Breast Cancer Classification

Myisha Chaudhry and Kayleigh Habib

CP322: Machine Learning, Wilfrid Laurier University

Wilfrid Laurier University

Dr. Sukhjit Sehra

April 06, 2023

Table of Contents

List of Figures

List of Tables

Abstract

Mammograms check for the presence of unusual masses in the breast, and doctors use these results to recommend whether further tests are required. Current assessment approaches are not very effective, resulting in a large number of unnecessary procedures. The purpose of this project is to determine whether machine learning models can successfully predict the possibility of breast cancer based on the attributes from patients' mammogram results.

As this project is a continuation of a previous project worked on in CP468, we decided to focus on an additional three models as well as the three looked at last time. Thus in total we analyzed six models: Logistic regression, Decision trees, K-nearest neighbours, Random forests, SVM and ANN. By applying cross-validation techniques, we trained models using a mammogram mass dataset retrieved from the UCI Machine Learning Repository. We then used each model type to predict outcomes for a test set of data, and compared the results across models, as well as to the BI-RADS assessment. We also decided to perform Principal Components Analysis but found no significant changes to our models and results, and thus for this reason, chose to not apply it to our final analysis.

The results indicated that each of the models outperformed the BI-RADS assessment with respect to accuracy and precision. These performance measures are moderate and further analysis could be performed to improve the predictive power of the machine learning approach, including expanding the data available, assembling models to strengthen predictions, and using ensemble of models, i.e. weighted combinations of different model types.

1.0 Introduction to Project

1.1 Overview

According to the Canadian Breast Cancer Society, about 1 in 8 women will develop breast cancer, but mortality rates due to this disease have been decreasing in recent years. This can be attributed to early detection through the use of mammograms. While confirming a breast cancer diagnosis requires a biopsy, breast cancer screening can be used to determine whether such a procedure is necessary. Consequently, accurate classification of mammogram results is necessary for both early detection and to avoid unnecessary invasive procedures.

1.2 Existing System

Mammograms use x-ray machinery to produce images of breast tissue which are then assessed by a doctor to determine whether further tests are required. The Breast Imaging Reporting and Data System (BI-RADS) is used, and according to the American Cancer Society, the assessment is interpreted as follows: 0 - incomplete i.e. further images are required; 1 - negative findings i.e. no traces of breast cancer; 2 - benign findings i.e. some cause for concern but it is not cancerous; 3 - likely benign (<2% probability of malignancy) i.e. low chance for cancer; 4 - suspected malignant and a biopsy test is recommended; 5 - highly likely to be malignant; 6 - identified as malignant previously, the patient has had a biopsy or is undergoing treatment.

1.3 Objectives of Project

The objective of this project is to use machine learning models to predict the possibility of breast cancer based on the attributes from patients' mammogram results. Creating a reliable prediction model will help improve the accuracy rate of breast cancer detection, so that positive cases can be detected early, while unnecessary procedures could be avoided in negative cases.

2.0 Pre-Processing and Exploratory Data Analysis

2.1 Dataset Collection

The data used was retrieved from the University of California, Irvine (UCI) machine learning repository. It consisted of mammogram mass data collected by the Institute of Radiology at the University Erlangen-Nuremberg in Germany.

2.2 Data Pre-processing

The data includes 961 records, with 6 columns as described in Appendix A. The data contained records with missing attribute information. We replaced missing Age values with the median age, and removed any remaining records that still had a missing value. Thus, 836 records remained in our data for further analysis.

2.3 Exploratory Data Analysis and Visualizations

We used visualization techniques to examine the distribution of the different attributes in the data. Looking at the target variable (severity) we found that 51% of the cases were classified as benign, while 49% were malignant. Thus, the data is very balanced with respect to outcomes. For the remaining attributes or predictors, we did not observe any outliers in the data. The variable distributions are shown in Appendix B.

We also looked at the relationship of each predictor to the target variable using bar plots and boxplots. For the Density variable, there seemed to be only a slightly higher proportion of cases in levels 1 and 2 that were malignant, while the other levels were more evenly split between malignant and benign.

Finally, we used a heatmap to look at the correlation between variables, where a correlation value close to +/-1 indicates a strong relationship. If two predictor variables are highly correlated, including both in the analysis may add complexity without material model improvement. This analysis did not indicate any significant correlations, so no variables were removed. The charts produced by these analyses are included in Appendices C and D.

2.4 Feature Engineering

2.4.1 Feature Selection

As the BI-RADS assessment variable is assigned by the attending physician based on the mammogram mass results, we determined that this attribute is not a predictive variable, so we removed it from the data for analysis. All of the other features i.e. Age, Density, Margin and Shape, were used as predictors in our models. As we would be using classification models in our analysis, we determined that it would be appropriate to scale the features so that the models would train more efficiently and have improved accuracy.

2.4.2 Principal Components Analysis

Principal components analysis (PCA) is an unsupervised learning method that can be used to identify key variables, or to identify the presence of outliers in the data. The mammogram mass data has 4 features, so it is not high-dimensional, but we explored the application of PCA to better understand the relationships between the features. Based on this analysis: for the first component - margin has the strongest effect; for the second and third components - age and density have a strong effect; and for the fourth component - shape has the most impact. The cumulative variance for the components is shown in Appendix E, which indicates that over 90% of the variance is explained by the first 3 principal components. While not a material decrease from all 4 features available, it is possible to use PCA to decrease model run time without material decrease in performance.

2.4.3 One-Hot Encoding

One-hot-encoding is applied to nominal variables to transform them, so each level of the feature becomes a new column with a value of 0 or 1 assigned, based on whether the level exists for a specific record. In order to remove collinearity, one level (i.e. new column) of the transformed variable is dropped from the data. We applied One-Hot Encoding to the nominal fields Shape and Margin. The resulting dataset was used for the remainder of this analysis.

3.0 Methodology

3.1 Introduction to Python for Machine Learning

To perform analysis on the data and create models, python classes were created. There were various different classes created to manipulate/clean the data, perform feature engineering, PCA, and EDA, produce training and test data, produce the actual models, evaluate the models, and produce an ROC curve. In addition to these classes, the various python libraries used for this project were Numpy, Pandas, MatplotLib and Sklearn.

3.2 Platform and Machine Configurations Used

The platform and machine configurations used were Syzygy (Jupyter Notebook) and Visual Studio Code.

3.3 Data Split

The first step involved splitting the data into two parts so that 80% of data (669 observations) was assigned to the training set used for building the models, and 20% of data (168 observations) assigned to the test hold-out data to be used only for evaluating the performance of the models. If all the data had been used for training the models, this would result in the models 'learning' this data so that they can be tuned to a high accuracy for this data but perform poorly on unseen data thus resulting in overfitting. For the train-test split, a random state of 42 was set to ensure reproducibility of results.

3.4 Model Planning

As the goal is to predict whether a mammographic mass is benign or malignant using structured data, this is a binary classification problem and supervised machine learning classification algorithms were considered. The previous analysis of this data explored 3 models: Logistic Regression, Decision Trees, and K-Nearest Neighbours. For this project, we examined these 3 model types, and extended the analysis to include 3 additional methods: Random Forests, Support Vector Machines and Artificial Neural Network.

3.5 Model Training

For each of the 6 models, the Training class first initializes the models by passing the parameters that would create an instance of the specific classifier using default parameters. The model was fitted using the training data that included a dataframe with the available features, and an array with the corresponding target variable for each record.

Each model training class includes a function to predict the outcomes for the training data. A scoring function then compares these predictions to their corresponding known outcomes. The output of this function is an accuracy measure determined as the proportion of correct predictions made using the model parameters.

3.6 Model Evaluation

The Predict class was used to evaluate the models using the test data that had been set aside. The tuned model was used to predict the outcomes on the test data, which were then compared to the known outcomes. A confusion matrix and a classification report displaying the relevant metrics are produced, and these are included in Section 4.3 of this report.

In order to classify predictions as benign or malignant, a threshold of 0.5 was used. Varying this threshold would alter the number of records assigned to each class. To explore the impact of changing the threshold level, a Receiver Operating Characteristic (ROC) curve is used. This is discussed further in Section 4.2.

3.7 Model Optimization

Each of the models was optimized using a function in the Training class to tune the appropriate hyperparameters based on model type. We applied a grid search with cross-validation, which uses the model, a grid which includes a range or list of parameters to be tested, and the number of folds to be used for cross-validation.

Under this cross-validation approach, the training data is further subdivided into train and validation data by resampling. The training data is randomly split into subgroups of equal

size, with each one having a turn as the validation data, while the remaining groups are used to fit the model. Thus, each record in the training data will be part of validation data one time and be part of the train data the remaining times.

The result of this optimization step would be the model fit that results in the highest training accuracy. For each model, the parameters that were tuned and the final parameters with the best fit for each classification algorithm are shown in Appendix F.

3.8 Final Model Building

For small datasets, some models can suffer from overfitting, when the model attempts to fit to the noise in the data. This results in low bias and high variance, and the model does not generalize well on unseen data. A material decrease in accuracy from training data to test data is an indication of possible overfitting. As a final assessment of the tuned models, the training data accuracy was compared to the test data accuracy.

Table 1 shows that for all the models produced the training accuracy proved to be higher than the test accuracy, so our models may be slightly overfitting our data. The model with the largest difference in training and test accuracy is the KNN model (84.3% vs 81.0%). Since the KNN model also has the highest test accuracy, this model may not actually be the best performing model as there is a risk that it simply memorized the patterns in the training data and thus is not performing as well on the unseen test data.

Table 1: *Comparison of Training versus Test Accuracy*

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

4.0 Results

4.1 Description of Models

4.1.1 Logistic Regression

Logistic regression uses a sigmoid function to determine the probability that a record belongs to one of the classes. To assign the record to a specific class, we set a cut-off or threshold, so if the probability exceeds this amount, the record will be labeled as "1" i.e. malignant. The threshold can be set to reduce the number of false positives or false negatives.

4.1.2 Decision Tree

A decision tree works by recursively dividing the inputs into decisions, where nodes represent features, and branches are the decisions made to divide the data. The leaves of the tree represent the outcome for each record. Decision tree models are very interpretable, but often suffer from overfitting. This can be overcome by pruning the tree, or by setting limits on the number of leaf nodes.

4.1.3 K-Nearest Neighbours

The K-nearest neighbours approach groups individual observations into categories based on its proximity to other similar data records. This methodology is relatively simple to implement due to the small number of parameters but does not perform well when data is high-dimensional as it tends to overfit.

4.1.4 Random Forest

Random Forests combine multiple decision trees, each of which uses a random subset of features, and then aggregate the results of the trees to produce the outcome. Random forest models help reduce overfitting and variance but can be time-consuming to run.

4.1.5 Support Vector Machine

Support Vector Machines categorize data by using kernels to map the data to a high-dimensional feature space. A separator between the categories is drawn as a hyperplane. Although SVMs are memory-efficient, they do not work well on large datasets.

4.1.6 Artificial Neural Network

Neural Networks are a form of artificial intelligence that use layers (input layer, output (or target) layer and a middle-hidden layer) that are connected by nodes to form a "network". Due to the hidden layers, neural networks may be more time-consuming and are often difficult to explain.

4.2 Performance Metrics

To evaluate and compare the model performances, we produced a confusion matrix, a classification report and the ROC curve for each of the algorithms.

A confusion matrix is used to categorize counts that compare predicted classes versus actual classes. Using these values, the following can be calculated:

- Accuracy = (TP+TN)/(TP+TN+FP+FN) i.e. proportion classified correctly
- Recall = TP/(TP+FN) i.e. proportion of malignant masses correctly
- Precision = TP/(TP+FP) i.e. proportion of masses classified correctly
- False alarm rate = FP/(FP+TN) i.e. proportion of benign masses incorrectly classified
- False negative rate = FN/(FN+TP) i.e. proportion of malignant masses incorrectly classified

There are two types of errors that could occur: False positive (FP) which represents a benign mass is misclassified as malignant and False negative (FN) which represents a malignant mass is misclassified as benign. While any misclassifications are undesirable, in the case of mammogram classification, a False Negative is more detrimental than a False Positive. Thus, the model performances will also be assessed on this measure.

Plotting the sensitivity measure against the false positive rate produces the Receiver Operating Characteristic (ROC) curve. This curve can be used to determine the optimal threshold at which the classes should be defined. The Area Under the Curve (AUC) provides the total measure of performance across all possible thresholds for a model. A higher value for AUC indicates a better performing model.

4.3 Results Table

Table 2: *Confusion Matrix metrics using test data*

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

Table 2 shows a summary of the classification metrics for each model. The KNN model has the highest test accuracy, while the Decision Tree model has the lowest. In addition, the false alarm rate and false negative rate for the KNN are among the lowest of all the models analyzed.

4.4 Interpretation of Results

4.4.1 Model Performance

As stated in section 4.3, the model that performed the best on the test data is the KNN model. KNN algorithms are some of the simplest to implement and are known to perform well with smaller well labeled datasets as in this analysis. For some of the other classification models explored, having more features (i.e. high-dimensional data) and records would have helped improve their performance.

4.4.2 ROC Curve

When analyzing the ROC curve, the models that have a higher Area Under the Curve (AUC) perform better than other models as they have a higher predictive power. For the ROC curve created based on the 6 models in this analysis, Logistic Regression had the highest AUC thus meaning this model performed the best. While the Decision Tree is the worst performer (0.79). The KNN model had the highest overall test accuracy, and an AUC of 0.84.

4.4.3 BI-RADS Assessment

In addition to comparing the models to one another, we also looked at how they compared to the BI-RADS assessment. The metrics for the BI-RADS assessment (on the original data after removing missing values) are shown in Table 3:

Table 3: *Metrics for BI-RADS Assessment*

Accuracy	Precision	False Alarm Rate	False Negative Rate
48.8%	46.6%	91.6%	1.6%

In all cases, the predictive models outperformed the BI-RADS assessment with respect to both accuracy and precision. In addition, each of the models had a much lower false alarm rate than the BI-RADS assessment, so fewer patients were misclassified as having breast cancer. The BI-RADS assessment had a very low false negative rate, although this is a consequence of classifying a large proportion of cases as malignant.

4.4.4 FAR and FNR for Models

Comparing the False Alarm Rate (FAR) and the False Negative Rate (FNR) for all models, we note that the Decision Tree has the highest FAR and predicts the most false positives (classifying benign masses as malignant). The Neural Network model has the lower FAR, so it is predicting the least false positives. Additionally, the Decision Tree has the highest FNR, and is incorrectly classifying a higher proportion of malignant classes as benign

than any other model. The KNN model has the lowest FNR, so this model is incorrectly classifying a lower proportion of malignant classes as benign than any other model.

4.5 Sensitivity Analysis

As part of our analysis, we explored using Principal Components Analysis to reduce the number of inputs to the model. Using 3 principal components would have accounted for over 90% of the variation in the data but would not have materially impacted the model run-time. In addition, preliminary results did not indicate any improvement in model accuracy level, and as such, we did not apply this feature reduction technique.

We also tested using different parameters for the model fitting process in order to determine which ones should be included in the tuning process. In all cases, this resulted in reducing the size of the parameter grid used in the grid search process, and thus helped to reduce model run-time while still resulting in optimal performance.

5.0 Conclusion

Our analysis indicates that machine learning methods appear to outperform BI-RADS assessment in successfully classifying mammogram masses as benign or malignant. Based on the AUC value, the logistic regression model has more predictive power than the other models , although the KNN model has the overall highest test accuracy. With more time, there are several approaches that can be used to get more predictive value:

1. Obtain more data so the models can be trained on a larger number of observations.
2. Try to expand the number of features. This may include either feature engineering, or additional measurements of masses, such as radius or location.
3. Use an ensemble of models, i.e. weighted combinations of different model types.

References

Gour, R. (2019, May 7). *Artificial Neural Network for machine learning - structure & layers*. Medium. Retrieved April 6, 2023, from https://medium.com/javarevisited/artificial-neural-network-for-machine-learning-structure-layers-a031fcb279d7

Habib, K., & Szeto, W. (2022). *Breast Cancer Classification*. Computer Science Department, Wilfrid Laurier University.

Lee, S. (n.d.). *Breast cancer statistics*. Canadian Cancer Society. Retrieved April 6, 2023, from https://cancer.ca/en/cancer-information/cancer-types/breast/statistics

Limitations of Mammograms. How Accurate Are Mammograms? (n.d.). Retrieved April 6, 2023, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html

McGregor, M. (2020, July 2). *SVM machine learning tutorial – what is the support vector machine algorithm, explained with code examples*. freeCodeCamp.org. Retrieved April 6, 2023, from https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

UCI Machine Learning Repository: Mammographic mass data set. (n.d.). Retrieved April 6, 2023, from https://archive.ics.uci.edu/ml/datasets/mammographic+mass

Understanding your mammogram report. Mammogram Results. (n.d.). Retrieved April 6, 2023, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/understanding-your-mammogram-report.html

Yiu, T. (2021, September 29). *Understanding random forest*. Medium. Retrieved April 6, 2023, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Appendix

Appendix A: Data Attribute information

Data Field	BI-RADS assessment	Age	Shape	Margin	Density	Severity (Target)
Data Type	ordinal	integer	nominal	nominal	ordinal	boolean
Data Description	Scale of 1 to 5 determined by medical examiner	Patient's age in years	Round = 1 Oval = 2 Lobular = 3 Irregular = 4	circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5	high=1 iso=2 low=3 fat-containing=4	benign=0 malignant=1

Table A1: Data attributes in mammogram mass data

Appendix B: Distribution of Features in Data

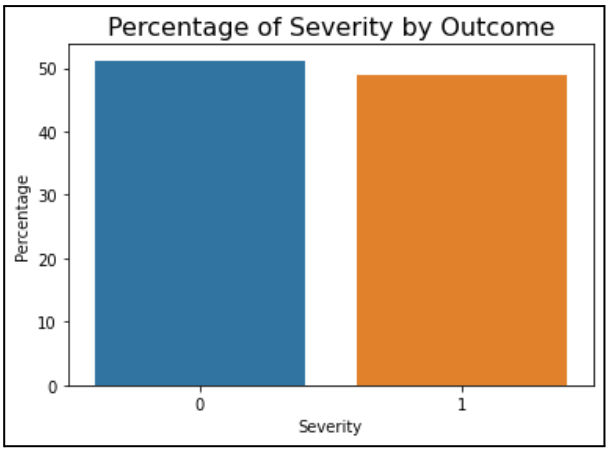
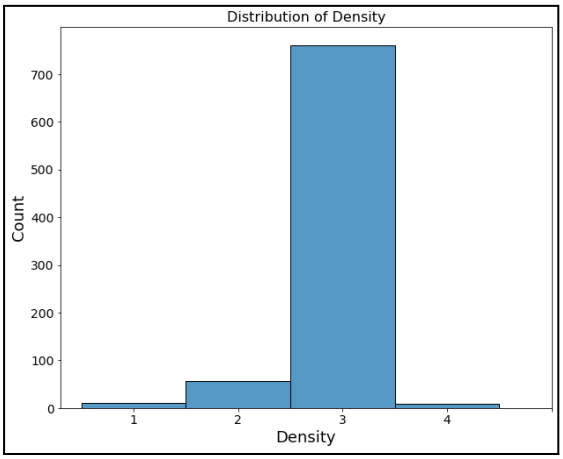
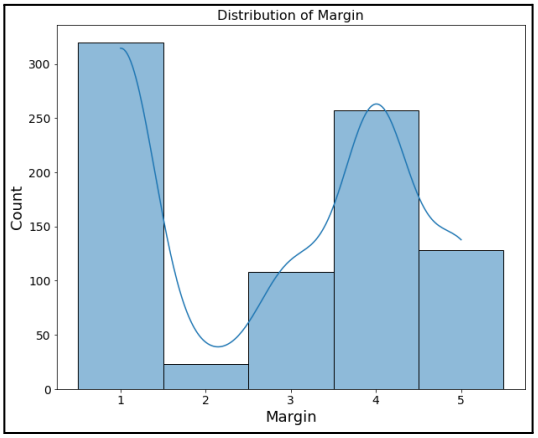
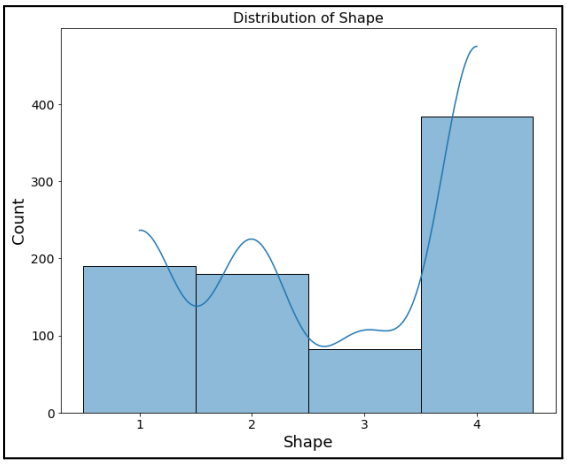


Figure B1: Distribution of Target Variable







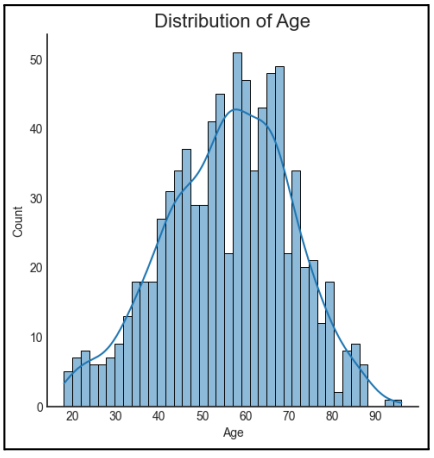


Figure B2: Distribution of Predictor Variables

Appendix C - Relationship of Features

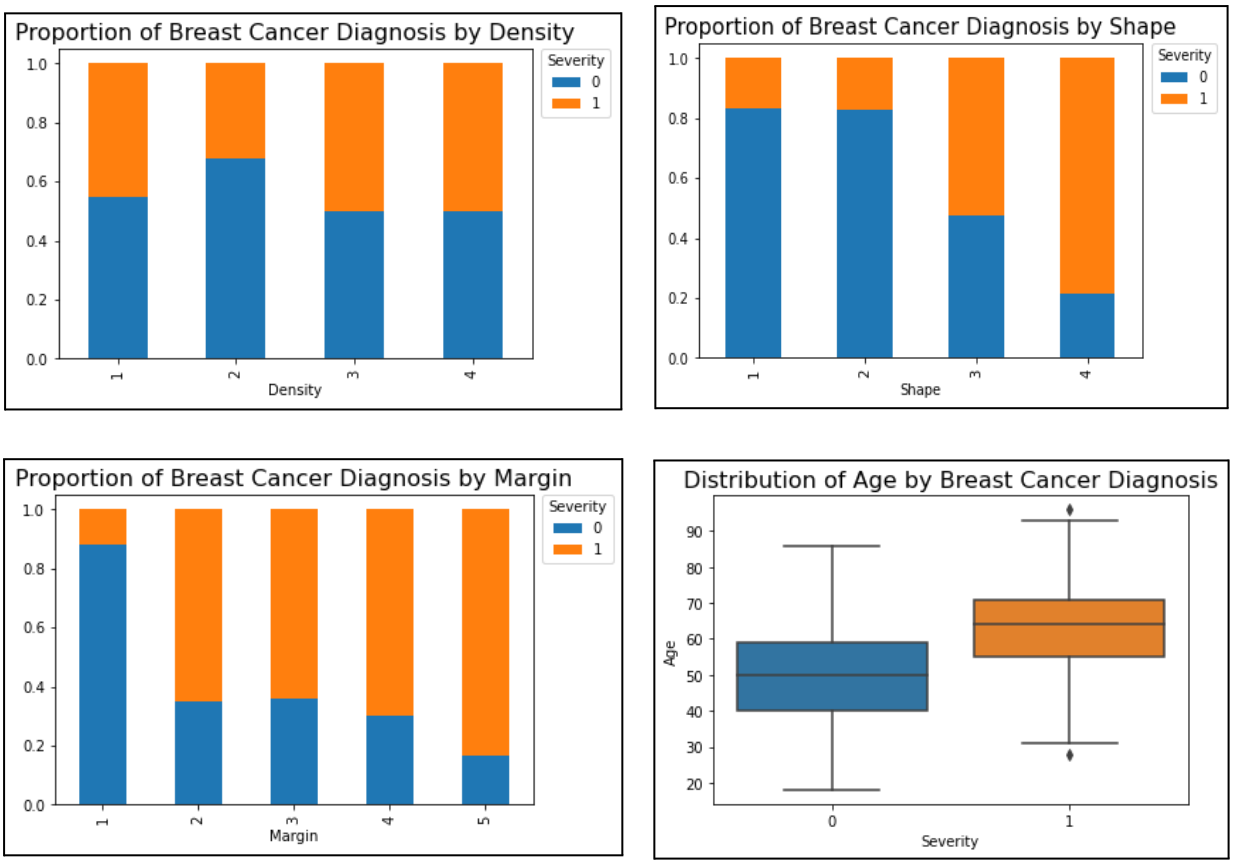


Figure C1: Relationship of Predictors to Target

Appendix D - Correlation Matrix

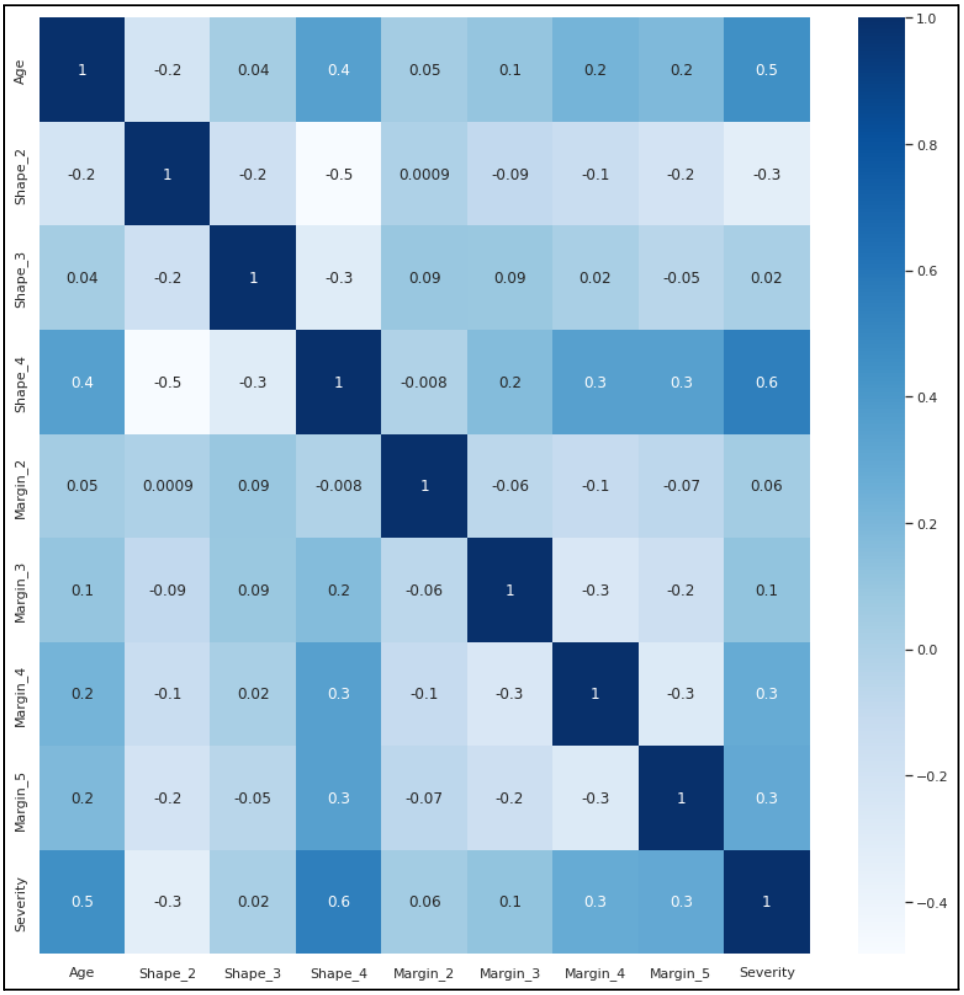


Figure D1: Correlation Matrix showing pairwise relationships

Appendix E - PCAs

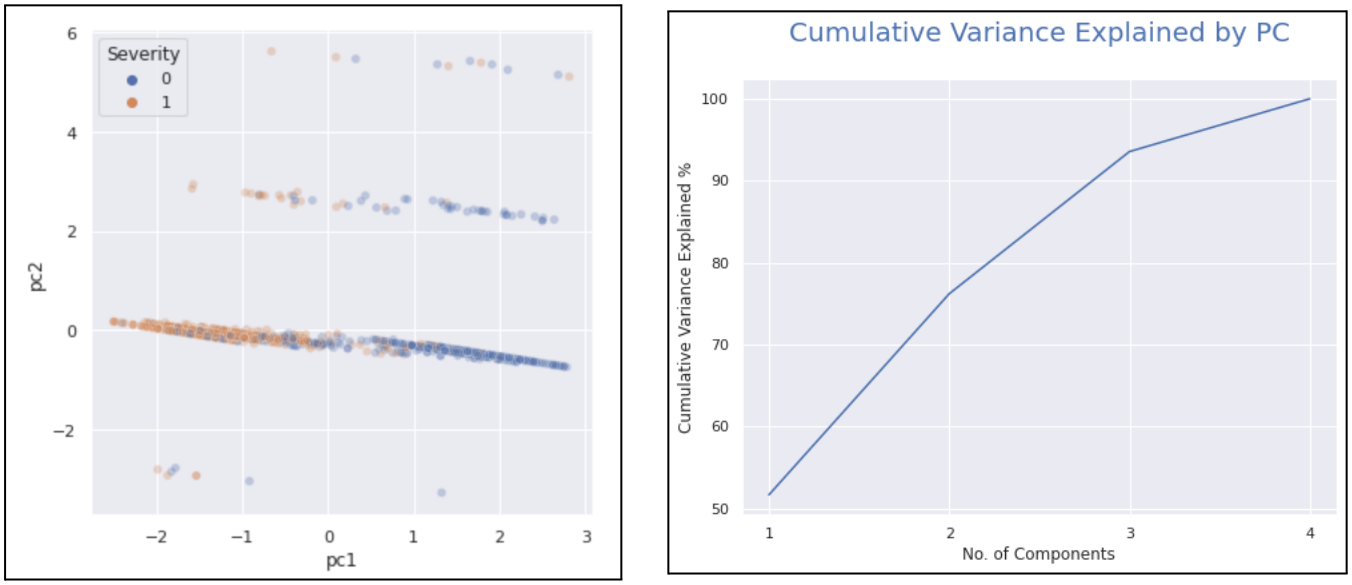


Figure E1: PCA

Appendix F - Hyperparameters for model tuning

Model	Logistic Regression	Decision Trees	K-Nearest Neighbours	Random Forest	Support Vector Machines	Artificial Neural Networks
Range of Parameters Tested	Penalty = ridge (l2) or lasso (l1) C = 0.1 to 1 Solver = (liblinear, newton-cg, lbfgs)	Maximum depth = 2 Minimum sample for split = 2 to 12 by increments of 1 Minimum sample leaf = 10, 15 Criterion = gini	k = 2 to 11 in increments of 3 Weight = uniform or distance Metric = minkowski, manhattan	Number of estimator = 2, 5, 12 Maximum depth = 1 to 5 Criterion = gini or entropy	C = [0.01, 3,10] Kernel = rbf, polynomial Gamma = scale, auto Degree = [3,5]	Hidden Layer Sizes = 9 and 15 Activation = identity, logistic, relu Solver = adam Maximum Iteration = 500, 1000
Final Parameters	Penalty = lasso (l1) C = 1 Solver = liblinear	Maximum depth = 2 Minimum sample for split = 2 Minimum sample leaf = 10 Criterion = gini	k = 5 Weight = uniform Metric = manhattan	Number of estimator = 12 Maximum depth = 3 Criterion = gini	C = 3 Kernel = rbf Gamma = scale Degree = 3	Hidden Layer Sizes = 9 Activation = logistic Solver = adam Maximum Iteration = 500

Table F1: Summary of Hyperparameters used in model tuning

Appendix G - Model Evaluation Metrics

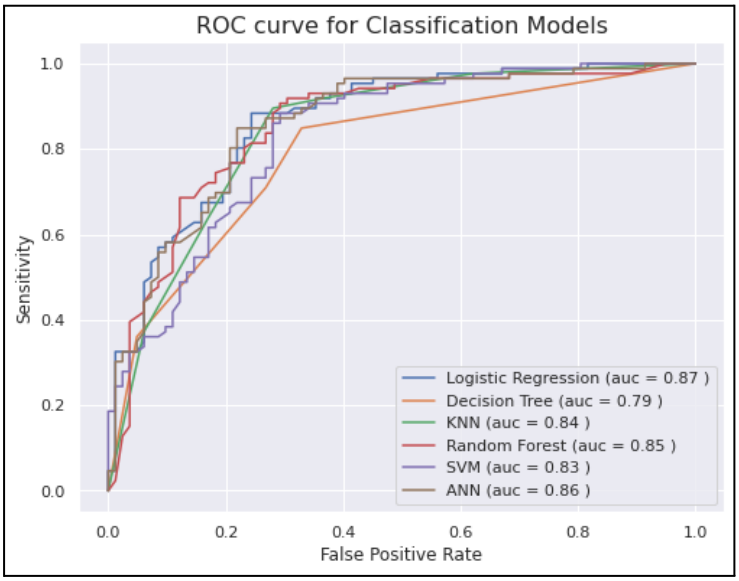


Figure G1: ROC Curves and AUC metrics for final models