Name: Amy Femal
Date: August 6, 2020
Title: Breast Cancer Analysis

Section 1

Introduction

Breast cancer is the most common type of cancer among women – 1 in 8 women in the United States will be diagnosed at some point in their life. Early detection is important as it provides more treatment options and a better chance for survival. Mammographic images can help doctors look for early signs of cancer. Suspicious mammographic may require a biopsy to determine a diagnosis. Since only approximately 20% of biopsies results in a malignant diagnosis and can be very painful, I would like to use the measurements taken from mammographic images to predict the diagnosis. The results could better determine whether the woman requires a biopsy or further monitoring through follow-up examinations. The only definite way to diagnose breast cancer is through a biopsy; however, fine needle aspiration biopsies also have problems with prediction accuracy, with a 4% false positive rate and 12% false negative rate (Smith, 1997). Therefore, I would also like to use the measurements taken from biopsies to predict and improve the diagnosis accuracy and recall. The accuracy of diagnosis allows for earlier detection, which mean the difference between life and death. Finally, I would like to determine which variables are significant to the survival of a patient diagnosed with breast cancer.

Research questions

Using different datasets, I would like to answer the following questions:

- What measurements taken during a mammographic image are significant to the diagnosis of breast cancer?
- Using these significant measurements, can I use a machine learning model to predict a diagnosis with high accuracy and high recall?
- What measurements of cell nuclei taken from a fine needle aspiration biopsy are significant to the diagnosis of breast cancer?
- Using these significant measurements, can I use a machine learning model to predict the diagnosis with high accuracy and high recall? Less than 4% false positive and 12% false negative?
- What characteristics contribute to a patient surviving longer than five years post-surgery?

Approach

For each dataset, I will perform exploratory data analysis. I will examine the variables to find which are the most significant in predicting breast cancer and the survival of a patient. I will use different classification models to identify which provides the best accuracy and recall in predicting a diagnosis.

How My Approach Addresses the Problem

Type II errors, or false negatives, allow breast cancer to remain undetected and treatment to be delayed. Delays in treatment increase the risk of metastasis, or the spread of cancer to other parts of the body, which can reduce life expectancy. I will use the survival data to determine which variables contribute most to the survival of a patient. I will use the data from mammographic images and fine

needle aspiration biopsies to determine how to predict breast cancer with high accuracy and perfect recall. This will ensure breast cancer does not go undetected and increase the chance of survival.

Data

All datasets were obtained from Kaggle.com and include the following information:

- Haberman Cancer Survival Data
  https://www.kaggle.com/krpiku/haberman.csv?select=haberman.csv

  This dataset was collected by the University of Chicago's Billings Hospital between 1958 and 1970 from patients who had undergone breast cancer surgery. The dataset contains 305 observations with four variables:
  - age: age of the patient at the time of surgery
  - year: year the surgery was performed
  - nodes: number of positive axillary lymph nodes
  - status: status of the patient 5 years post-surgery, where 1 = survived and 2 = died

- Mammographic Mass Data
  https://www.kaggle.com/overratedgman/mammographic-mass-data-set

  This dataset was collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006 from mammographic images. The original dataset was comprised of 961 observations but attributes with missing values were removed. This dataset includes 830 observations and six variables:
  - BI-RADS: standard system used to describe findings where:
    - 0 = Incomplete - Additional imagining is necessary
    - 1 = Negative - No abnormalities to report
    - 2 = Benign - No cancerous findings
    - 3 = Probably benign finding
    - 4 = Suspicious abnormality
    - 5 = Highly suggestive of malignancy
    - 6 = Known biopsy-proven malignancy
  - age: age of patient
  - shape: shape of the mass, where 1 = round, 2 = oval, 3 = lobular, 4 = irregular
  - margin: description of the mass's edge, where 1 = circumscribed, 2 = microlobulate, 3 = obscured, 4 = ill-defined, 5 = spiculated
  - density: density of the mass, where 1 = high, 2 = isodense, 3 = low, 4 = fat-contained
  - severity: diagnosis of the mass, where 0 = benign, 1 = malignant

- Breast Cancer Prediction Dataset
  https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset

  This data was collected by Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian in the General Surgery and Computer Science Departments as the University of Wisconsin in 1992

from fine needle aspiration biopsies. The original dataset is comprised of 569 observations with 32 variables. This dataset contains all 569 observations but has been reduced to the following six variables:

- radius: mean of distances from the center of the cell nuclei to the points on its perimeter
- texture: mean texture of the nuclei, described by the spatial arrangement and variation of grey values observed
- perimeter: mean distance around the nuclei
- area: mean area of the nuclei
- smoothness: mean of the local variation in radius lengths
- diagnosis: diagnosis of the mass, where 0 = malignant, 1 = benign

.

Required Packages

I will be using the following packages to perform the analysis:

- ggplot2
- gridExtra
- MASS
- tidyverse
- caret
- rpart
- rattle
- rpart.plot
- RColorBrewer
- Class
- randomForest
- caTools
- mlbench
- e1071

Plots and Table Needs

I will be using histograms to determine distributions, scatter plots and correlation matrices to identify relationships, density plots and boxplots to compare different groups, and confusion matrices to summarize prediction results.

Questions for Further Steps

I will conduct further research of classification models to determine which are the most appropriate to use for the datasets. During this course, I have learned about logistic regression and k-nearest neighbors. I will be exploring other classification models, specifically decision tree and random forest models.

Section 2

How to import and clean my data

First, I will import each dataset.

```
mam <- read.csv('mammography.csv')
biop <- read.csv('biopsy.csv')
hab <- read.csv('haberman.csv')
```

Then I will look at the structure and data summary for each. In doing so, I have verified there is no missing data. The summary of the *mam* dataset shows an observation that has a value of 55 for BI.RADS. As BI.RADS is an ordinal, categorical variable, I have removed this observations from the dataset. No other outliers were evident in the datasets. I have converted categorical variables, indicating ordinal variables where appropriate.

What does the final data set look like?

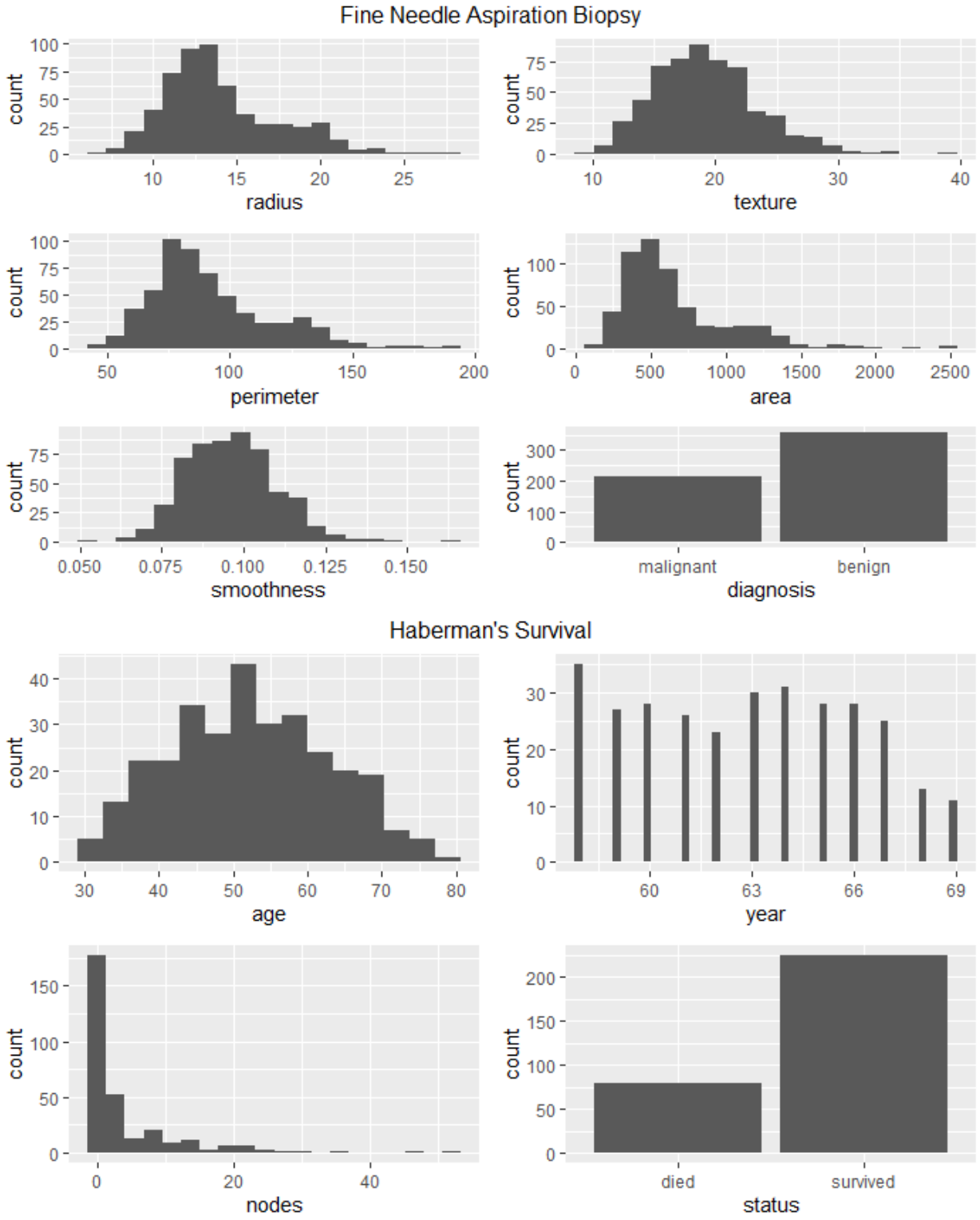
The clean datasets have the following summaries:

```
summary(mam)
BI.RADS      Age         Shape   Margin  Density Severity
0:  5   Min.   :18.00   1:190   1:320   1: 11   1:402
2:  7   1st Qu.:46.00   2:180   2: 23   2: 56   0:427
3: 24   Median :57.00   3: 81   3:105   3:754
4:468   Mean   :55.79   4:378   4:254   4:  8
5:316   3rd Qu.:66.00           5:127
6:  9   Max.   :96.00
```

```
summary(biop)
texture           area          smoothness      diagnosis
Min.   : 9.71   Min.   : 143.5   Min.   :0.05263   0:212
1st Qu.:16.17   1st Qu.: 420.3   1st Qu.:0.08637   1:357
Median :18.84   Median : 551.1   Median :0.09587
Mean   :19.29   Mean   : 654.9   Mean   :0.09636
3rd Qu.:21.80   3rd Qu.: 782.7   3rd Qu.:0.10530
Max.   :39.28   Max.   :2501.0   Max.   :0.16340
```

```
summary(hab)
age             year           nodes             status
Min.   :30.00   Min.   :58.00   Min.   : 0.000   died    : 80
1st Qu.:44.00   1st Qu.:60.00   1st Qu.: 0.000   survived:225
Median :52.00   Median :63.00   Median : 1.000
Mean   :52.36   Mean   :62.87   Mean   : 4.033
3rd Qu.:60.00   3rd Qu.:66.00   3rd Qu.: 4.000
Max.   :78.00   Max.   :69.00   Max.   :52.000
```

The variables have the following distributions:



The variables in the *biop* and *hab* datasets have skewed-right distributions.

Questions for future steps

Can I transform non-normal variables to make them normal? Can I remove any variables from the datasets?

What information is not self-evident?
The *biop* dataset has variables radius, perimeter, and area. Assuming a nucleus is perfectly round, these measurements are related: ½ * radius * perimeter (circumference) = area. In this case, I could remove radius and perimeter from the dataset. However, the nuclei in the dataset may not be round. I need to determine how these measurements are related to decide the best approach.

What are different ways you could look at this data?
I will examine how the variables compare when the data is split between the binary variables. This will provide a better understanding about what variables are significant to the dependent variables.

How do you plan to slice and dice the data?
I will split the data between the binary variables into two sample groups per dataset. This will help determine if there is a difference between the groups.

How could you summarize your data to answer key questions?
I will the non-parametric Mann-Whitney U Test to determine if there are differences between the groups. This will help determine significant variables to use in the classification models.

What types of plots and tables will help you to illustrate the findings to your questions?
I will use boxplots and density plots to compare the two groups from each dataset and correlation matrices to identify relationships between variables.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
Once I have identified the significant variables, I will use several classification algorithms to predict the diagnosis in the *mam* and *biop* datasets to decide which is the most accurate with the greatest recall. These models include logistic regression, decision trees, k-nearest neighbors, and random forests.

Questions for future steps.
I would like to explore the BI.RAD variable from the *mam* dataset. What variables (excluding severity) are significant? How accurate is BI.RADS in determining diagnosis? What classification model predicts the BI.RAD class the most accurately? Can I use this model to predict the BI.RAD value for the observation I removed from the dataset?

Section 3

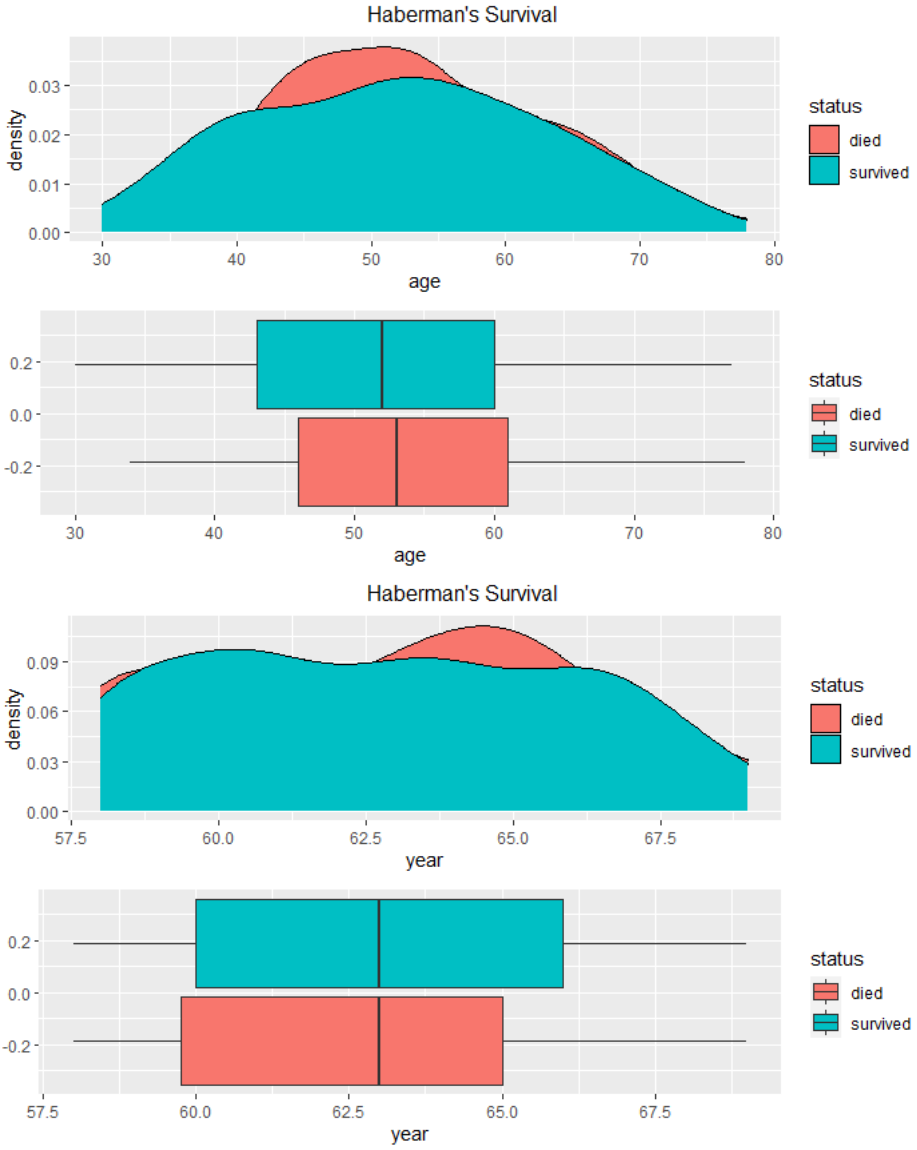
Breast cancer is the most common type of cancer among women – 1 in 8 women in the United States will be diagnosed at some point in their life. Mammographic images are used to detect abnormal masses in breast tissue. Based on the observed attributes of the mass, a physician may recommend a biopsy to determine a diagnosis. A fine needle aspiration (FNA) biopsy is a common form of the procedure, in which a physician uses a needle attached to a syringe to withdraw tissue or fluid from the mass. Mammographic images and FNA biopsies tend to have diagnosis inaccuracies, yielding false negative rate of 20% and 12%, respectively. False negative, or Type II, diagnosis results are the most

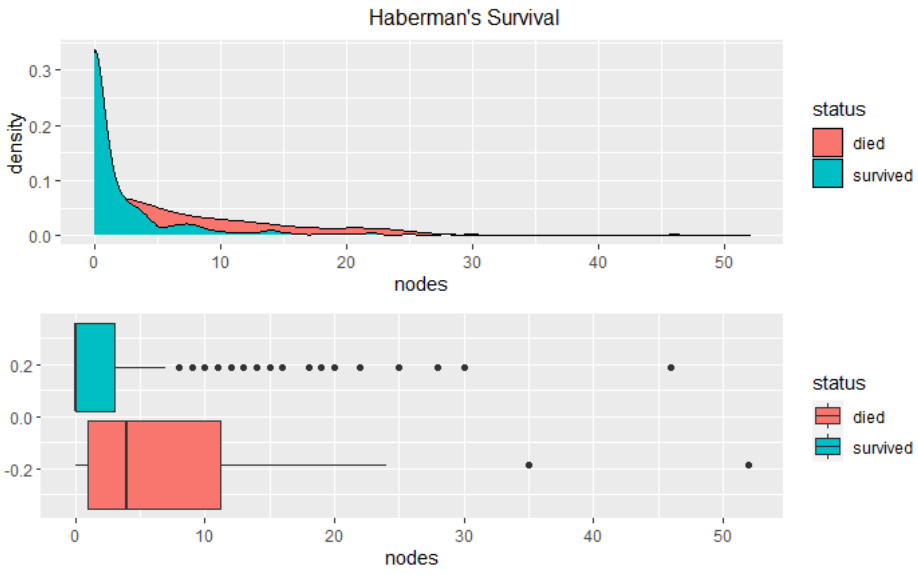
damaging as they allow breast cancer to remain undetected and treatment to be delayed. Delays in treatment increase the risk of metastasis, which can reduce life expectancy.

I will use the Haberman's Survival dataset to determine which variables are most significant to the survival of a patient. I will use the Mammographic Imaging and Biopsy datasets to determine which variables are most significant to predict breast cancer with high accuracy and perfect recall, avoiding any false negative predictions. This will ensure breast cancer does not go undetected and increases survival.

Haberman's Survival

First, I will determine which variables are significant in determining the survival status. I will compare the variables between the two groups, died and survived, and use density and boxplots for visual aids.





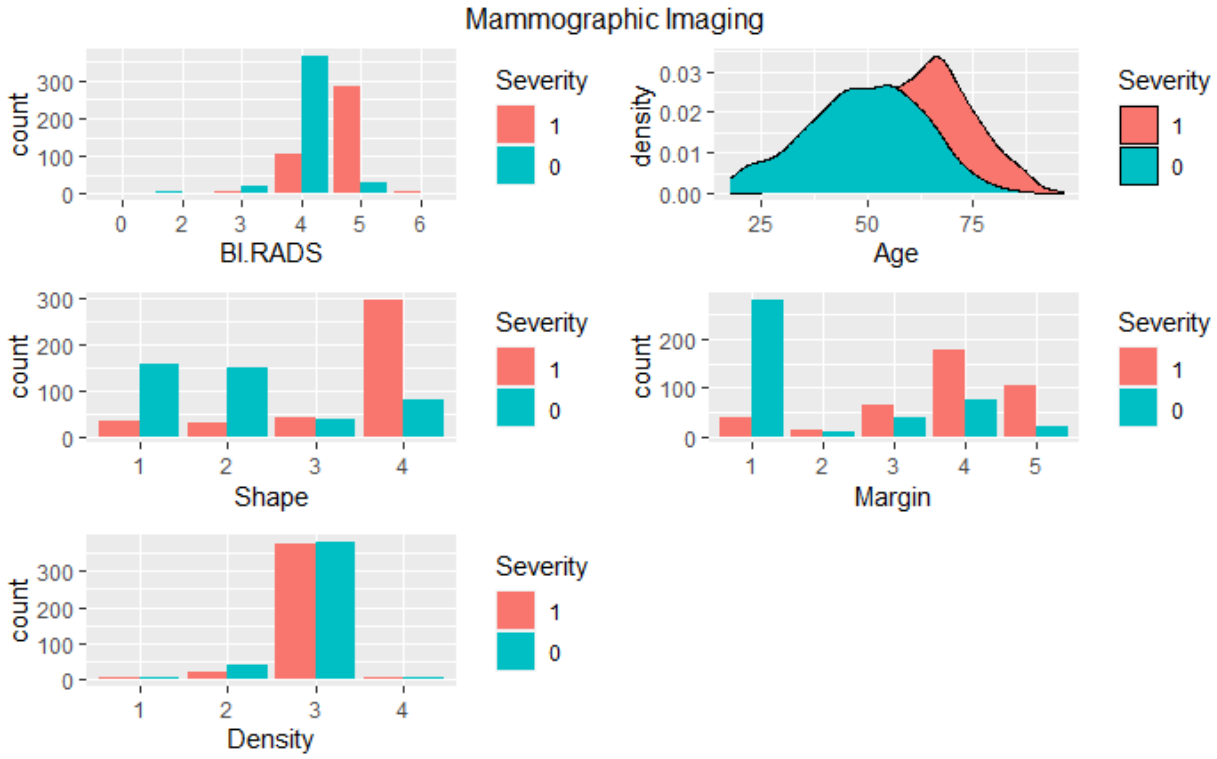
Based on the density and boxplots, it appears the only significant variable is nodes. As the variables are skewed-right, I will use the Mann-Whitney U Test to verify this claim.

```
        Wilcoxon rank sum test with continuity correction

data:  hab$age by hab$status
W = 9554, p-value = 0.4137
alternative hypothesis: true location shift is not equal to 0


        Wilcoxon rank sum test with continuity correction

data:  hab$year by hab$status
W = 9011, p-value = 0.9876
alternative hypothesis: true location shift is not equal to 0


        Wilcoxon rank sum test with continuity correction

data:  hab$nodes by hab$status
W = 12674, p-value = 1.275e-08
alternative hypothesis: true location shift is not equal to 0
```

The results show that there is a significant difference in nodes between the two groups. It also indicates that there is no difference in age and year. Therefore, I conclude nodes is the only significant factor in the dataset in determining whether a patient will survive longer than five years after breast cancer surgery.

Mammographic Imaging

I will determine which variables are significant in determining the diagnosis of breast cancer. I will compare the variables between the two groups, malignant and benign, and use density and boxplots for visual aids.



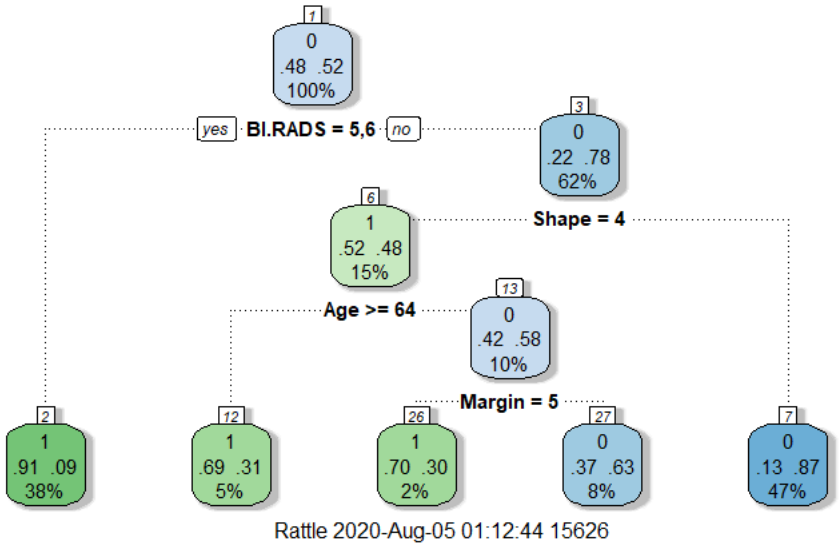
It appears that all variables, except Density, are significant variables in predicting Severity. To decide which variables to use in the logistic model, I used all variables in the dataset to determine the most significant. The full model summary indicates Age, Shape, and Margin are the most significant.

```
Call:
glm(formula = Severity ~ ., family = binomial(), data = mam)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5287  -0.4314   0.2127   0.4942   2.7931

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.614385  82.938529   0.080  0.93644
BI.RADS.L    -6.835378 178.431113  -0.038  0.96944
BI.RADS.Q    -2.661839  54.300146  -0.049  0.96090
BI.RADS.C     7.793335 259.630952   0.030  0.97605
BI.RADS^4    -6.772345 282.121556  -0.024  0.98085
BI.RADS^5     4.884801 156.734615   0.031  0.97514
Age          -0.047880   0.008875  -5.395 6.85e-08 ***
Shape2        0.204566   0.345441   0.592  0.55373
Shape3       -0.561793   0.426035  -1.319  0.18728
Shape4       -1.006812   0.378385  -2.661  0.00780 **
Margin2      -0.989340   0.645154  -1.533  0.12515
Margin3      -0.548119   0.403025  -1.360  0.17383
Margin4      -0.918193   0.341560  -2.688  0.00718 **
Margin5      -0.995241   0.432110  -2.303  0.02127 *
Density.L     1.496222   0.806466   1.855  0.06356 .
Density.Q     0.674573   0.634993   1.062  0.28809
Density.C     0.780747   0.395909   1.972  0.04861 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1148.48  on 828  degrees of freedom
Residual deviance:  603.15  on 812  degrees of freedom
AIC: 637.15

Number of Fisher Scoring iterations: 14
```

I used these variables in a new logistic regression model. The AIC is much lower, which indicates this model is a better fit than the full model.

```
Call:
glm(formula = Severity ~ Age + Shape + Margin, family = binomial(),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5814  -0.6541   0.2034   0.5600   2.7090

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.146865   0.564682   9.115  < 2e-16 ***
Age         -0.060890   0.008961  -6.795 1.08e-11 ***
Shape2       0.279764   0.357516   0.783  0.43391
Shape3      -0.592098   0.412716  -1.435  0.15139
Shape4      -1.501761   0.372058  -4.036 5.43e-05 ***
Margin2     -1.183966   0.643647  -1.839  0.06585 .
Margin3     -0.998893   0.391575  -2.551  0.01074 *
Margin4     -1.253405   0.332070  -3.775  0.00016 ***
Margin5     -2.173722   0.430883  -5.045 4.54e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 919.90  on 663  degrees of freedom
Residual deviance: 572.38  on 655  degrees of freedom
AIC: 590.38

Number of Fisher Scoring iterations: 5
```

I trained the model using training data and ran the test data through the model. Using a decision threshold of 50%, the model provides a prediction accuracy of 80% and a recall of 79%. To eliminate Type II Errors, I have decided to decrease the decision threshold to a value that will produce 100% recall. Using a decision threshold of 6%, the model provides a prediction accuracy of 58.2% but 100% recall. The full decision tree model indicates BI.RADS, Age, Shape, and Margin are the most significant variables.



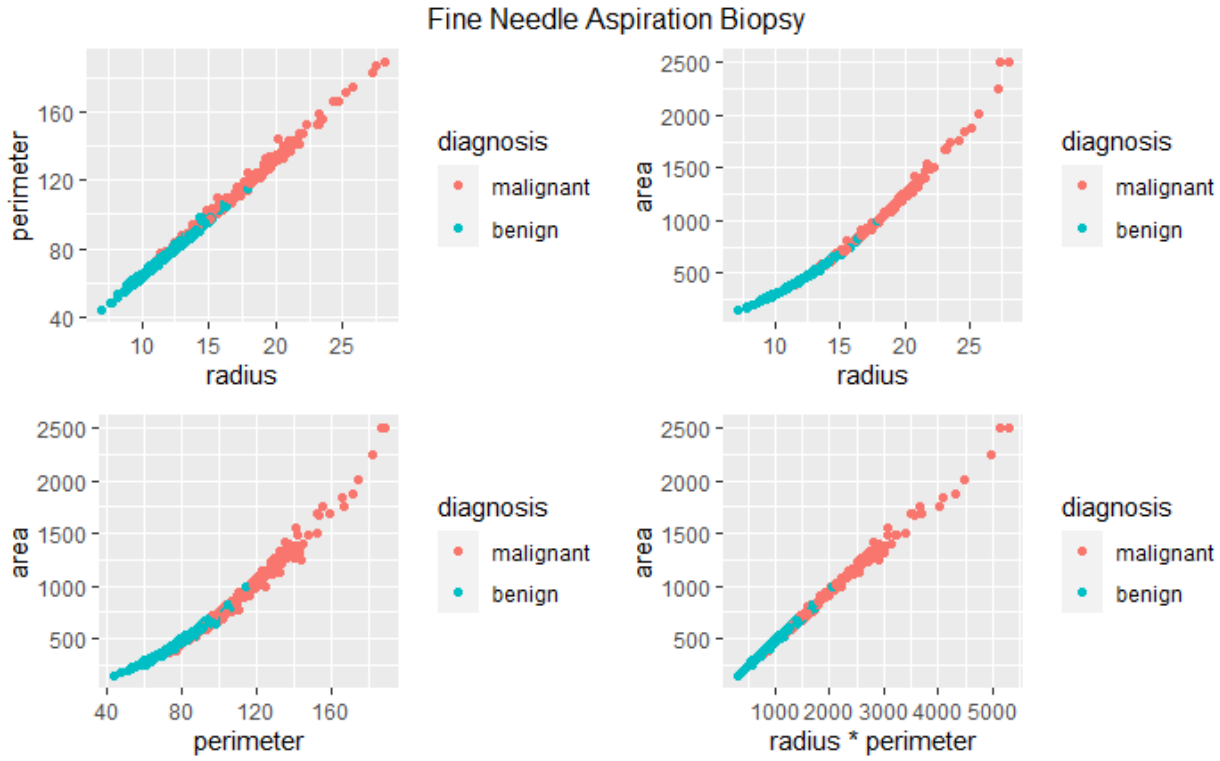
Using the training and testing data, the decision tree model provides a prediction accuracy of 83.6% and a recall of 79.8%. The k-nearest neighbors model, when k = 29, provides a prediction accuracy of 80% and a recall of 85%. Using cross validation on a random forest model, BI.RADS is indicated as the only

significant variable. The random forest model provides a prediction accuracy of 81.8% and a recall of 84.7%. The decision tree model provided the most prediction accuracy; however, the best model is the logistic regression model with decision threshold of 6% with 100% recall, as seen in the table below.

	Accuracy	Recall
Logistic Reg - 50%	0.8000000	0.7901235
Logistic Reg - 6%	0.5818182	1.0000000
Decision Tree	0.8363636	0.7977528
kNN	0.8000000	0.8500000
Random Forest	0.8181818	0.8472222

Fine Needle Aspiration Biopsy

I need to decide how to proceed with the radius, perimeter, and area variables. I will begin by using scatter plots to examine the relationship between them.



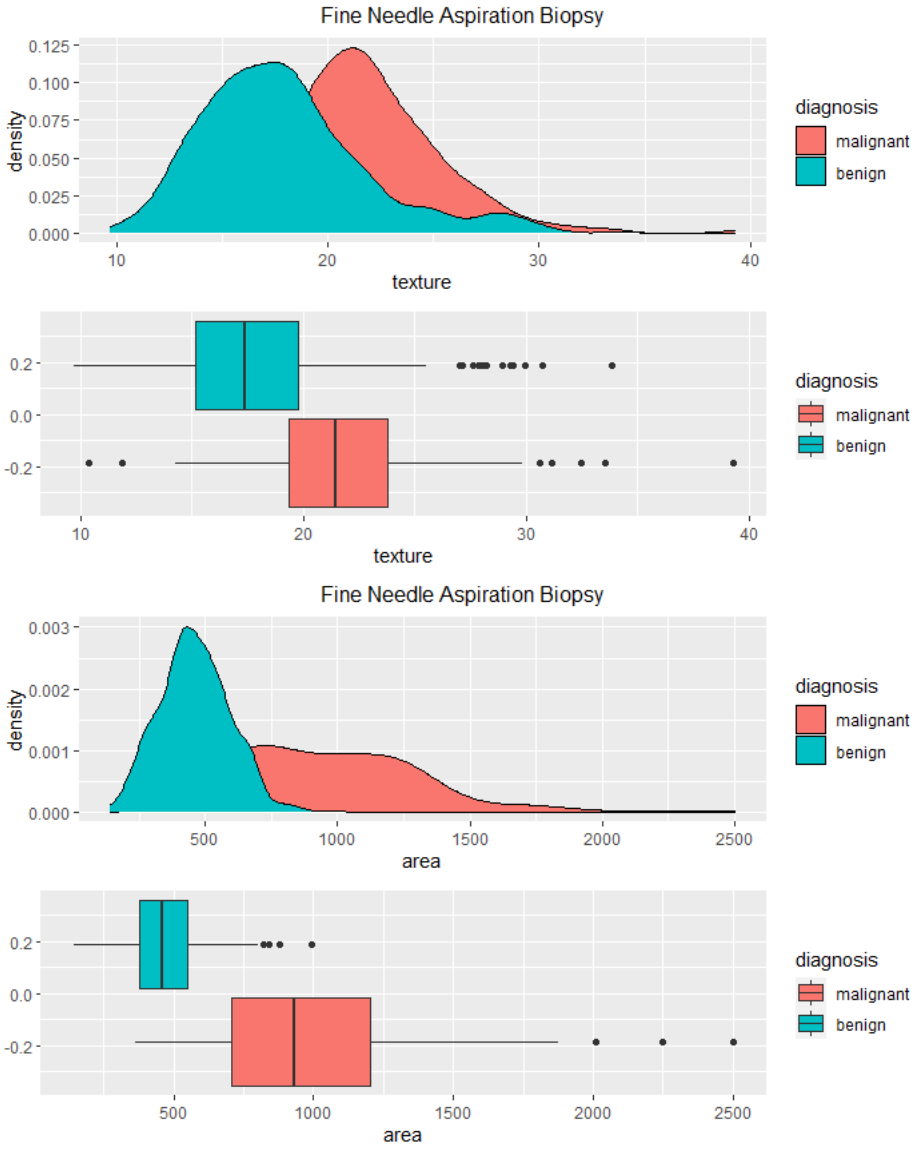
The variables are highly correlated, as suspected. To help decide which two variables to removed from the data frame, I will examine their correlation coefficients.

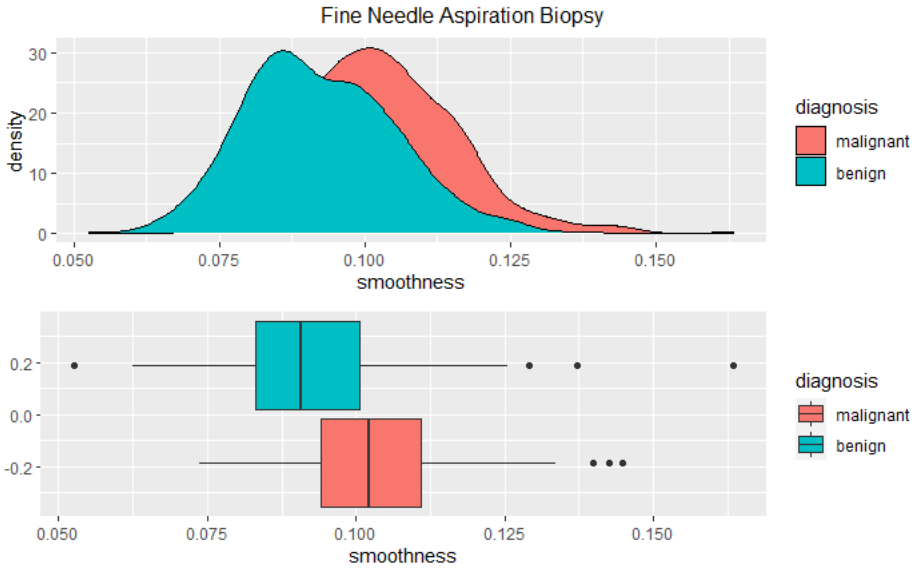
	Correlation
radius vs perimeter	0.9978553
radius vs area	0.9873572
perimeter vs area	0.9865068
radius*perimeter vs area	0.9982195

The greatest correlation is between the product of the radius and perimeter variables and the area variable. Therefore, I will remove the radius and perimeter variables. I want to examine the correlation between the remaining variables.

	texture	area	smoothness
texture	1.00000000	0.3210857	-0.02338852
area	0.32108570	1.0000000	0.17702838
smoothness	-0.02338852	0.1770284	1.00000000

None of the remaining variables are highly correlated with the others. Next, I want to compare the variables between the two diagnosis groups, malignant and benign. I will use density and boxplots for visual aids.





Based on the density and boxplots, it appears there is a significant difference in texture, area, and smoothness between the two diagnosis groups. I will use the Mann-Whitney U Test to show verify.

```
        Wilcoxon rank sum test with continuity correction

data:  biop$texture by biop$diagnosis
W = 58718, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0


        Wilcoxon rank sum test with continuity correction

data:  biop$area by biop$diagnosis
W = 71016, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0


        Wilcoxon rank sum test with continuity correction

data:  biop$smoothness by biop$diagnosis
W = 54647, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

The Mann-Whitney U Test verifies these claims. I will use these variables in a logistic regression model.

```
Call:
glm(formula = diagnosis ~ ., family = binomial(), data = biop_new)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.85800  -0.01878   0.04419   0.20032   2.14463

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.262e+01  3.608e+00   9.040  < 2e-16 ***
texture     -3.811e-01  5.768e-02  -6.607 3.92e-11 ***
area        -1.626e-02  1.827e-03  -8.901  < 2e-16 ***
smoothness  -1.468e+02  1.922e+01  -7.636 2.25e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 182.53  on 565  degrees of freedom
AIC: 190.53

Number of Fisher Scoring iterations: 8
```

I trained the model using training data and ran the test data through the models. Using a 50% decision threshold, the logistic regression model provides a prediction accuracy of 92% and a recall of 86.7%. To eliminate Type II Errors, I decreased the decision threshold to 26%. This provides a prediction accuracy of 94.7% and 100% recall. The decision tree model provides a prediction accuracy of 90.3% and a recall of 83%. The k-nearest neighbors model, when k = 23, provides a prediction accuracy of 86.7% and a recall of 71.4%. The random forest model provides a prediction accuracy of 92.9% and a recall of 90.5%. The logistic regression model with a 26% decision threshold provides the best prediction accuracy with perfect recall, as seen in the table below.

	Accuracy	Recall
Logistic Reg - 50%	0.9203540	0.8666667
Logistic Reg - 26%	0.9469027	1.0000000
Decision Tree	0.9026549	0.8297872
kNN	0.8672566	0.7142857
Random Forests	0.9292035	0.9047619

The only variable found to be significant in predicting the survival of a patient was the number of positive axillary lymph nodes. Axillary lymph nodes are taken from the axilla, or the armpit. When cancer cells are detected in the axillary lymph nodes, it indicates metastasis. As observed in the analysis, smaller amounts of positive lymph node increased the probability of surviving for longer than five years after surgery. These finding emphasizes the importance of recall accuracy, early detection, and timely treatment

The best recall results were obtained by decreasing the decision threshold in the logistic regression models. There may similar methods available to use on the other classification models; however, time limitations did not allow for further methods to be explored. The survival analysis conducted were based on data gathered from patients who had surgery between 1958 to 1970. Using current data would likely produce different results.

I used this data to find the classification models that predicted a diagnosis with high accuracy and recall. As breast cancer continues to affect so many, I feel further research into risk factors associated with its development would provide some interesting insights. These insights could decrease false negative results and increase survival.

References

Bell, D., Weerakkody, Y., et al (2013). Breast imaging-reporting and data system (BI-RADS): Radiology Reference Article. Retrieved July 29, 2020, from https://radiopaedia.org/articles/breast-imaging-reporting-and-data-system-bi-rads?lang=us

Biswas, K. (2019, July 19). Haberman Cancer Survival Data Set. Retrieved July 20, 2020, from www.kaggle.com/krpiku/haberman.csv?select=haberman.csv

Limitations of Mammograms: How Often are Mammograms Wrong? (n.d.). Retrieved August 06, 2020, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html

Ovsen. (2016, October 31). Mammographic Mass Data Set. Retrieved July 20, 2020, from www.kaggle.com/overratedgman/mammographic-mass-data-set

Singh Suwal, M. (2018, September 26) Breast Cancer Prediction Dataset. Retrieved July 20, 2020 from, www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset