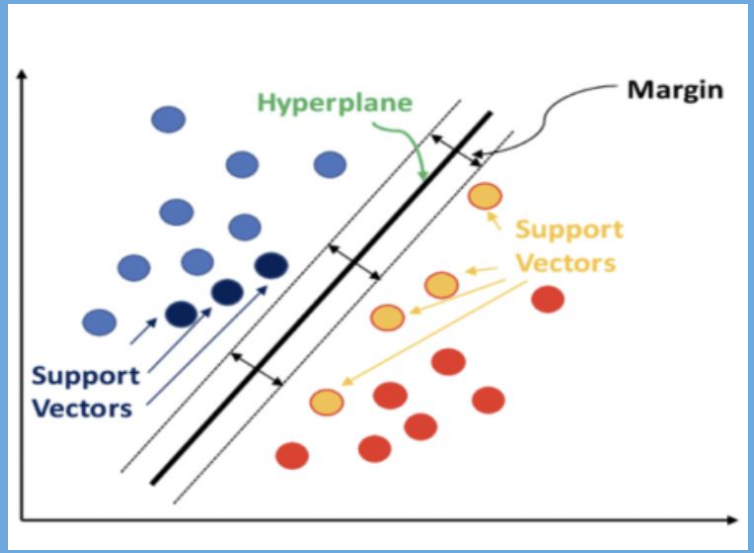
- Uses kernels to map data to high dimensional feature space

Why we chose it?

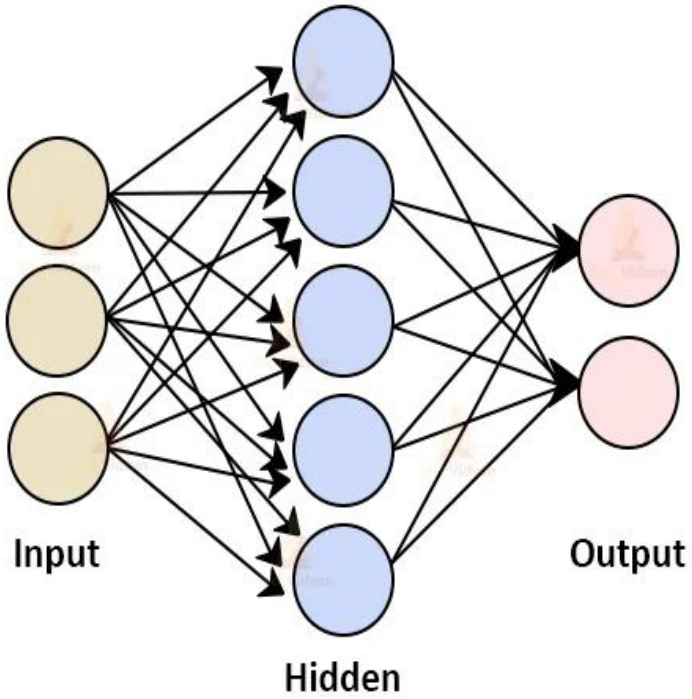
- Memory efficient
- Clear margin of separation between classes

Results

- Hyperparameters: C = 3, Degree = 3, Gamma = scale, kernel = rbf
  - AUC = 0.83
  - Test Accuracy = 79.8%



Model 6: Artificial Neural Networks



What is Artificial Neural Networks?

- Form of AI that uses layers (input, output, hidden) and is connected by nodes

Why we chose it?

- Learns from the data to perform better
- However, can be time-consuming

Results

- Hyperparameters: Hidden layer sizes = 9, Activation = logistic, Solver = adam, Max iteration = 500
  - AUC = 0.86
  - Test Accuracy = 80.4%

4. Results

Metrics Using BI-RADS Assessment

- Current method for classifying mammogram masses based on BI-RADS assessment assigned by physician:
  - BI-RADS assessment of 1, 2, 3 : benign
  - BI-RADS assessment of 4 and 5 : malignant (biopsy recommended)

Accuracy	48.8%
Precision	47.7%
False Alarm Rate	91.6%
False Negative Rate	1.6%

Comparing the Models using Metrics

Model	Accuracy	Precision	False Alarm Rate	False Negative Rate
Logistic Regression	80.0	76.0	29.0	12.0
Decision Tree	76.0	73.0	33.0	15.0
K-Nearest Neighbours	81.0	77.0	28.0	10.0
Random Forest	80.0	77.0	28.0	12.0
Support Vector Machine	80.0	76.0	29.0	12.0
Artificial Neural Network	80.0	77.0	27.0	13.0

Comparing Training vs Test Accuracy

Train accuracies are slightly higher than test accuracies

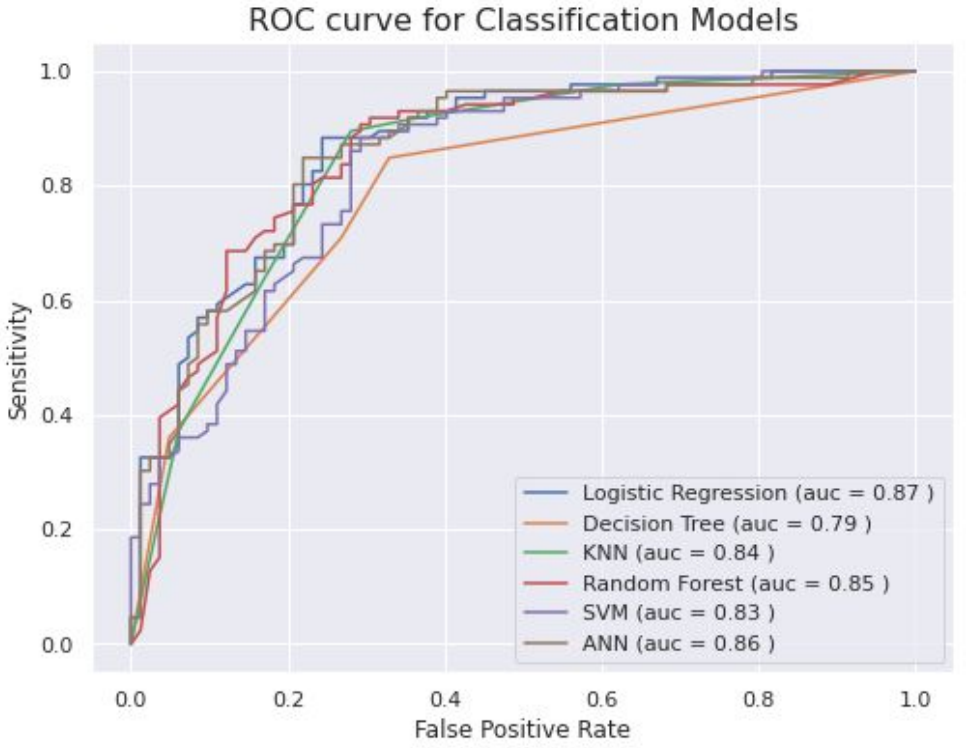
- Overfitting
- Low Bias, High Variance

Model	Training Accuracy	Test Accuracy
Logistic Regression	81.9	79.8
Decision Tree	79.3	76.2
K-Nearest Neighbours	84.3	81.0
Random Forest	81.4	80.4
Support Vector Machine	82.8	79.8
Artificial Neural Network	82.2	80.4

ROC Curve

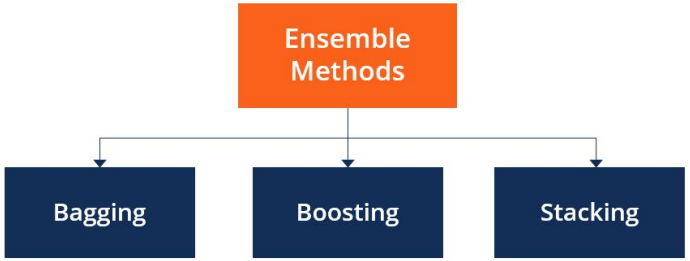
ROC Curve: plots false positive rate against the true positive rate

- Logistic has the highest Area under the Curve
- Decision tree has the lowest Area under the Curve



5. Conclusion

Findings



Machine Learning tools outperforms the previous way of classifying BI-RADS

Logistic regression performed the best with respect to AUC

Future Enhancements:

- Acquire more data
- Expand number of features
- More Ensemble methods

Contributions

Myisha Chaudhry	Worked on Pre-processing, EDA, Comparing models, created ROC curve Created classes for Decision Tree, Random Forest and SVM Worked on the report Worked on slides presented
Kayleigh Habib	Worked on Pre-processing, EDA, Comparing models, created ROC curve Created classes for Logistic Regression, KNN and ANN Worked on the report Worked on slides presented

Live Demonstration