Breast Cancer Classification

Kayleigh Habib and Winnie Szeto
200370580 200553800

Table of Contents

Overview

1 Purpose
Data Description

Model Planning

2 Data Preparation and Analysis

Model Building

3 Model Building and Comparisons

Conclusion

4 Findings
Future Enhancements

1. Overview

Introduction

Breast Cancer

- About 28,600 women and 270 men will be diagnosed in Canada in 2022
- Deaths account for 14% of all cancer deaths in Canada
- Can be successfully treated with the help of early detection

About Mammograms:

- Performed to check for presence of unusual masses in the breast
- X-ray machine to produce images of your breast tissue by flattening the breasts to look for breast cancer cells

Goal of the Project:

- How effectively can supervised models be used to determine whether a mass indicated in a mammogram is benign or malignant?

Data Description

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

Source Of Data:
Elter, M., & Schulz-Wendtland, D. R. (n.d.). Mammographic Mass Data Set. UCI Machine Learning Repository: Mammographic mass data set. Retrieved July 17, 2022, from https://archive.ics.uci.edu/ml/datasets/mammographic+mass

	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

```
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Age       836 non-null    int32 
 1   Shape     836 non-null    object
 2   Margin    836 non-null    object
 3   Density   836 non-null    object
 4   Severity  836 non-null    int64 
```

Data Cleaning

Data contained missing values

- 5 records with missing age were replaced with median age
- Remaining rows with missing values were removed from the data

Fix data types to match attribute description

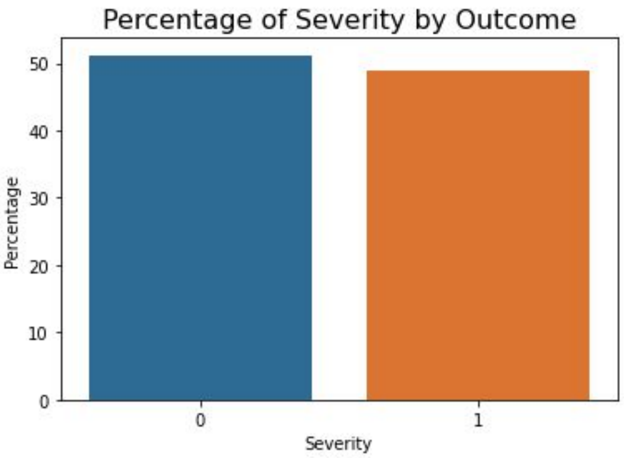
- Used one-hot encoding to create dummy variables for nominal variables, and drop one level to reduce redundancy

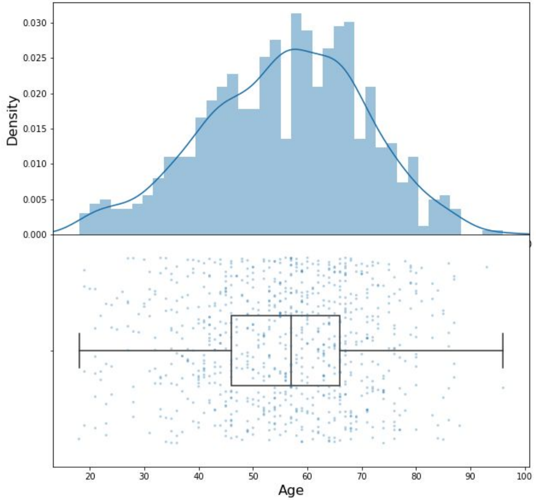
Remove non-predictive feature

- Comparing performance of models to BI-RADS assessment, so this variable should not be used as a predictor

2. Model Planning

Exploratory Data Analysis



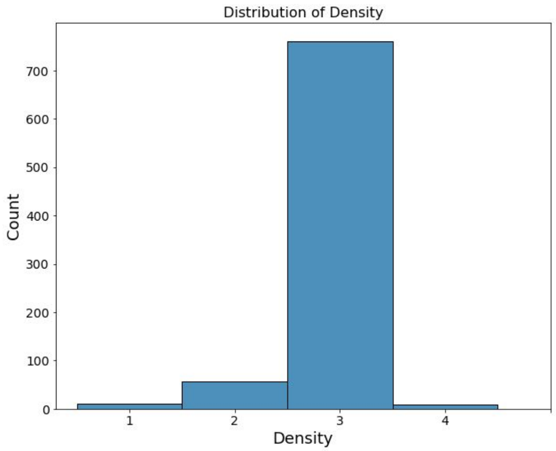


Data is balanced

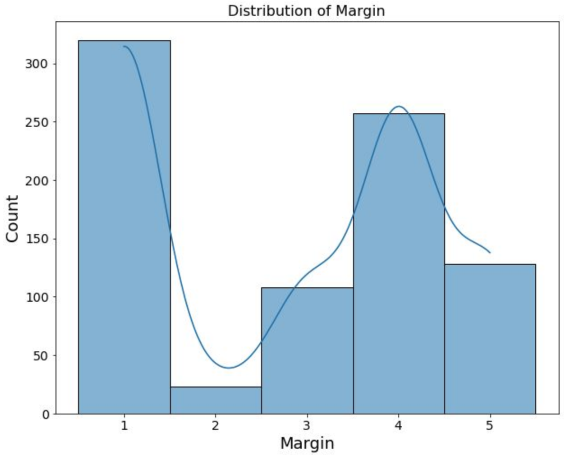
- 51% benign cases
- 49% malignant cases

50% of Ages are between 45 and-65

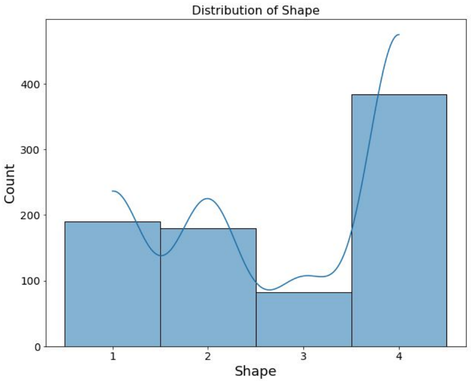
Exploratory Data Analysis



Majority of masses classified as Density 3 (low)



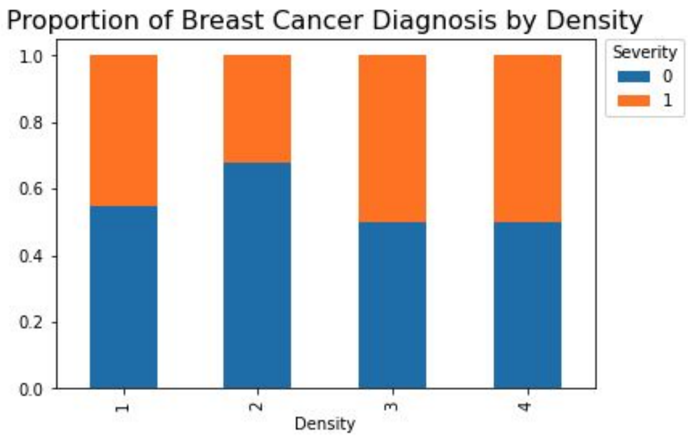
Masses are either the 1 (circumscribed) or 4 (ill-defined) category



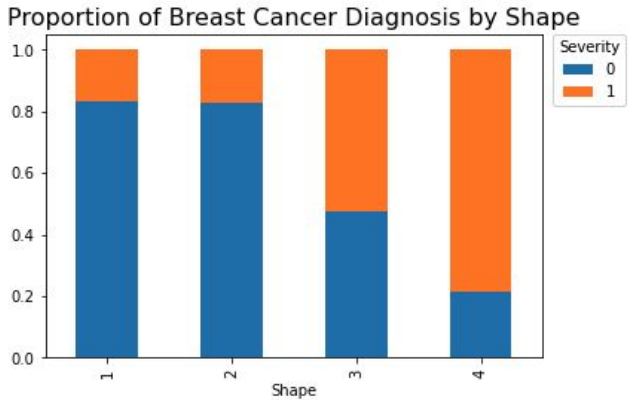
Category 3 (lobular) contains least number of records

Relationship between Predictors and Response

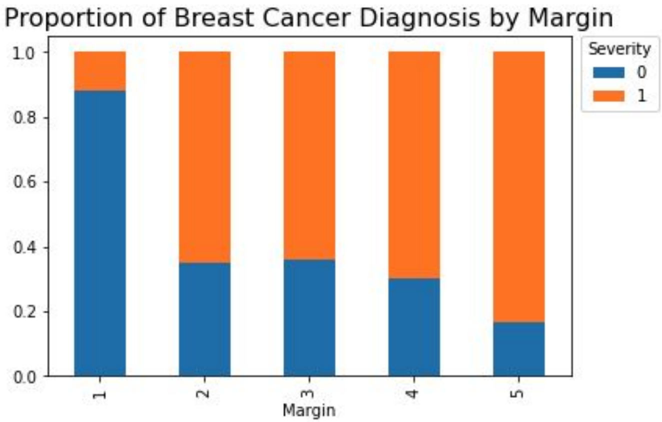
Density category 2 (iso) has lower incidence than other categories



Shape categories 1 and 2 have lowest incidence
Incidence rate is about 50% for category 3
Increases almost 80% for category 4

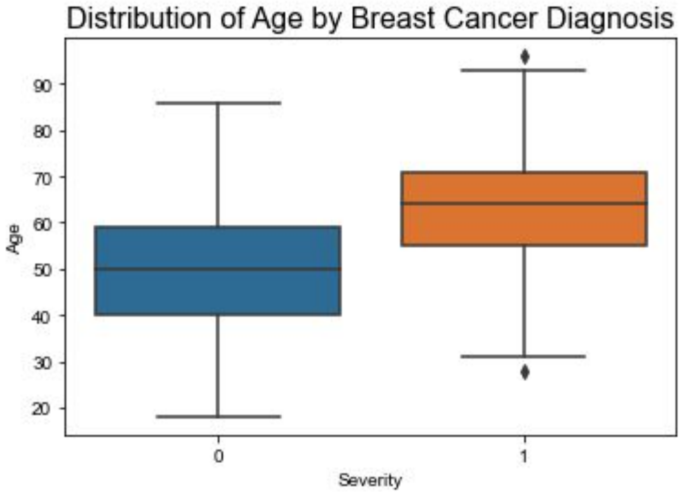


Relationship between Predictors and Response

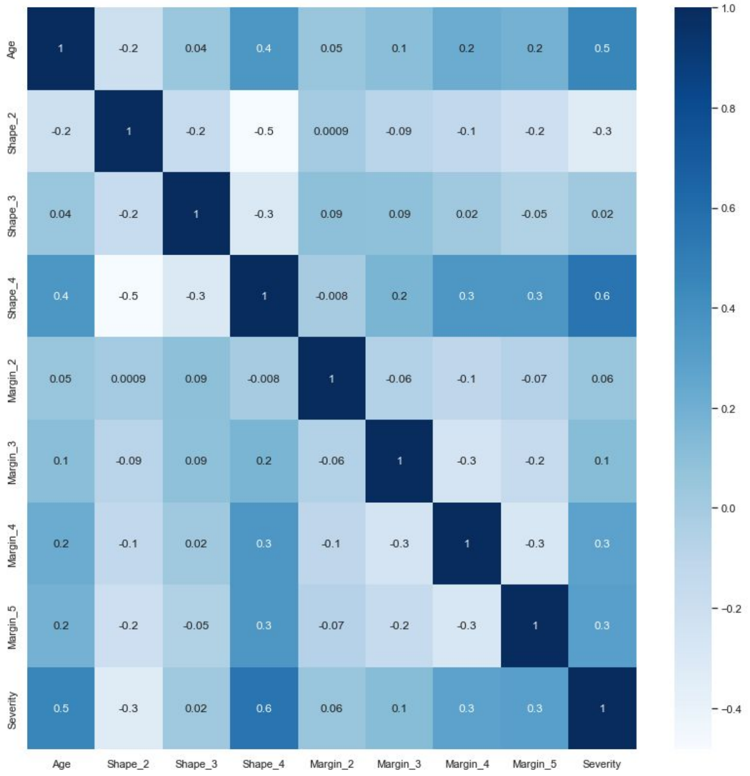


Margin category 1 has low incidence.
Incidence rate increases with each category
Highest for category 5

Incidence of breast cancer appears to be more likely at higher ages



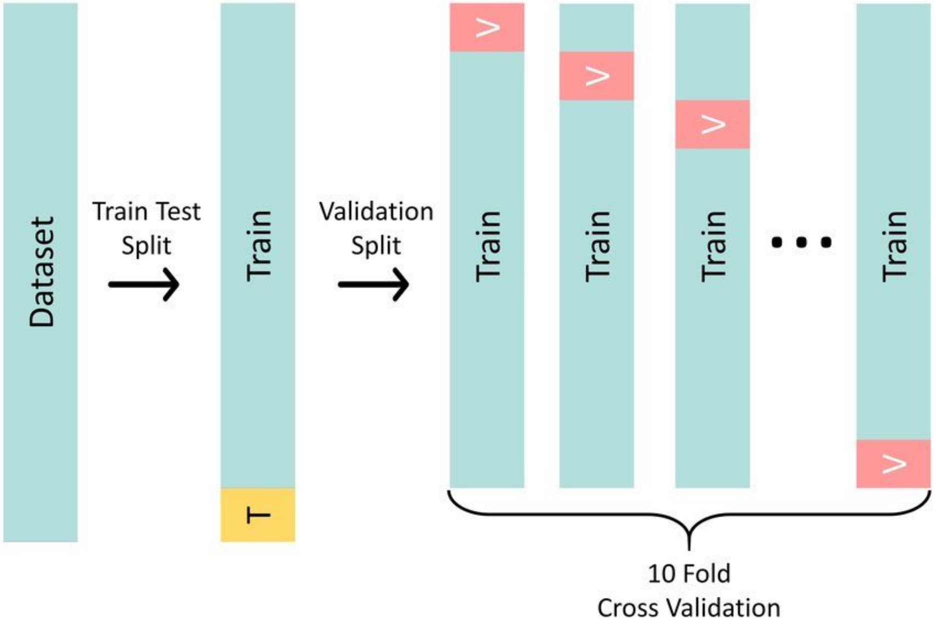
Correlation Matrix



- Identify correlations between features
- Stronger relationship: represented by darker blue shade
- Variables highly correlated, may add complexity without model improvement
- Matrix does not appear to be any strong correlations between features

3. Model Building

Splitting the data



Need to split data into two parts:

1. 80% assigned to training
2. 20% assigned to testing

10- fold Cross - validation

- Training data is subdivided
  - Train : split into 10 equal groups, with 9 used to fit the model at each iteration
  - Validation: used to assess fit of the model

Model 1 : Logistic Regression

What is Logistic Regression?

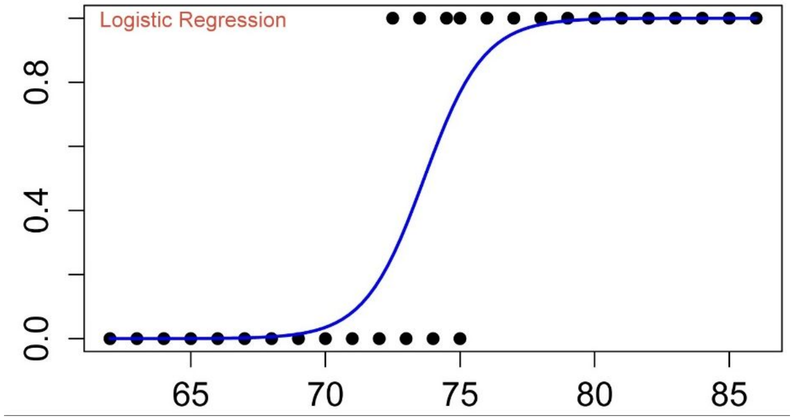
- Uses log odds transformation to determine probability event would occur
- Threshold set to determine which class will be assigned based on probability

Why choose Logistic Regression?

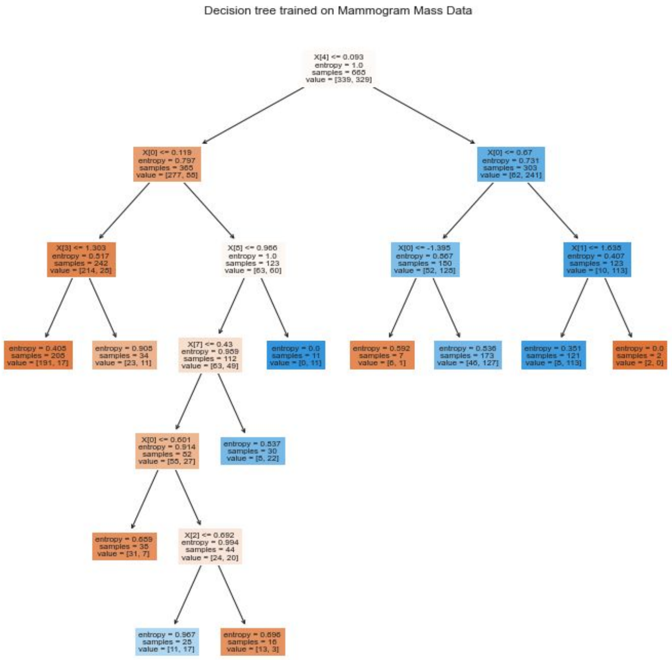
- Easy to implement and interpret
- Linear boundary between categories

Our Logistic Regression Model

- Used ridge penalty so less important predictors have coefficient close to zero
- AUC = 0.84



Model 2 : Decision Trees



What is a Decision Trees?

- Categorize data based on series of decisions that are made at each node
- Nodes: attributes, Banches: decision paths, Leaves: classification

Why choose Decision Trees?

- Easy to interpret and explain
- Suffer from overfitting but can be pruning

Our Decision Tree

- Used entropy as criterion for splitting
- Set max depth = 7, and max leaf nodes = 11
- AUC = 0.80

Model 3: K- Nearest Neighbours

What is K-Nearest Neighbours?

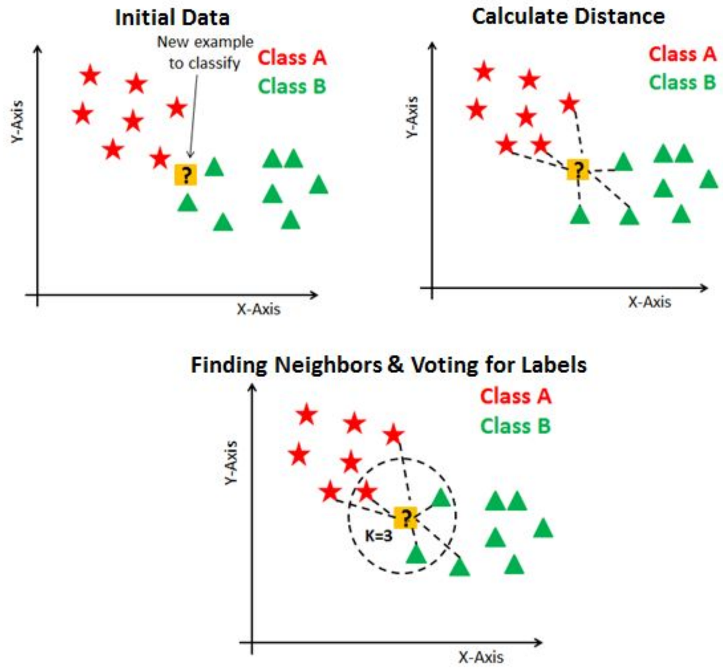
- Groups individual observations into categories based on proximity to other observations

Why we chose K-Nearest Neighbours?

- Relatively simple
- Data is not high-dimensional so not likely to overfit

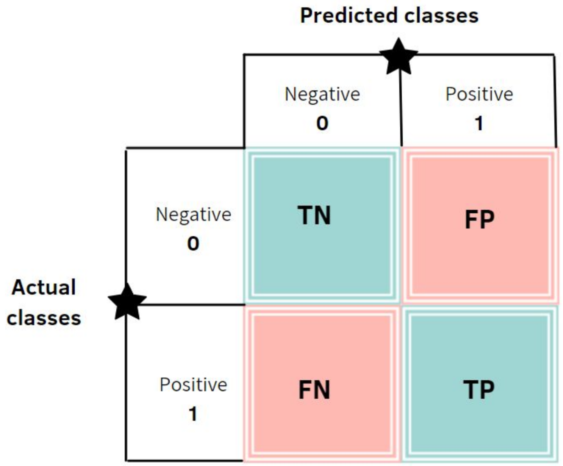
Our KNN model

- Used 5 nearest neighbours and uniform weighting
- AUC = 0.79



Comparing Using Confusion Matrix

- Confusion matrix used to display counts by predicted versus actual outcomes for classification problems
- Can then be used to calculate metrics for comparison



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

Precision = TP/(TP+FP)

False Alarm Rate = FP/(TP+FP)

False Negative Rate = FN/(FN+TP)

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)

Metrics using BI-RADS assessment in mammogram mass data:

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

Note the higher proportion of cases that are incorrectly classified as malignant

Metrics Using Models

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

		Predicted	
		0	1
Actual	0	63	26
	1	11	68

		Predicted	
		0	1
Actual	0	58	31
	1	13	66

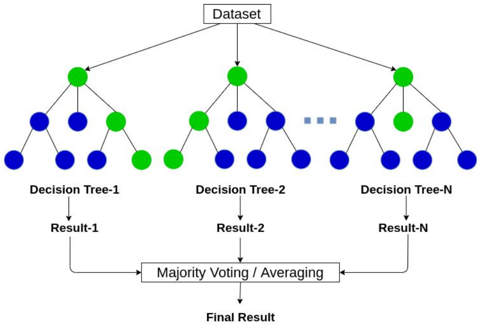
		Predicted	
		0	1
Actual	0	61	28
	1	8	71

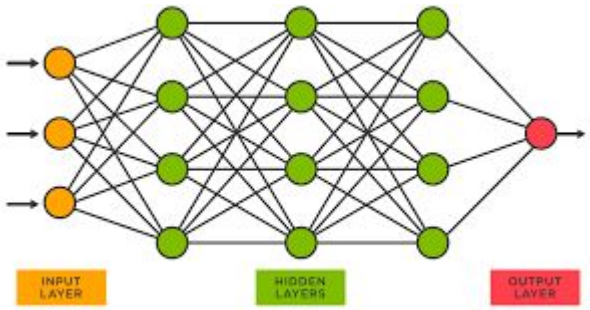
4. Conclusion

Findings

We Found that:

- Machine learning methods outperform BI-RADS assessment in successfully classifying mammogram masses as benign or malignant in accuracy and precision
- Logistic model performed best with an AUC of 0.84





Future Enhancements

- Obtain more data (observations, features)
- Use ensemble of Models
- Explore other classifications
    - Random Forests
    - Support Vector Machines
    - Neural Network

Thank you!

Contribution

Name	Contribution
Kayleigh Habib - 200370580	Coding, Write-up, Presentation
Winnie Szeto - 200553800	Write-up, Presentation, Review