Breast Cancer Classification

Kayleigh Habib and Winnie Szeto

CP468: Artificial Intelligence

Wilfrid Laurier University

Dr. Sukhjit Sehra

July 19, 2022

Table of Contents

Executive Summary

Breast cancer accounts for 14% of all cancer deaths in Canada but can be successfully treated if detected early. 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.

While many machine learning methodologies are available, we focused on three types : Logistic regression, Decision trees, and K-nearest neighbours. 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. The results indicated that each of the models outperformed the BI-RADS assessment with respect to accuracy and precision, and greatly reduced the proportion of cases incorrectly classified as malignant.

Metric	BI-RADS Assessment	Logistic Regression	Decision Tree	K-Nearest Neighbours
Accuracy	49%	78%	74%	79%
Precision	47%	72%	68%	72%
False Alarm	92%	29%	35%	31%
False Negative	4%	14%	16%	10%

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 exploring other model types.

1.0 Introduction

1.1 Purpose

According to the Canadian Breast Cancer Society, it is estimated that in 2022 alone, approximately 28,600 women and 270 men will be diagnosed with breast cancer. In addition, breast cancer deaths account for 14% of all cancer deaths in Canada. Breast cancer can be successfully treated, and early detection is key. While the only definitive means of confirming a breast cancer diagnosis is through a biopsy, breast cancer screening can be used to determine whether such a procedure is necessary. A mammogram can be performed to check for the presence of unusual masses in the breast and based on these results a doctor will recommend whether further tests are required using an approach known as Breast Imaging Reporting and Data System (BI-RADS). Mammograms use x-ray machinery to produce an image of your breast tissue by flattening the breasts. The image produced by this technique is used to look for breast cancer cells.

The purpose 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 detect the cancer in the earlier stages and reduce the need for unnecessary invasive procedures.

1.2 Scope

The mammogram mass data used included information about each patient's age, as well attributes about the mass, including its shape, margin, and density. The data also includes the severity, which is the diagnosis (i.e. benign or malignant) for each record. Using this data, we will explore three different classification models that can be used to predict the category to which each mammogram mass record will belong based on the attributes provided. These models include logistic regression, decision trees, and K-nearest neighbours. All analysis was performed using pandas.

1.3 Constraints

As full-time students with other course commitments, we had limited time to complete this project so we were restricted in the range of models that could be explored. In addition, we have limited domain knowledge and relied solely on the validity of the data that was used. We also have limited knowledge of the range of machine learning techniques, so we chose to explore those that we were able to understand and interpret.

2.0 Data

2.1 Data 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.

2.1.1 Data Description

The data includes 961 records, with 6 columns as described below:

Data Field	Data Type	Data Description
BI-RADS assessment	ordinal	Scale of 1 to 5 determined by medical examiner
Age	integer	Patient’s age in years

Shape	nominal	round=1 oval=2
		lobular=3 irregular=4

Margin	nominal	circumscribed=1 microlobulated=2
		obscured=3 ill-defined=4
		spiculated=5

Density	ordinal	high=1 iso=2
		low=3 fat-containing=4

Severity	Boolean	Target variable benign=0 malignant=1

Table 1: Data Attributes

2.1.2 Missing Data

The data contained records with missing attribute information, specifically 2 missing BI-RADS assessment, 76 missing density, 48 for margin, 31 missing shape, and 5 with missing age. We decided to replace any missing Age values with the median age, then removed any remaining records that still had a missing value. Thus, 836 records remained in our data for further analysis.

2.1.3 Attribute Type

We converted the data fields to align with the attributed descriptions provided. This included using one-hot-encoding to convert the nominal attributes into dummy variables and dropping one level of each of these nominal variables to reduce collinearity.

2.2 Exploratory Data Analysis

Using visualization techniques, we were able to examine the distribution of the different attributes contained 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. We observed that the density attribute was concentrated in the level 3 category, with almost 90% of the records falling under this level. The variable distributions are shown in Appendix 1.

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 Appendix 2.

2.3 Feature Selection

After some investigation, we learned that the BI-RADS assessment variable is assigned by the attending physician based on the mammogram mass results. Thus, this attribute is not meant to be a predictive variable, and we removed this attribute from the data for analysis. We retained all of the other features i.e. Age, Density, Margin and Shape, to be 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.

3.0 Methodology

3.1 Training and Test Data

The first step involved splitting the data into two parts :

1. 80% of data (669 observations) assigned to the training set to be used for building the models, and
2. 20% of data (168 observations) assigned to the test set will be 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. This is known as overfitting.

In addition, 10-fold cross-validation was used when building the models. Under this approach, the training data is further subdivided into train and validation data by resampling. The training data is randomly split into 10 groups of equal size, with each one having a turn as the validation data, while the other 9 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 9 times. This approach was used given the small size of the training data (669 records).

3.2 Models Used

This is a classification problem, and while there are many machine learning methodologies that could be used, we focused on three model types : Logistic regression, Decision trees, and K-nearest neighbours.

3.2.1 Logistic Regression

The logistic regression model uses a log odds transformation to determine the probability that an event would occur. In the case of mammogram mass data, the model determines the coefficients to be applied to the independent features (density, shape, margin, and age) that maximize the log-likelihood function. These coefficients are then applied to each record to calculate a probability (between 0 and 1). A threshold can then be set to determine which class (0 or 1) that observations will be assigned to base on the calculated probability. A threshold closer to 1 may result in fewer malignant cases being identified, while a threshold closer to 0 may result in too many false positives. For this model, hyperparameters include the solver to be used, the type of regularization penalty to be applied, and a weight C, that is applied to the penalty. Logistic regression is easy to implement and interpret, but because it creates a linear boundary between categories, it assumes linearity between the target variable and the features.

3.2.2 Decision Trees

A decision tree model is used to categorize data based on a series of tests or decisions that are made. The nodes of the tree represent the attributes, while the branches represent the decision paths that are taken. The leaves of the tree represent the outcome or classification for that observation. For this model, hyperparameters include maximum depth, the minimum number of samples required to split a node, and the maximin leaf nodes. Decision trees are useful in machine learning as they are easy to interpret and explain, but they often suffer from overfitting. This can be reduced by pruning the tree (limited the depth of the tree) or adjusting the number of samples that can be in each leaf.

3.2.3 K-Nearest Neighbours

The K-nearest neighbours approach is a classification approach that groups individual observations into a category based on its proximity to other similar data records. Thus each data point is assigned a category based on the category assigned to the majority of its nearest neighbours. For this model, two hyperparameters are required : k (the number of neighbours to include) and a distance measure (e.g. Euclidean distance, uniform etc.) . The k-nearest neighbours approach 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.

3.3 Tuning Hyperparameters

As discussed above we need to determine the optimal parameters that would produce the best fit. To achieve this, we used a grid search with cross-validation, which takes as its parameters :

a. the model being used,

b. a grid which includes a range or list of parameters to be tested,

c. a scoring measure (we used accuracy as the measure for all models), and

d. the number of folds to be used for cross-validation (we set cv=10 for all models)

Model	Range of Parameters Tested	Final Parameters Used

Logistic Regression	Penalty = ridge or lasso
	C = -2 to 4 so that 20 values were tested
	Solver = (liblinear, newton-cg, lbfgs)
		Penalty = ridge
		C = 0.02
		Solver = newton-cg

Decision Tree	Maximum depth = 5, 7, 9 Minimum sample for split = 2, 3, 4 Maximum leaf nodes = 2 to 12 in increments of 1 Criterion = gini or entropy	Maximum depth = 7 Min sample for split = 2 Max leaf nodes = 11 Criterion = entropy

K-nearest neighbours	k = 1 to 20 in increments of 1
	Weight = uniform or distance
		k = 5
		Weight = uniform

Table 2: Hyperparameters

4.0 Results / Experimental Analysis

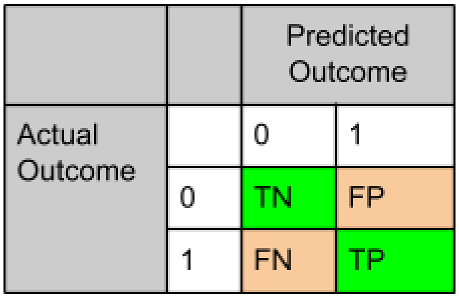
As mentioned in section 3.1, a test dataset was held aside for the purpose of evaluation of the models. The final version of each model was used to predict the outcome for each observation in the test data. Receiver Operating Characteristic (ROC) curves were plotted for each set of model predictions and the Area Under the Curve (AUC) was calculated and used to compare model performance. In addition, predicted outcomes were compared to the actual outcomes and a confusion matrix was created which allowed the determination of several metrics, including precision, accuracy, false alarm rate and false negative rate.

4.1 ROC Curves and AUC

The ROC curve is used to plot the true positive versus the false positive rates to illustrate the sensitivity of the model. An ROC curve that is closer to the top left corner indicates that the model is doing better at classifying data into the appropriate categories. The AUC was calculated for each model, and it was determined that with an AUC of 0.84, the logistic regression model outperformed both the decision tree (AUC = 0.80) and the K- nearest neighbours (AUC = 0.79). The plotted ROC curves and the calculated AUC for the 3 models used are shown in Appendix 3.

4.2. Confusion Matrix

A confusion matrix summarizes the predicted outcomes for classification problems. The columns represent the predicted outcomes, while the rows represent the actual outcomes. The cells indicate the number of outcomes that are included in the intersection of the respective row and column. The values in the confusion matrix can then be used to calculate several useful metrics.



		Predicted Outcome	
Actual Outcome		0	1
	0	TN	FP
	1	FN	TP

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

False Alarm Rate = FP/(TP+FP)

False Negative Rate = FN/(FN+TP)

While accuracy and precision are useful in assessing model performance, both the false alarm rate and the false negative rate are important in this case. The false alarm rate indicates the percentage of cases predicted to be malignant but were actually benign. These patients would have been subjected to unnecessary procedures, and anxiety due to an incorrect prediction. The false negative rate is also of concern, as this would indicate the percentage of cases that were predicted to be benign, but actually turned out to be malignant. In these cases, patients with cancer would be incorrectly told the masses were benign and may miss out on medical interventions and this could lead to poorer prognosis when eventually detected.

4.2.1 Confusion Matrix and Metrics for Models Used

The following is a summary of the results for each of the three models used in our analysis.

Model	Logistic Regression	Decision Tree	K-Nearest Neighbours
Accuracy =	78%	74%	79%
Precision =	72%	68%	72%
FAR =	29%	35%	31%
FNR =	14%	16%	10%

Table 3 : Metrics from Confusion Matrix

As can be seen in the above table, the k-nearest neighbours performed best in terms of accuracy and had the lowest false negative rate. The performance measures are moderate however, as a 10% false negative rate and 31% false alarm rate still seem to be quite high. The confusion matrix for each model is shown in Appendix 4.

4.3 Comparison to BI-RADS Assessment

In addition to comparing the models to one another, we also looked at how machine learning performed compared to the BI-RADS assessment completed by a physician. Using the original data (i.e. after removal of rows with missing values) we calculated similar metrics to those determined by the confusion matrix. The metrics for the BI-RADS assessment are:

Accuracy	49%
Precision	47%
FAR	92%
FNR	4%

Table 4: BI-RADS Metrics

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, meaning that fewer patients were mis-classified as having breast cancer, and therefore fewer were subjected to unnecessary biopsies. 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.

5.0 Conclusion

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 the more predictive power than either the decision tree or the K-nearest neighbours models.

With more time, there are several approaches that can be used to get more predictive value:

1. Obtain more data so that 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.
4. Explore other classification model types:
    - Random Forests combine multiple decision trees, each of which uses a random subset of features, and then aggregates the results of the trees to produce the outcome. Random forest models help reduce overfitting but can be time-consuming to run.
    - 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.
    - Neural networks use layers (input layer, output (or target) layer and a middle-hidden layer) that are connected by nodes to form a “network”. These can be useful particularly if the mammogram images are available for analysis. Due to the hidden layer, neural networks are often difficult to explain.

6.0 References

7 types of classification algorithms in machine learning. ProjectPro. (2022, July 8). Retrieved July 16, 2022, from https://www.projectpro.io/article/7-types-of-classification-algorithms-in-machine-learning/435

Breast cancer statistics. Canadian Cancer Society. (2022, May). Retrieved July 14, 2022, from https://cancer.ca/en/cancer-information/cancer-types/breast/statistics

Elter, M., & Schulz-Wendtland, D. R. (n.d.). *Mammographic Mass Data Set*. UCI Machine Learning Repository: Mammographic mass data set. Retrieved July 10, 2022, from https://archive.ics.uci.edu/ml/datasets/mammographic+mass

Lee, S. (n.d.). *Mammography*. Canadian Cancer Society. Retrieved July 16, 2022, from https://cancer.ca/en/treatments/tests-and-procedures/mammography

Appendix 1 - Distribution of Features in Data

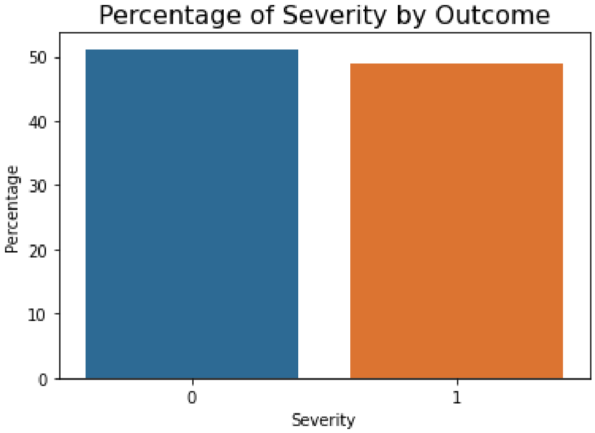


Figure 1: Distribution of Target Variable

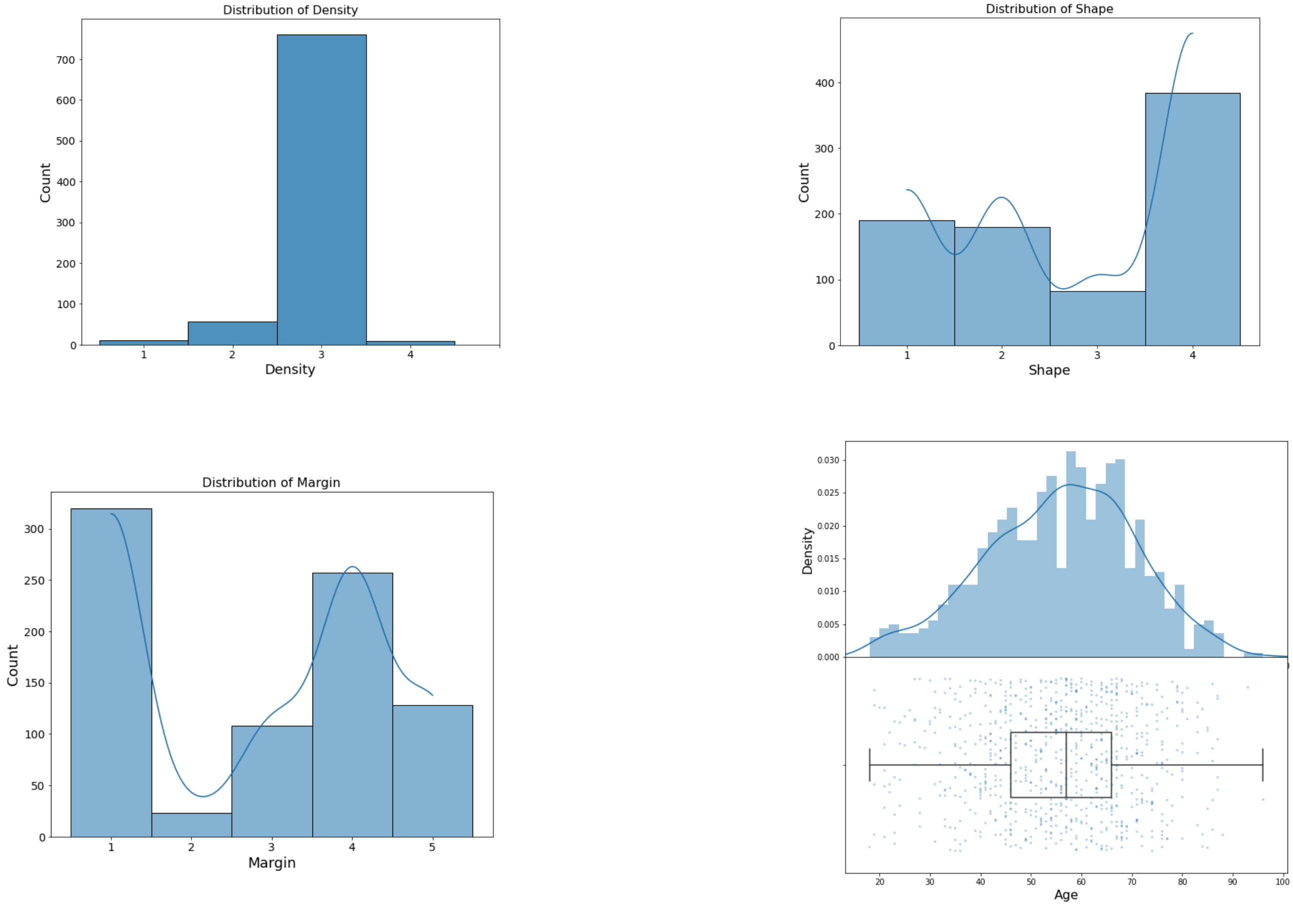


Figure 2: Distribution of Predictor Variables

Appendix 2 - Relationship of Features

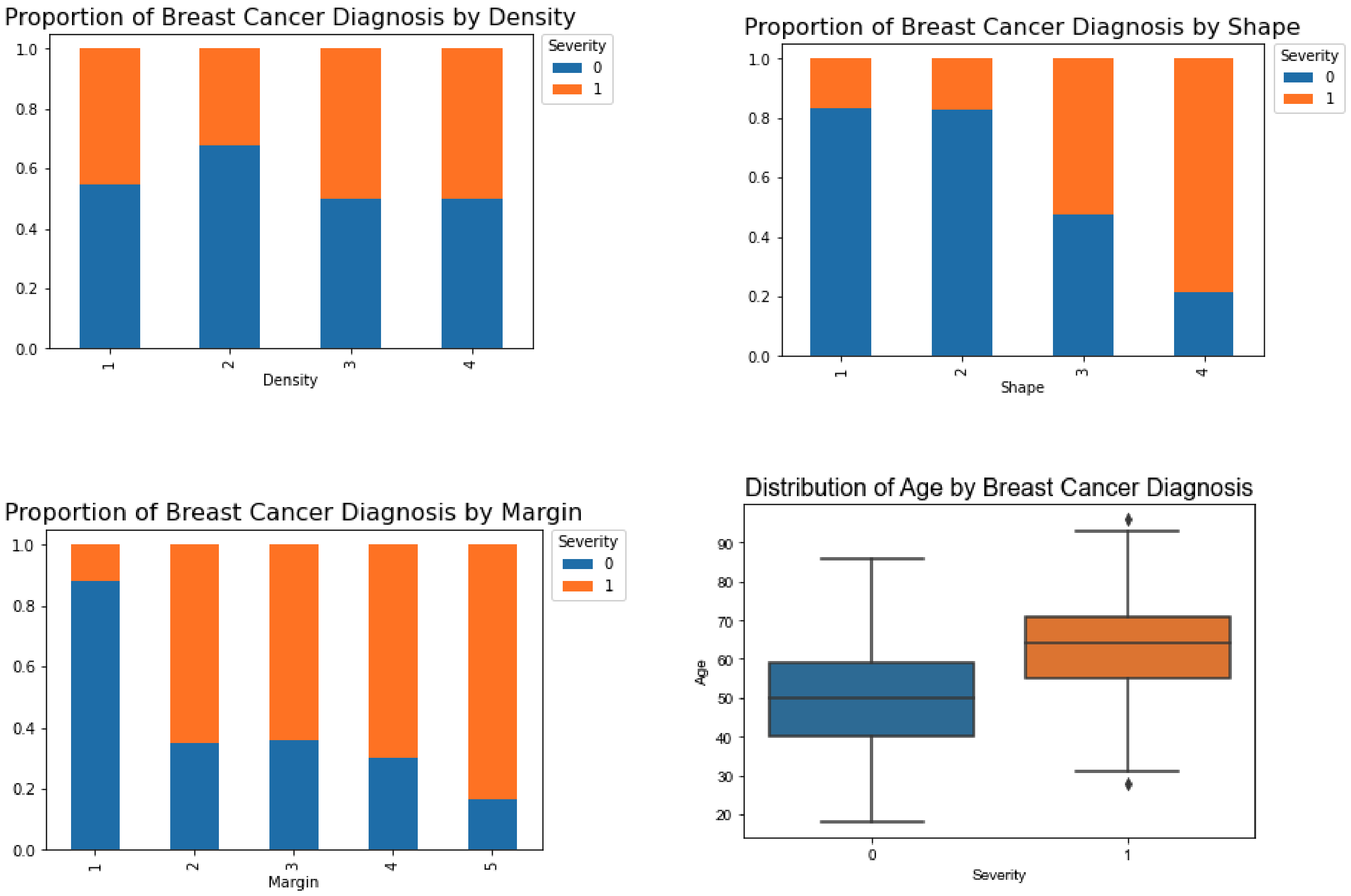


Figure 3: Relationship of Predictors to Target

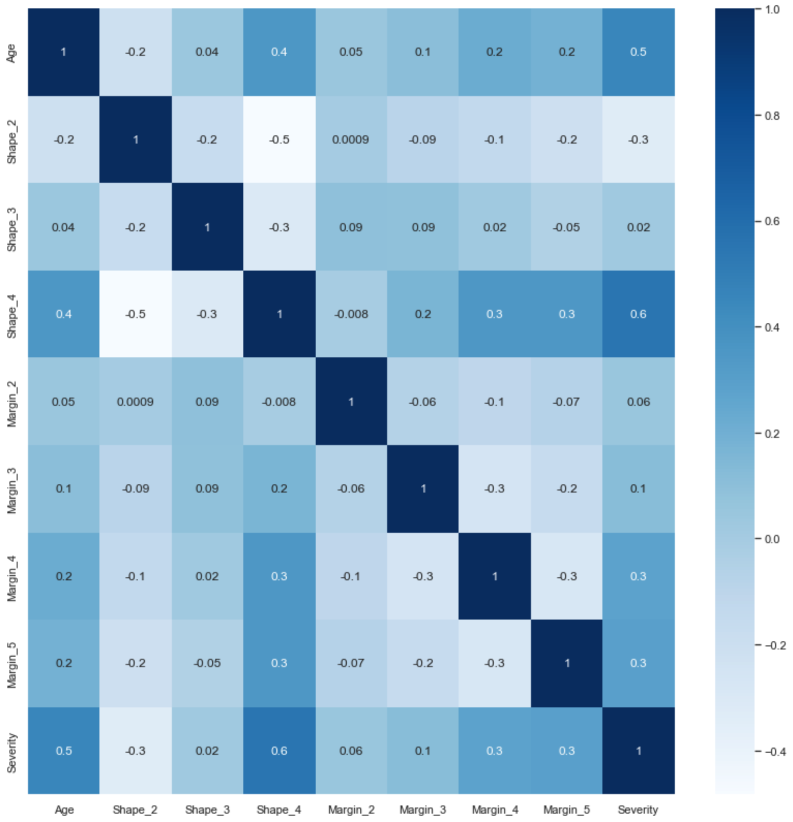


Figure 4: Correlation Matrix showing pairwise relationships

Appendix 3 - ROC Curves and AUC Values

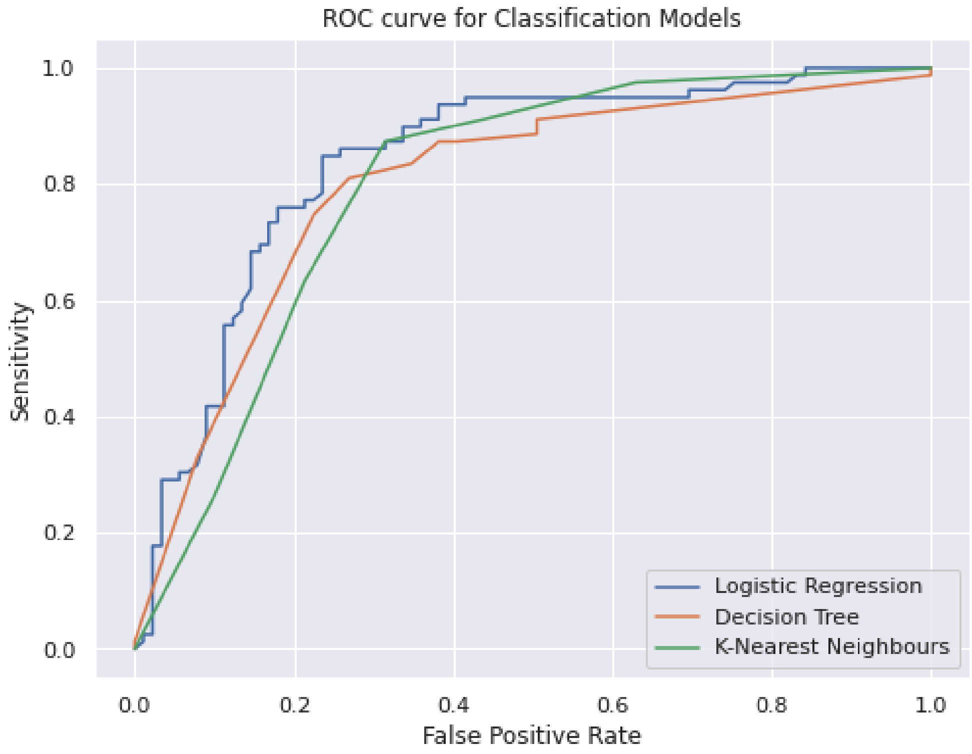


Figure 5: ROC Curve for Models

Logistic Regression	0.84
Decision Tree	0.80
K-Nearest Neighbours	0.79

Table 5: AUC Values

Appendix 4 - Confusion Matrix Results

Logistic Regression Model

		Predicted	
		0	1
Actual	0	63	26
	1	11	68

Decision Tree Model

		Predicted	
		0	1
Actual	0	58	31
	1	13	66

K-Nearest Neighbours Model

		Predicted	
		0	1
Actual	0	61	28
	1	8	71

Figure 6: Confusion Matrices for Models