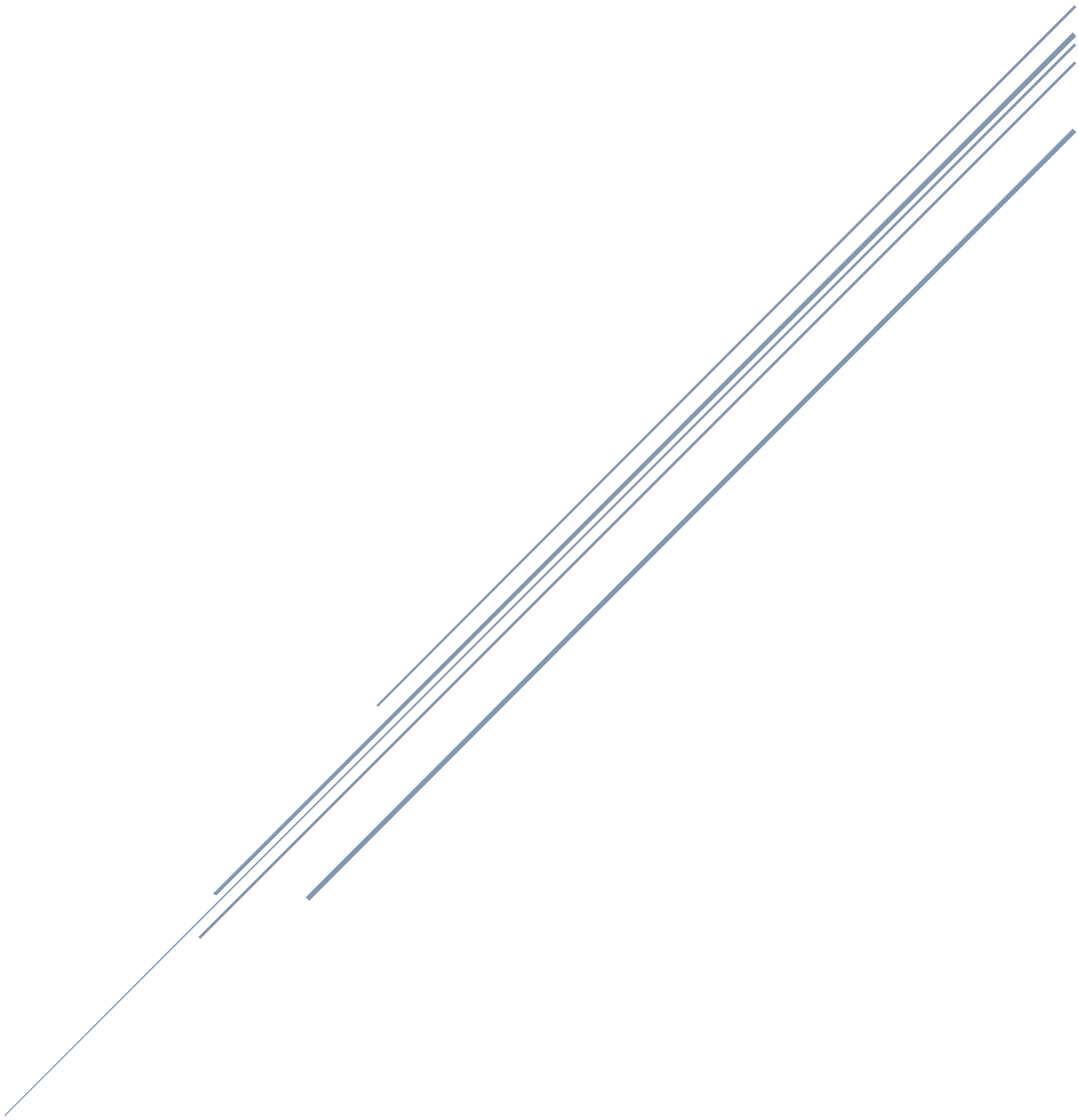
BREAST CANCER PREDICTION USING MULTIVARIANT TECHNIQUES

Understanding the Use of Breast Cancer Cell Measurements in the Prediction of Breast Cancer

Yolanda Lewis, Daniel O'Brien, Jose Guzman, Will Ranick, Ross Gibson



Abstract

Breast cancer is one of the most common types of cancer and early detection is important for survival. Our goal is to create a model to predict whether a breast tumor is benign or malignant based on known measurements of the tumor. More specifically, we would like to determine how breast cancer cell measurements affect the predictability of a breast cancer diagnosis. Principal Component analysis as well as Common Factor Analysis were used as exploratory measures to evaluate the relationship between the independent variables and logistic regression, LDA was used to predict the response variable to determine breast cancer diagnosis. The PCA identified 5 components with a cumulative variance of 84.7%. An association between radius, perimeter, and area measurements, is one example of a relationship identified by PCA. The factor analysis demonstrated that the shape (smoothness, compactness, and symmetry) of a cell is the most important factor in classification. Logistic regression showed that the odds of breast cancer diagnosis are increased with the use cell measurements such as worse mean symmetry, texture standard error, and smoothness mean. Based on the different analyses, strong relationships between the independent variables' radius, perimeter, and area were identified which allowed us to better understand the data and enabled us to create a model to predict if a breast mass was benign or malignant with 94.69% accuracy.

Executive Summary

According to Dr. Koriech in his 1996 journal article titled *Breast Cancer and Early Detection*, "breast cancer is the most frequent female cancer in developed and many developing countries" [2]. Dr. Koriech goes on to say that the survival rate of cancer can improve with the early detection of malignancy and "early detection of cancer or a precursor can result in less radical treatment and improve prognosis of some cancers" [2]. Our objective is to create a model that will allow us to predict whether a breast cell mass is benign or malignant based on measurements from a digitized image of a fine needle aspirate (FNA) biopsy. In addition, we would like to better understand the relationship between the different cell measurements using classification and dimension reduction techniques. More specifically, we would like to determine how breast cancer cell measurements affect the predictability of a breast cancer diagnosis.

The dataset we used for our analysis is a public dataset from Kaggle that contains FNA biopsy features as well as whether the biopsy was determined to be benign or malignant (represented as either a 'M' or a 'B'). The dataset contains 569 observations, or FNA biopsies, and contains one categorical diagnosis variable as well as 30 variables representing different features of the FNA biopsy. There are 10 distinct measures/features extracted from the digital FNA biopsy image and are represented in the dataset in three different ways: the mean, the standard error, and the worst for each feature measurement.

Fine needle aspirate (FNA) is a method where a thin needle draws cells from a mass underneath the skin. This dataset is comprised of different measures of cancer cells collected through FNA. There are several different measurements taken from the cancer cells, and these different measurements are represented as our independent variables.

The Principal Component Analysis (PCA) was used to determine relationships between the different independent variables. Through the PCA, we will determine which variables are associated with each other, and which variables share little association with the other independent variables.

Factor Analysis was used to test the theory that increases in various cell features led to a higher likelihood for a cell to be classified as malignant. The analysis was used to solidify this belief by discovering associations between the measurements and the ultimate classification of a cell.

Logistic regression was used to determine the probability that the measurements provided in the dataset would have an impact on determining breast cancer diagnosis. The log odds were used to determine the likelihood that an increase or decrease in the odds of predicting a breast cancer would occur for every unit increase in the predictor variables.

Linear Discriminant Analysis (LDA) was used to create a model that would classify an FNA biopsy image as benign or malignant based on the features of the biopsy image. This type of model requires that the independent variables, or the measurements taken from the image, are independent of each other and are normally distributed.

Through the PCA associations between radius, perimeter and area measurements became apparent, as well as associations between smoothness and compactness measurements and the standard error of compactness, concavity, and fractal dimensions. This presents information on how the variables relate to one another.

Factor analysis ultimately led to the confirmation that there, in fact, are associations between several measurements and the classification of the cell. The overall shape of a cell has a significant impact on the classification.

Based on the logistic regression, the worst mean measurement of a fine needle aspirate is the most predictive at determining if a patient has a malignant breast tumor. It was also observed that there is the least likelihood of a breast cancer diagnosis when the standard error increases.

We were able to successfully create a statistical model using an analysis called Linear Discriminant Analysis (LDA) that was able to predict with 94.69% accuracy whether a FNA biopsy image was benign or malignant.

The dataset focuses on a small sample set of patients from Wisconsin. In future works, it would be helpful to expand the dataset to more patients from other areas. More health factors such as smoker status, family history, and diet should also be included in the dataset to include more variables for predicting a breast cancer diagnosis. Additionally, the data would benefit from a larger number of tumor samples as well as more samples from the same tumor.

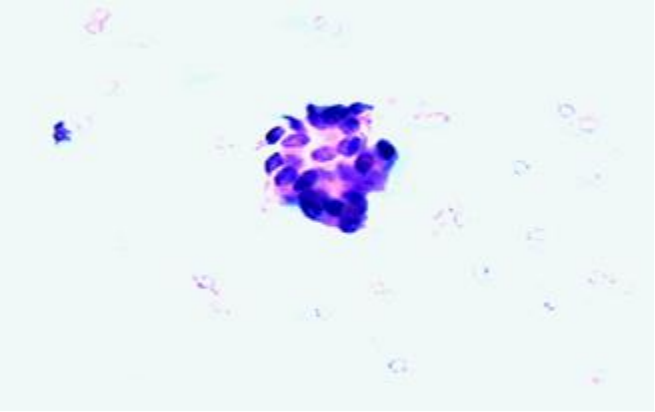
From the analyses performed, it appears that factors that indicate cell irregularity, such as texture, lack of smoothness, and lack of symmetry, are more common in the cells of a malignant tumor. These analyses enabled us to create a model that could predict whether a biopsy image of a breast cancer tumor was benign or malignant with 94.69% accuracy. We also determined that the likelihood of determining a breast cancer diagnosis is increased when features such as symmetry, texture and smoothness are determined from FNA.

Introduction

The dataset used for analysis contained 569 observations and 32 variables from patients undergoing breast cancer testing in the state of Wisconsin. The observations in the dataset represent patient data from digitized images of a fine needle aspirate (FNA) of a breast mass. According to Casaubon et al, a fine needle aspiration is a type of biopsy that is performed with a small needle to obtain samples of tissue and fluid from a breast lesion [6]. Digitized images were taken of the cells obtained from the tissue and fluid samples, and measurements were taken of the cell nuclei to determine a breast cancer diagnosis. More specifically, the computerized images contained measurements of the center of the cells nuclear membrane known as the cell nuclei. According to Casaubon et al, fine needle aspiration is controversial for diagnosis of malignant tumors because of the small sample size of breast tissue that is sampled and the high rate of inadequate samples [6].

Previous research has found that evaluating and diagnosing breast cancer using an interactive computer system to scan FNA slides can have accuracy over 95% (Wolberg 1994). Many other studies on breast cancer use different information regarding breast cancer cells, such as percent density of cancer cells and texture features. These studies have the same objective of identify the presence of breast cancer (Manduca 2009). The features in the data are 10 types of measurements of the cell nuclei that contain 3 distinct measurements. The measurements include the mean, standard error, and the worst or largest mean of the three largest values. These features allowed for a comprehensive analysis to determine how breast cancer cell measurements affect the predictability of a breast cancer diagnosis.

Below is an example image that represents a single observation in our dataset [5]



Literature Review

Machine Learning for Cancer Diagnosis

[1] The authors, Wolberg, Street, and Mangasarian, were able to prove in their March 1994 journal article published in *Cancer Letters* that a machine learning algorithm can diagnose breast cancer from processed images of fine needle aspirates. These authors proved that a statistical model could in fact diagnose an FNA biopsy as benign or malignant with a high degree of accuracy, which is what we are trying to test as well.
https://www-sciencedirect-com.ezproxy.depaul.edu/science/article/abs/pii/030438359490099X

Comparing Data Mining Techniques

[3] This is a 2013 article comparing multiple data mining techniques and their accuracy in correctly classifying the Wisconsin Breast Cancer observations (Our dataset). The authors compare multiple different approaches to classifying this dataset. It is evidence that there are multiple ways to solve this problem. Some examples in this article include, logistic regression, decision trees and K-Nearest Neighbors. Since we are trying to find the best way to predict whether a breast cell mass is benign or malignant, we can use this previously done comparison as a foundation for comparing our methods.
https://pdfs.semanticscholar.org/b5fd/6d83d2e0a2e96152522e1df50052b8c84fd0.pdf

Cell Texture Features to Predict Breast Cancer Risk

[4] Percent density along with image texture features are used to determine if these features can be used to predict breast cancer risk in this 2010 study by Manduca, Carston, Heine and their team from the Mayo Clinic. The authors concluded that texture features originating from mammographic images can predict breast cancer risk at the same extent as percent density.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674983/

Methods

In preparation to analyze the data, a correlation matrix and a correlation plot were used to determine any associations between the variables in the dataset. Variables displaying values of 0.7 or higher were considered to have a strong association. Checking for multicollinearity was appropriate as it can cause inaccuracies in the computation of the betas when performing certain types on analysis such as regression.

Principal Component Analysis (PCA) was used as an exploratory step to assess associations between independent variables. Understanding the relationships between the independent variables can provide insight when building a prediction model using these variables. The number of components selected will assist in accounting for a certain level of variance in our variables. To find the number of components necessary for our PCA a variety of factors will be taken into consideration, including eigenvalues, parallel analysis, the scree plot, and the cumulative variance.

Common Factor Analysis was used to reduce the number of variables to a smaller number of factors as well as to concisely describe the relationship among those variables. We want to reduce the number of variables to interpret them. We want to take the three variations of each observation and minimize them to interpret them more easily. The dataset consisted of 3 variations of the same 10 measurements. The complexity of 3 sets of 10 very similar measurements was not ideal in understanding the relationship between the measurements.

Linear Discriminant Analysis (LDA) was used to create a model using the tumor cell measurements (independent variables) to predict whether the tumor was benign or malignant (dependent variable). An LDA analysis attempts to predict a categorical variable based on a set of independent variables by creating a line, or “decision boundary”, that is used to classify new prediction data.

To perform an LDA analysis, there are a few assumptions that need to be validated before a successful and accurate model can be created. There are three assumptions that cannot be violated:

predictors must be normally distributed, independent, and homoscedastic. We also need to make sure that there is no multicollinearity between our predictors. So, the first step that was completed was to run some exploratory graphs and correlation tables in R to make sure that none of these assumptions were violated. After reviewing the correlation tables and VIF scores of all predictors in our dataset, it was apparent that many predictors were correlated to one another based on correlations greater than 0.7 and high VIF scores above a value of 10. I removed one predictor at a time (the one with the highest VIF score) and then reran the VIF scores until I was left with a set of predictors that had VIF scores all below 10. After this check, we were left with 17 predictors that did not exhibit multicollinearity and 1 response variable which was our diagnosis variable (refer to 'Figure 10' below which shows the predictors that were left in the LDA dataset after solving for multicollinearity).

After we confirmed that all assumptions to run an LDA were satisfied, we split our data into a training set (80% of our data) and a testing set (20% of our data). We then used our training dataset to create the actual LDA model using the 'MASS' package in R. We can see a plot in 'Figure 11' showing the distribution of diagnosis using this training dataset. After the LDA model has been created we used 'R' to predict the diagnosis of the predictors in our test dataset and calculated a confusion matrix to show the number of false positives and negatives in our prediction ('Figure 12'). Finally, we calculated the accuracy of our LDA model in terms of its ability to predict the dependent diagnosis variable using our test dataset.

Logistic regression was used to determine the variables in the dataset that were most significant when predicting a breast cancer diagnosis. Logistic regression is a method used to fit a regression curve when the response variable is categorical. Binomial logistic regression was used as the response variable was binary. Multicollinearity was checked for by observing a correlation plot (see figure 1) and checking the correlation matrix for the dataset. Variables representing a correlation value of 0.7 or higher were removed from the dataset. The data was split into 80% training and 20% testing. The training set was used to fit the model, which was later used for testing over the test set. Backwards elimination was used for feature selection. The accuracy of the model was determined using cross validation, a confusion matrix, and calculation of the misclassification error. An ROC plot was created to test the model performance. The specificity and sensitivity were also calculated for the model. The odds ratios of the final model were used to interpret the betas coefficients and the variance inflation factor of the final was determined to ensure there was no multicollinearity present.

Results

When performing the PCA, Bartlett's Test of Sphericity returned with a p-value less than 0.05. This demonstrates there are samples that are not equal to zero and there is enough variation to run the model. Cronbach's alpha test returned a raw alpha value of 0.58, with a 95% confidence level of 0.58-0.59, this value, remaining above 0.5 shows that we can assume a reasonable level of reliability. The Kaiser-Meyer-Olkin (KMO) factor adequacy returned a value of 0.83, this could be an indication that the data is appropriate for factor analysis.

The number of components used in the PCA were selected through parallel analysis, eigenvalues, the scree plot, and cumulative variance. The parallel analysis suggested using 5 components, 6 components had eigenvalues greater than one, and according the scree plot, the 'knee' appears to indicate that 3 or 4 components may be appropriate. The decision to use 5 components was reached by reviewing the PCA with 4, 5 and 6 components. The PCA with 6 components included two

components with only 2 variables, and both the PCA with 4 and 5 components had 1 component with only two variables. Since both 4 and 5 components resulted in a single component with two variables, and the proportional variance was higher with 5 components, 5 components were used in the final PCA.

The 5 components are Measurement Summary, made up of the mean measurements of radius, perimeter, area, concavity and concavity points as well as the standard error of radius, perimeter, area and the worst measurements of radius, perimeter, area, and concave points. The second component, smoothness/compactness, is made up of the mean and worst of smoothness, compactness, and symmetry as well as the worst fractal dimension measurement. The third component, concavity standard error, included the standard error of compactness, concavity, concave points, and fractal dimensions. The fourth component, texture standard error, is made up of the standard error of texture, smoothness, and symmetry. And the fifth component, texture, is made up of the mean and worst measurement of texture. The cumulative variance for these 5 components is 84.7% (Figure 7).

Correlation matrices were created to evaluate correlations within the 3 variations of the data set. The 3 variations of these measurements are mean, standard error and "worst" which is the mean of the three largest values. With this information in mind, three subsets of the data were created to check for multicollinearity within the subsets. (Figures 2-4) With this information, the decision was made to remove all variations of the area, radius, fractal, concave points, and concavity measurements. These measurements contained a high degree of multicollinearity. (Figure 5)

To begin the factor analysis, the following tests were completed. The KMO test was performed and resulted in an MSA of .63. This means the sample size is adequate but not great. The Barlett's Test of Sphericity resulted in a p-value that was about 2.22e-16. This means there is enough variation in the dataset to run the model. Cronbach's alpha test resulted in a value of .60. with a 95% confidence interval between .59 and .61. The number of factors were chosen with the use of a scree plot using the Knee method. The Eigenvalues greater than 1 method also supported the suggested number of factors which is 3. (Figure 8)

The 3 factors were determined to be Shape, Perimeter and Texture of the cell. Shape contains the means of smoothness, compactness, and symmetry. It also contains the standard error of compactness along with the worst measurements of smoothness, compactness, and symmetry. Perimeter contains the mean, standard error, and worst measurement of perimeter. Texture contains the mean, standard error, and worst measurement of texture. (Figure 9). Our Linear Discriminant Analysis model that used FNA biopsy measurements from a digital image scan was able to predict the diagnosis of the cell mass as benign or malignant with 94.69% accuracy. The LDA model's false positive rate was 0.0% and its false negative rate was 14.3% (refer to 'Figure 12').

The logistic regression analysis was used to predict the probability that the measurements used in the model would lead to a diagnosis of a malignant or benign tumor. To deal with the multicollinearity in the data, all measurements representing the least amount of multicollinearity (correlation values of 0.7 or lower) were selected to be used in the model (refer to figure 1). These measurements included the means of smoothness and symmetry, the standard errors of texture, smoothness, symmetry, and fractal dimension, and the worst measurement of symmetry. The features in the final model were determined using backwards elimination. The following variables were identified using backwards elimination: smoothness mean, texture standard error, symmetry standard error, and the worst symmetry measurement. Refer to figure 13.

The p-values for all the variables showed significance, with all p-values less than 0.05. The worst mean symmetry measurement was the most significant variable with a p-value of 2e-16. Based on the p-values, the null hypothesis was rejected and all the estimates in the model should be used to determine the probability of a breast cancer diagnosis. The log odds of each estimate were calculated to determine the odds of the given variables ability to predict the diagnosis variable. The log odds of the predictor variables were determined by holding the corresponding predictor variables at a fixed value.

The estimate for smoothness mean was 0.62. The odds ratio of smoothness mean was calculated to be 1.87. So, for every one unit increase in the smoothness mean, there is an 86% increase in the odds of diagnosing breast cancer. The estimate of texture standard error was 0.51. The odds ratio of the texture standard error was calculated to be 1.66. So, for every one unit increase in the texture standard error, there is a 66% increase in the odds of diagnosing breast cancer. Symmetry standard error had a negative estimate of -0.97. The odds ratio was calculated to be 0.37 which represents a less likelihood of being diagnosed with breast cancer based on this measurement. Based on the log odds ratio there is a negative relationship between the symmetry standard error and having a malignant tumor. Based on this, we would see a 62% decrease in the odds of a malignant diagnosis. Lastly, the worst symmetry measurement had an estimate of 1.45 and the odds ratio was 4.25. Based on the log odds we would see a 325% increase in the odds of a breast cancer diagnosis with every one unit increase in the worst symmetry measurement. Refer to figures 14 and 15 for logistic regression model and odds ratios.

The accuracy of the model was determined using 10 -fold cross validation which yielded a 76% accuracy. A receiver operating characteristic (ROC) curve (refer to figure 16) was used as an additional measure to determine the logistic model performance. The area under the curve (AUC) showed that the model has an AUC of 75 % which indicated that the model was performing at an acceptable rate for predicting a breast cancer diagnosis. The confusion matrix based on the test set yielded 81% accuracy (refer to figure 17). The misclassification error was calculated based the testing data predictions. The misclassification error was 19%. The specificity of the model was 97% and the sensitivity was 21%.

Discussion

The PCA confirmed that an association between radius, perimeter and area measurements existed in the dataset. Additionally, measure of texture was found to share little association with the other variables present in our dataset. Uncovering these relationships led to a better understanding of the dataset and variables.

Before beginning the factor analysis, we had a suspicion that a cell's shape had a strong correlation with its classification. This is because a cell being classified as malignant is caused by irregularities in the cell. After completing the factor analysis, we found that a cell's shape (smoothness, compactness, and symmetry) is the strongest indicator of classification. The factor analysis confirmed our initial thoughts, and this information was used in conjunction with other findings to create our final model.

For the PCA and CFA, a different number of components and factors were used. 5 components were used for the PCA and 3 factors were used for the CFA. There are some similarities regarding the grouping of the different variables. The first factor and second component both include the mean of smoothness, compactness, symmetry, and the worst measurement of smoothness, compactness, and symmetry. Additionally, the mean and worst measurements for texture both belong to component 5 and factor 3. There are also a few differences in terms of variable groupings and levels of proportional and cumulative variance between PCA and CFA. (Figures 7 & 9).

The logistic regression model showed that the likelihood of predicting a malignant breast mass appeared to increase when using cell nuclei measurements to determine a breast cancer diagnosis. It

was determined that the smoothness mean, texture standard error, and the worst symmetry have a positive relationship with breast cancer diagnosis. The standard error of symmetry had a negative correlation which might suggest that the likelihood of being diagnosed with a malignant tumor decreases with the use of this measurement. The inclusion of the worst measurement of the symmetry of the cell nuclei appeared to be the most significant variable when determining a breast cancer diagnosis.

All three prediction models that were used in our analysis had several variables in common after completing our exploratory analysis and checking for multicollinearity. This means that we were all using similar predictors in our analyses to try to determine the diagnosis of FNA biopsies. Because we used similar inputs in our models, we were able to compare the results and predictability of our models and determined that LDA was more accurate when during a breast cancer diagnosis.

Limitations and Future Works

There are innumerable possibilities of measurements that could be taken which could relate to tumor malignancy. Although the data already accounts for several factors (radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry), there are more measurements that might be helpful in determining malignancy, such as aspirate mass. It is likely that several measurements were excluded due to the inability to compute them in the same way that the present measurements were computed.

Additionally, it is unclear if multiple aspirates were taken from the same patient. Since tumor structure can be unpredictable, it is possible that different parts of the same tumor may possess significantly different measurements. Particularly, if nuclei were not aspirated from the tumor at complete random, it is possible that the data could be skewed in a way that causes the diagnosis criteria to be either too broad or too narrow. The greater the number of aspirates measured as well as the number of individuals from which samples were collected, the more likely it is that the data will be truly representative of the population.

Future work to be conducted on the dataset could include using the components from PCA or factors from the CFA and using those components or factors in the logistic regression or the LDA. Additionally, a hierarchical cluster analysis should be conducted to provide a better understanding of patients who are diagnosed with breast cancer versus those that are not.

The dataset also focuses on a small sample set of patients from Wisconsin, IL. In future works, it would be helpful to expand the dataset to more patients from other areas. More health factors such as smoker status, family history, and diet should also be included in the dataset to include more variables for predicting a breast cancer diagnosis.

Conclusion

Overall, it appears that the measurements provided were extremely useful in determining the breast cancer diagnosis of the patients undergoing fine needle aspiration of a breast mass. Based on the analysis performed, we were able to identify strong relationships between the cell measurements representing the radius, perimeter, and area which allowed us to better understand the data and enabled us to create a model that could predict whether a biopsy image was benign or malignant with 94.69% accuracy.

From the analyses performed, it appears that factors that indicate cell irregularity, such as texture, lack of smoothness, and lack of symmetry, are more common in the cells of a malignant tumor. Logistic regression helped to determine the odds of the different measurements predicting a patients'

breast cancer diagnosis. The odds of being diagnosed with a malignant tumor appear to increase when measurements such as texture standard error, smoothness mean, and the worst symmetry measurements are provided for a fine needle aspirate of a breast cancer cell. Therefore, it appears that the inclusion of breast cancer cell measurements taken from a fine needle aspirate affect the predictability of a breast cancer diagnosis.

Appendix I (Citations)

[1] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, *77*(2-3), 163-71.

[2] Koriech O. M. (1996). Breast cancer and early detection. Journal of family & community medicine, 3(1), 7–9.

[3] Kumar, G. R., Ramachandra, G. A., & Nagamani, K. (2013). An efficient prediction of breast cancer data using data mining techniques. *International Journal of Innovations in Engineering and Technology (IJIET)*, *2*(4), 139.

[4] Manduca, A., Carston, M. J., Heine, J. J., Scott, C. G., Pankratz, V. S., Brandt, K. R., Sellers, T. A., Vachon, C. M., & Cerhan, J. R. (2009). Texture features from mammographic images and risk of breast cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *18*(3), 837–845. https://doi.org/10.1158/1055-9965.EPI-08-0631

[5] 92_4964. (1994, November 3). Retrieved August 20, 2020, from ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/cancer_images/

[6] Casaubon JT, Tomlinson-Hansen S, Regan JP. Fine Needle Aspiration of Breast Masses. [Updated 2020 Aug 12]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470268/

Appendix II (Figures/Plots)

Figure 1. Correlation Plot

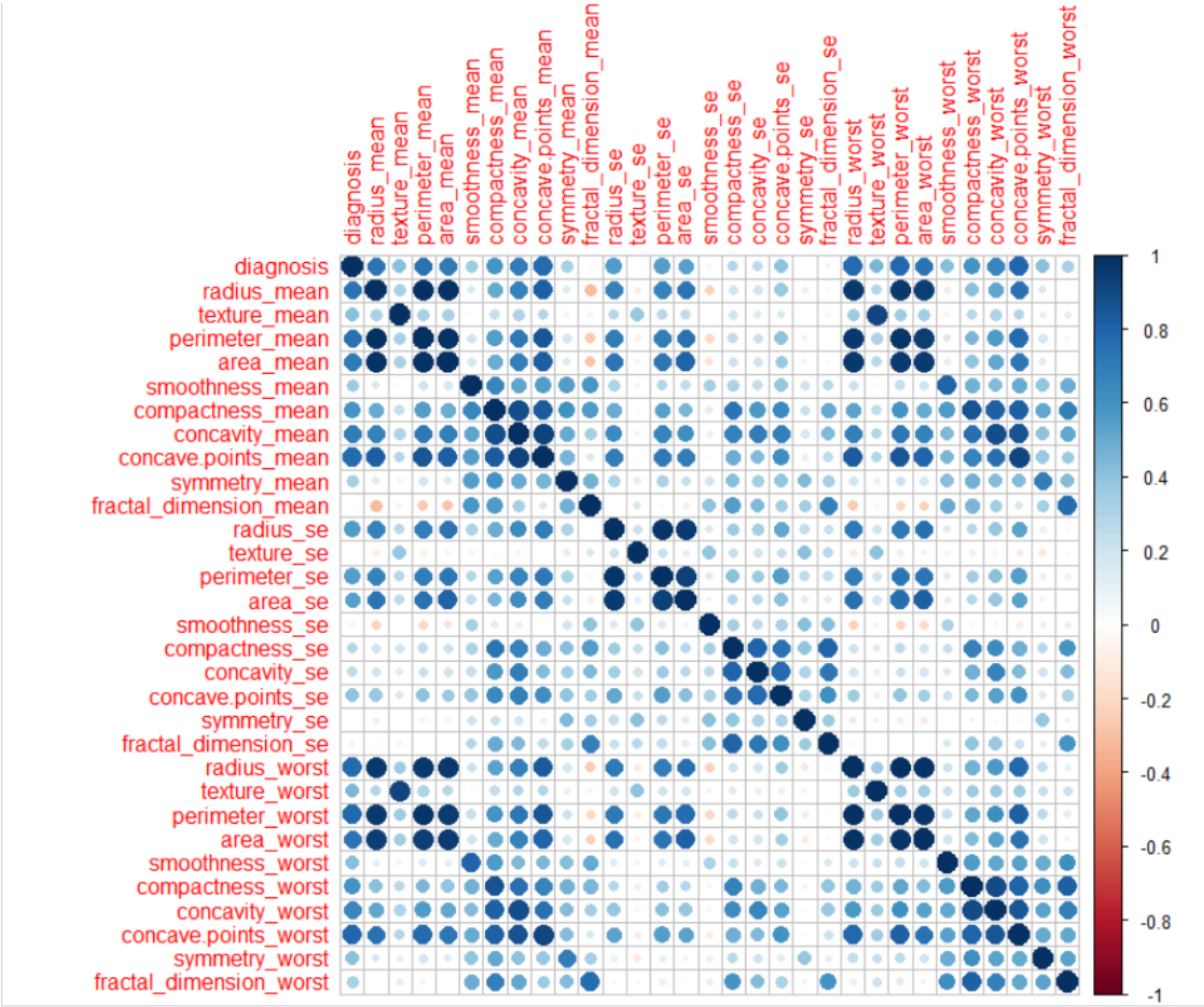


Figure 2. Mean Correlation Plot

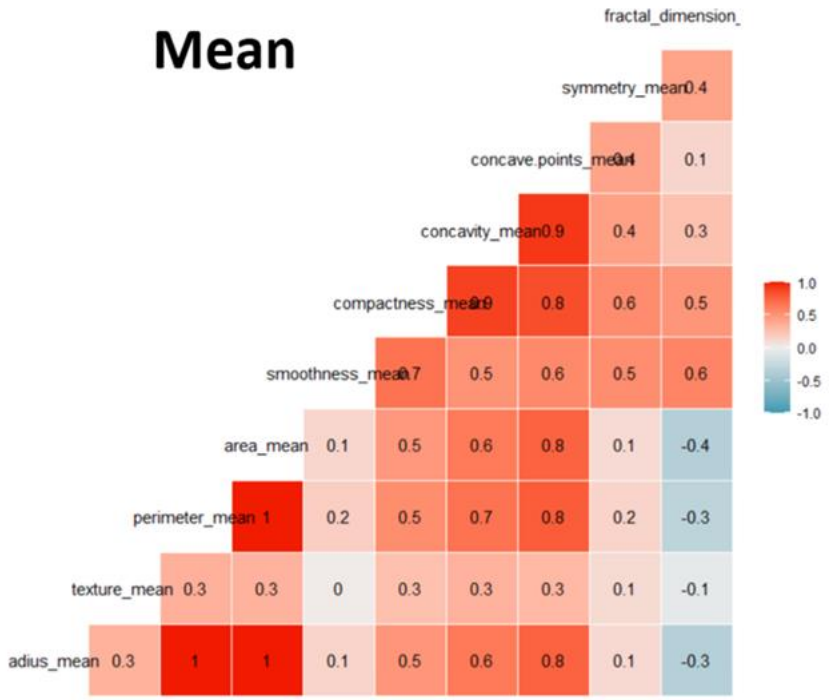


Figure 3. Standard Error Correlation Plot

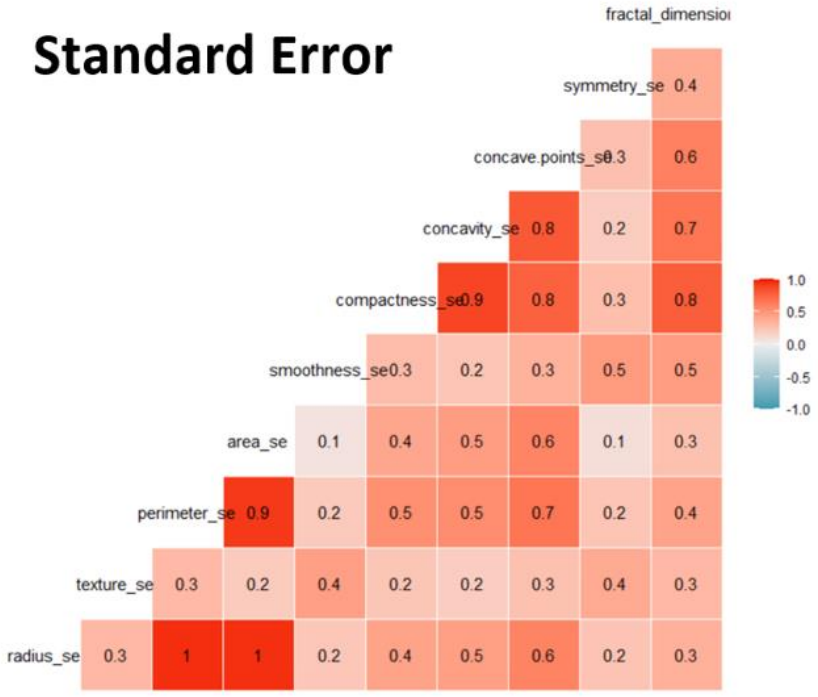


Figure 4. Worst Correlation Plot

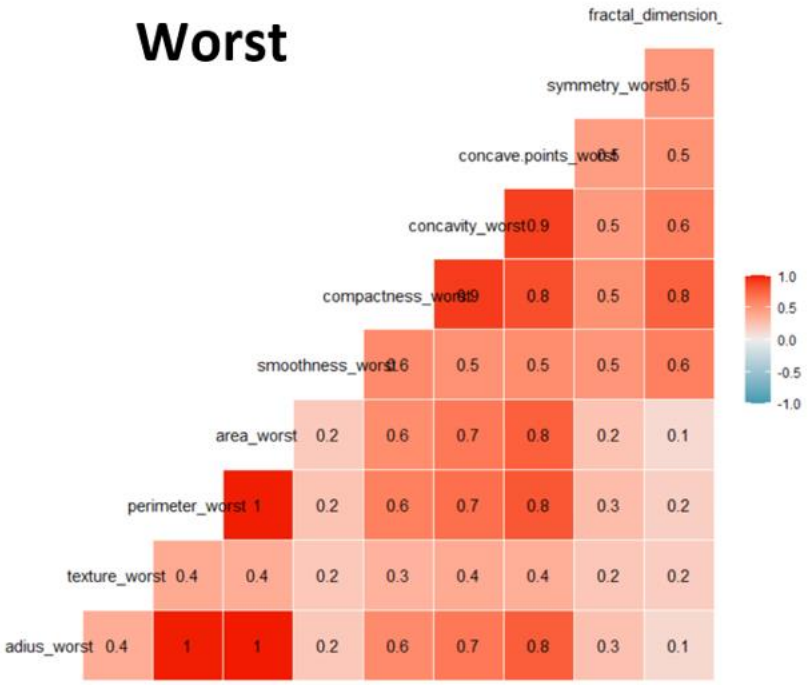
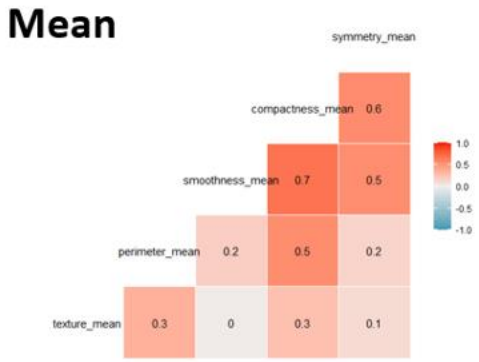
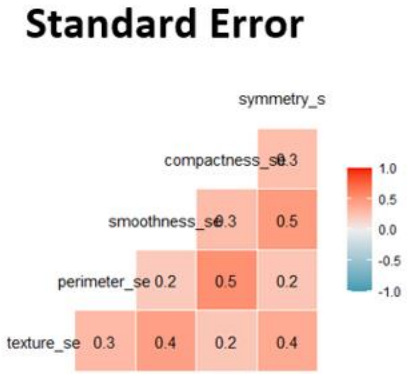


Figure 5. Final Correlation Plots Within Variations





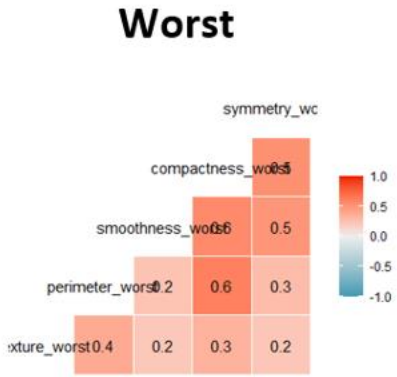


Figure 6. Scree Plot

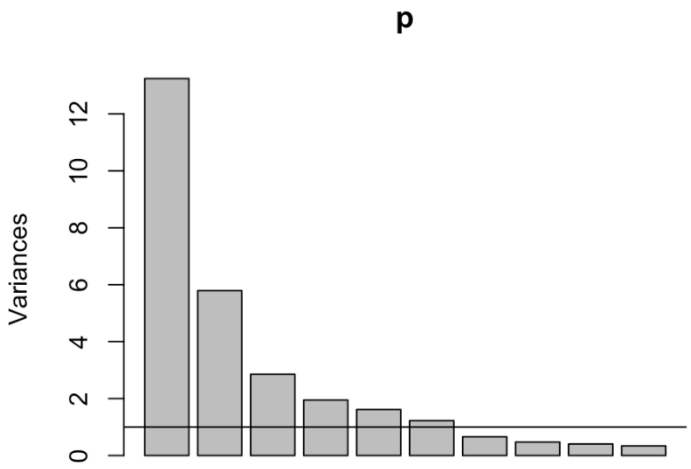


Figure 7. PCA Summary

```
Loadings:
                        RC1    RC5    RC2    RC3    RC4
radius_mean              0.952
perimeter_mean           0.950
area_mean                0.968
concavity_mean           0.653
concave.points_mean      0.805
radius_se                0.834
perimeter_se             0.821
area_se                  0.877
radius_worst             0.946
perimeter_worst          0.941
area_worst               0.949
concave.points_worst     0.684
smoothness_mean                0.780
compactness_mean               0.625
symmetry_mean                  0.674
fractal_dimension_mean
smoothness_worst               0.858
compactness_worst              0.652
symmetry_worst                 0.773
fractal_dimension_worst        0.702
compactness_se                        0.875
concavity_se                          0.875
concave.points_se                     0.713
fractal_dimension_se                  0.841
concavity_worst
texture_se                                   0.665
smoothness_se                                0.723
symmetry_se                                  0.701
texture_mean                                        0.910
texture_worst                                       0.948

                 RC1   RC5   RC2   RC3   RC4
SS loadings    10.158 5.220 5.006 2.765 2.271
Proportion Var  0.339 0.174 0.167 0.092 0.076
Cumulative Var  0.339 0.513 0.679 0.772 0.847
```

Figure 8. Parallel Analysis Scree Plot

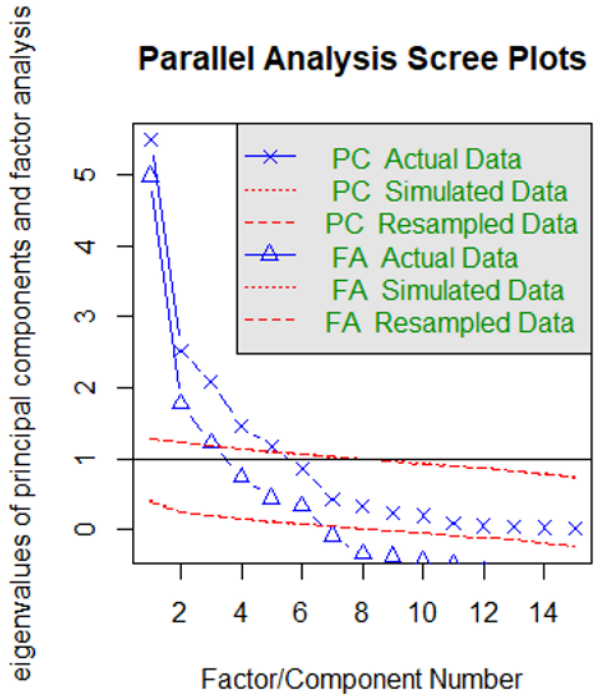


Figure 9. Factor Analysis Loadings

```
Loadings:
                  Shape Perimeter Texture
smoothness_mean   0.689
compactness_mean  0.896
symmetry_mean     0.639
compactness_se    0.760
smoothness_worst  0.583
compactness_worst 0.778
symmetry_worst    0.510
perimeter_mean            0.957
perimeter_se              0.675
perimeter_worst           0.968
texture_mean                    0.879
texture_worst                   0.966
texture_se                      0.459
smoothness_se
symmetry_se


SS loadings       3.852   2.914   2.063
Proportion Var    0.257   0.194   0.138
Cumulative Var    0.257   0.451   0.589
```

Figure 10. Variables in Final LDA Model

```
> names(LDADataset)
 [1] "diagnosis"              "texture_mean"             "smoothness_mean"        "concave.points_mean"    "symmetry_mean"          "fractal_dimension_mean"
 [7] "texture_se"             "perimeter_se"             "smoothness_se"          "compactness_se"         "concavity_se"           "concave.points_se"
[13] "symmetry_se"            "fractal_dimension_se"     "area_worst"             "smoothness_worst"       "symmetry_worst"         "fractal_dimension_worst"
```

Figure 11. LDA Plot

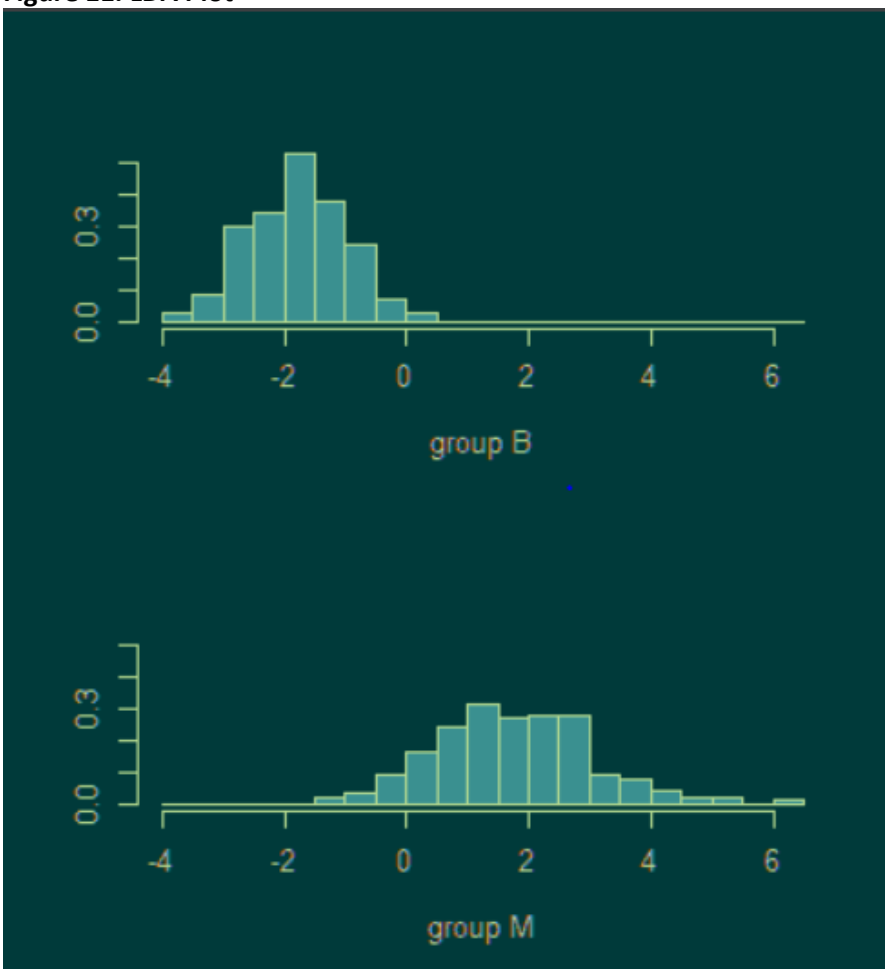


Figure 12. LDA Confusion Matrix

p	B	M
B	71	6
M	0	36

False Positive:	**0.0%**
False Negative:	**14.3%**

Figure 13. Logistic Regression Model and Odds Ratios and Confidence Interval Output

```
Call:
glm(formula = diagnosis ~ smoothness_mean + texture_se + symmetry_se +
    symmetry_worst, family = binomial(), data = clean_breastCancerData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0346  -0.8162  -0.4446   0.7503   3.4000

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.6744     0.1055  -6.392 1.63e-10 ***
smoothness_mean  0.6238     0.1207   5.167 2.38e-07 ***
texture_se       0.5066     0.1291   3.926 8.65e-05 ***
symmetry_se     -0.9693     0.1555  -6.232 4.60e-10 ***
symmetry_worst   1.4480     0.1712   8.458  < 2e-16 ***
---
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: 567.18  on 564  degrees of freedom
AIC: 577.18

Number of Fisher Scoring iterations: 5
```

Figure 14. Odds Ratios and Confidence Intervals

```
                    OR      2.5 %    97.5 %
(Intercept)     0.5094605 0.4129403 0.6247946
smoothness_mean 1.8660813 1.4804605 2.3782131
texture_se      1.6596617 1.2890201 2.1397454
symmetry_se     0.3793560 0.2771752 0.5108512
symmetry_worst  4.2546395 3.0815319 6.0341882
```

Figure 15. Logistic Regression ROC Curve

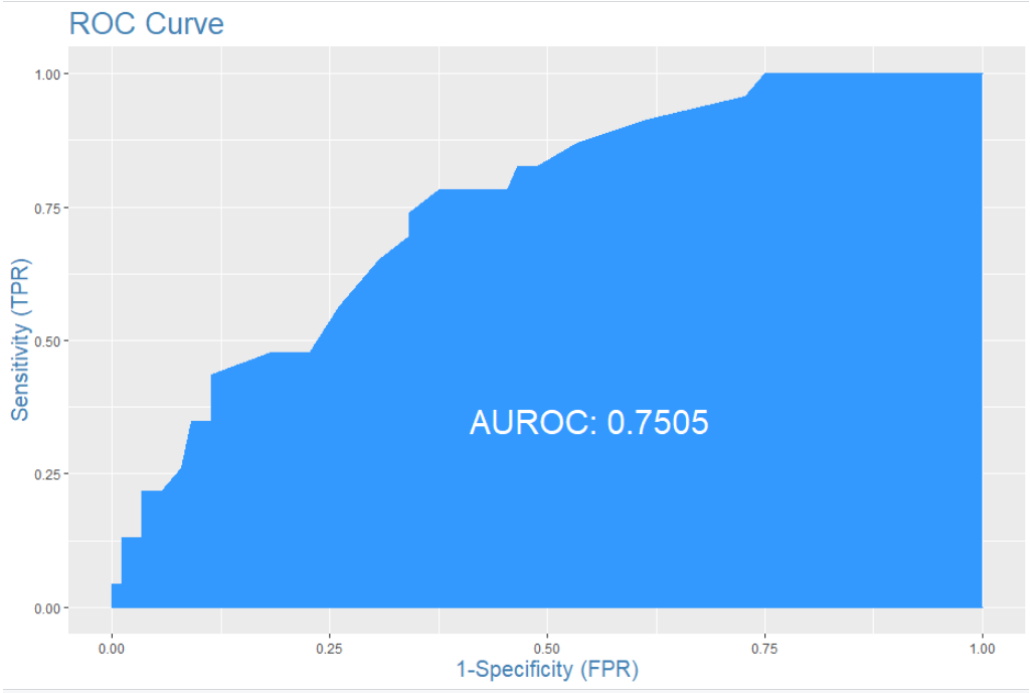


Figure 16. Confusion Matrix

Actual Values	Predicted Values	
	0 = "Benign"	1 = "Malignant"
0	85	18
1	3	5