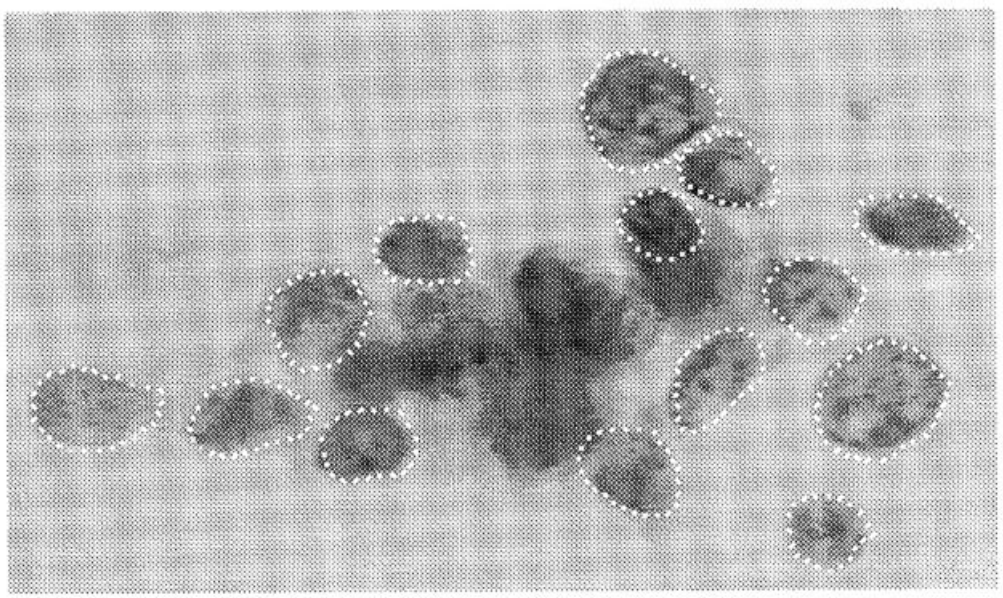
Breast Cancer Clustering

Isis Ramirez

December 18, 2019

Introduction



The accuracy of image processing techniques in determining the diagnosis for breast tumors has been an area of interest in the biomedical field. Image processing is much less invasive than biopsies, which is a common procedure used to diagnose tumors. Image processing can capture different angles and contours to extract information about a tumor's size, shape, and texture. In 1992, scientists used image processing techniques to acquire measurements for cell nuclei of breast cancer tumors. The motivation was to determine if these techniques can improve the accuracy and speed of the breast tumor diagnosis process.

The following measurements were recorded for 569 tumors:

Radius
Perimeter
Area
Compactness
Smoothness
Concavity
Concave Points
Symmetry
Fractal Dimension

Each tumor contains a cluster of cells; the mean, standard errors, and worst or highest values seen were recorded. By design, higher feature values "indicate a higher likelihood of malignancy"[1].

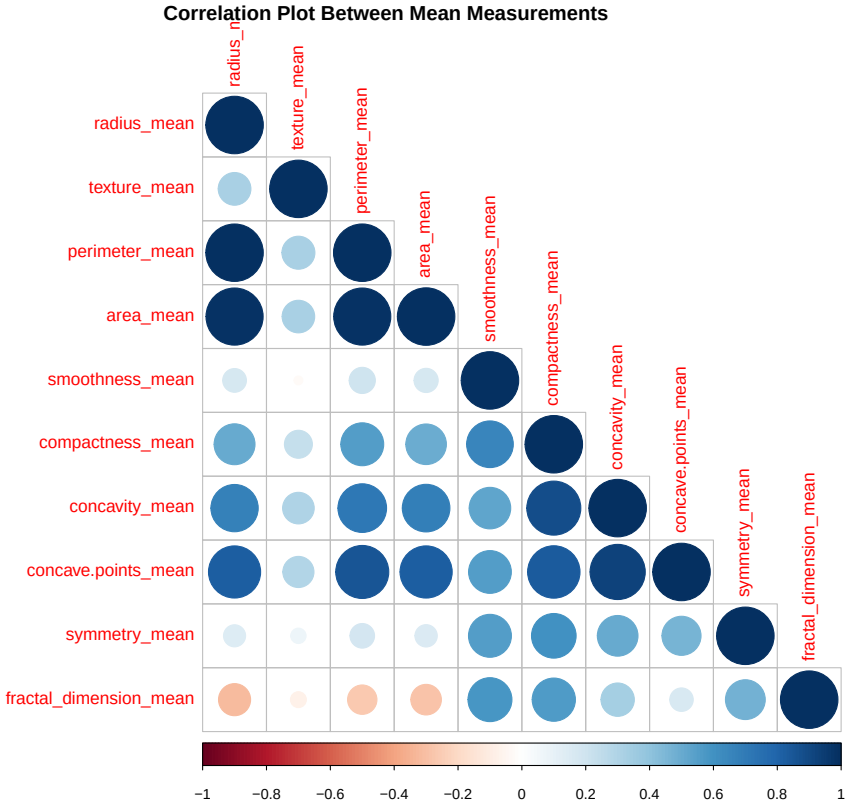
Exploratory Data Analysis

Below are the counts for benign (B) and malignant (M) tumors. About 63% of the data corresponds to benign tumors and 37% correspond to malignant tumors.

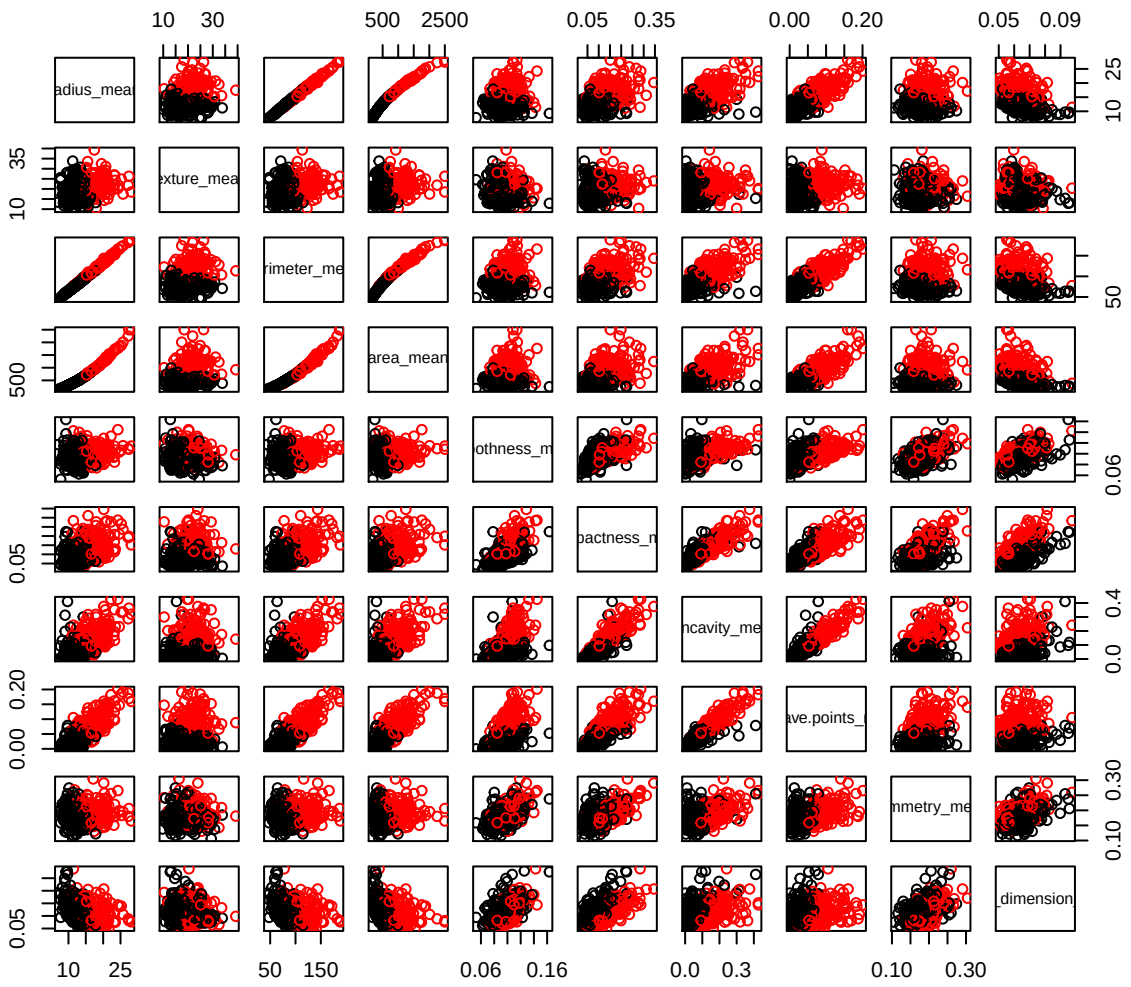
Table 1: Observation Counts

Diagnosis	Count
B	357
M	212

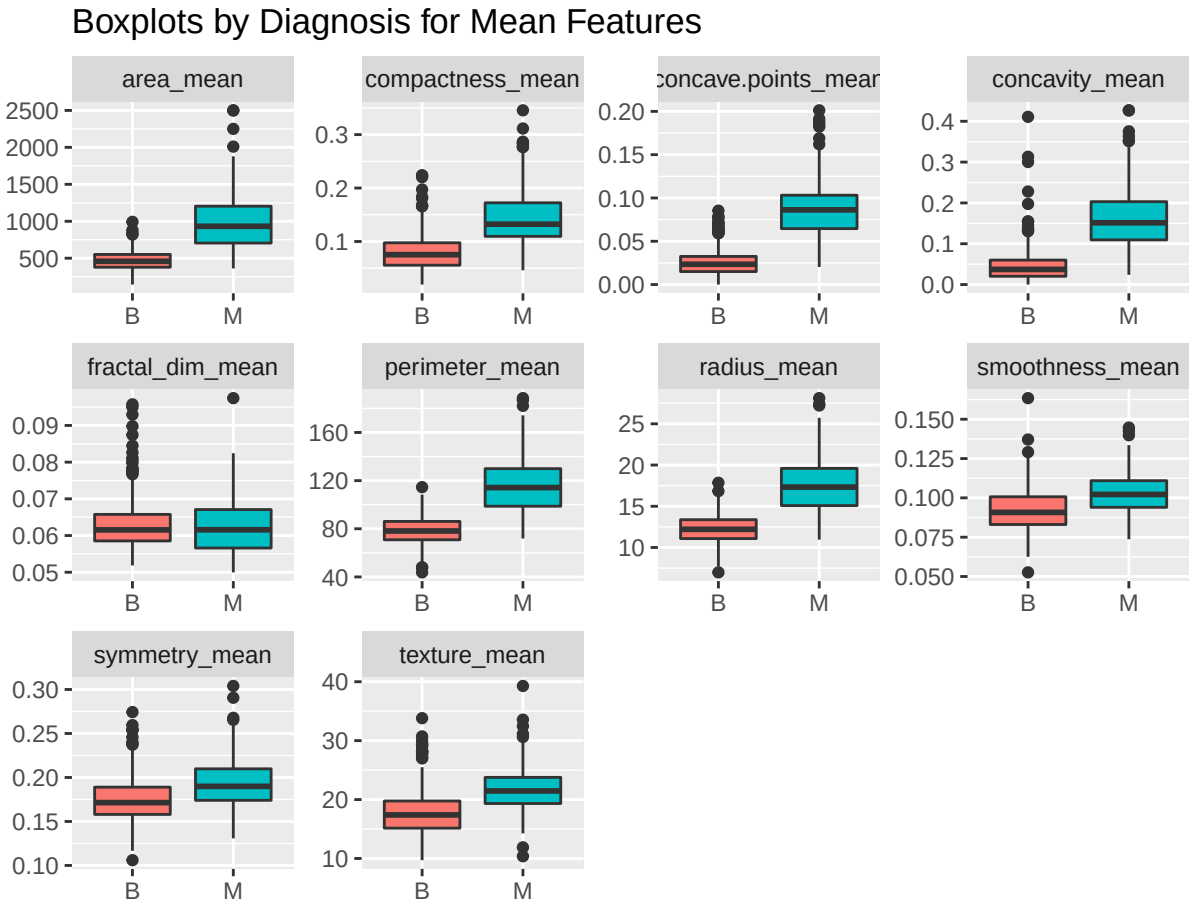
Below is the correlation plot for the mean variables. Unsurprisingly, we see strong linear correlations between radius, perimeter, and area. Similar trends in correlations are seen for standard errors and worst values; measures pertaining to the size of cells are highly correlated.



The scatterplots below for the mean values show clear separations between the 2 classes (malignant and benign) for several features; this indicates that there are variables that may be helpful in classifying tumors. We also see strong linear correlations as indicated by the correlation plot above. In the clustering analysis, we will use principal components to remove some of the redudancy that comes from the several correlated features.



Below are the boxplots for the mean features. We see that several features have distinct distributions depending on the diagnosis such as the radius, perimeter, and concavity mean parameters. The distributions for fractal dimensions seem to be the most similar between benign and malignant; this suggests that this feature may not be very helpful in classifying tumors as compared to other features. We will use principal components analysis to reduce redundancy in the data and the number of dimensions.



The clustering analysis will be done using the first principal components, which explain 72% of the variance.

	eigenvalue	variance.percent	cumulative.variance
Dim.1	1.328e+01	4.427e+01	
Dim.2	5.691e+00	1.897e+01	
Dim.3	2.818e+00	9.393e+00	
Dim.4	1.981e+00	6.602e+00	
Dim.5	1.649e+00	5.496e+00	
Dim.6	1.207e+00	4.025e+00	
Dim.7	6.752e-01	2.251e+00	
Dim.8	4.766e-01	1.589e+00	
Dim.9	4.169e-01	1.390e+00	
Dim.10	3.507e-01	1.169e+00	

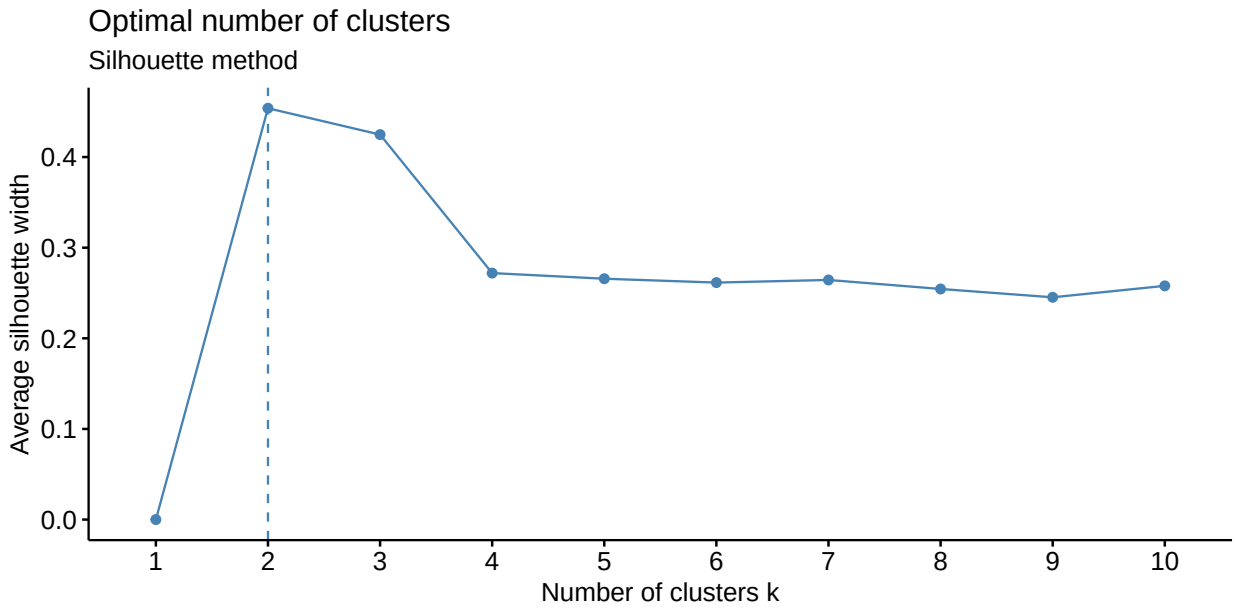
Clustering

It is vital to be able to distinguish whether patients have malignant tumors accurately and quickly. The quicker a tumor is accurately diagnosed, the quicker the patient can undergo treatment. We will test different clustering methods that could possibly help identify the diagnosis of breast tumors.

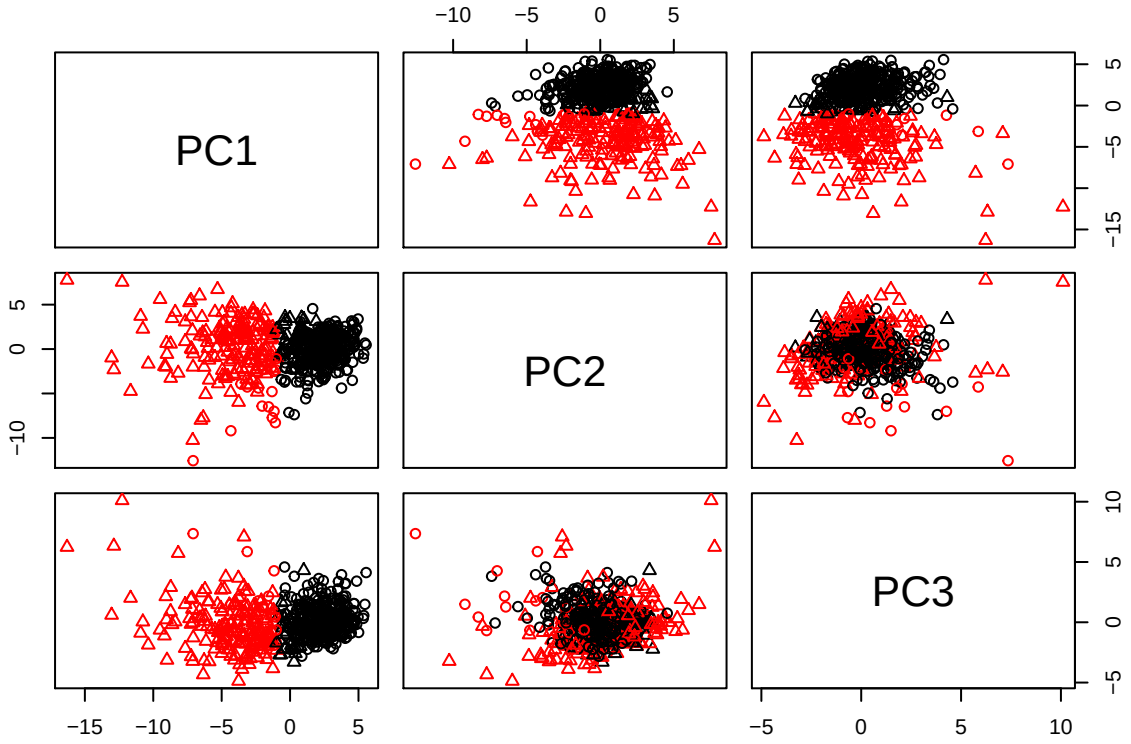
K-means

First, we will use the K-means clustering method and review it's accuracy to determine how well the model classifies benign and malignant tumors.

Using the silhoutte method, the K-means algorithm suggests k=2; this nicely aligns with the benign and malignant classes.



We see that the 2 K-means clusters are clearly distinguished in the first and second principal components. However, the clusters are not as clearly defined when looking at the first or second principal component against the third principal component; this suggests that the third principal component does not provide the K-means algorithm with much additional useful information.



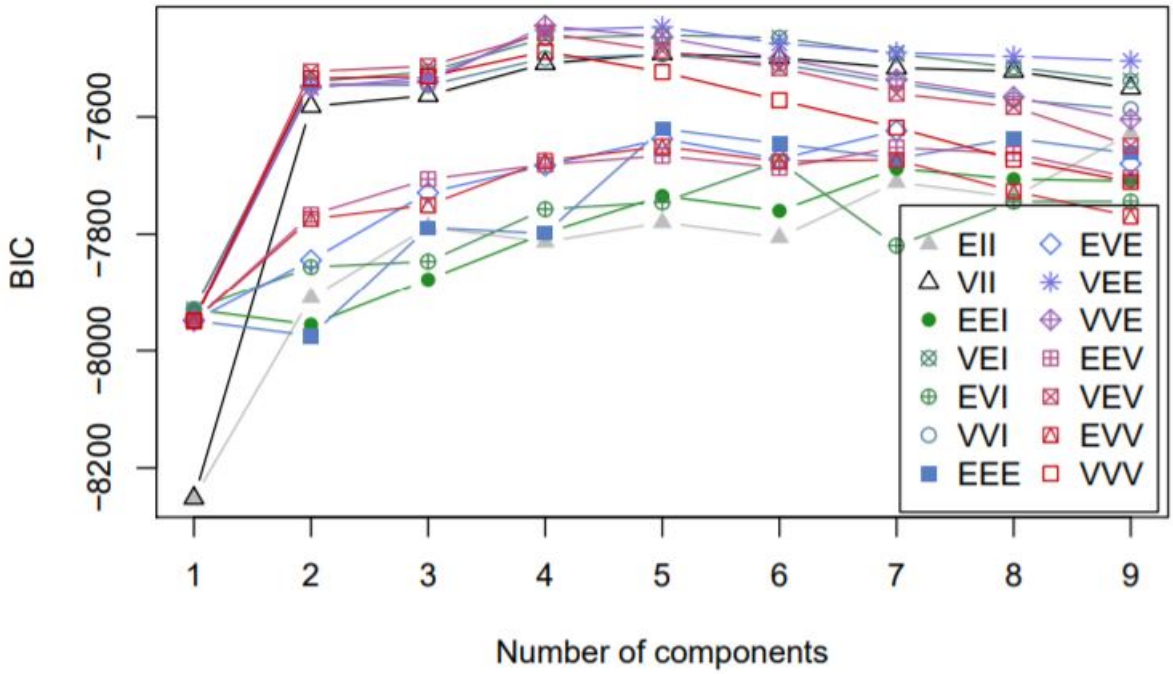
Overall, the K-means algorithm yielded an accuracy of 91%. The malignant class was harder to identify as suggested by a lower accuracy of 83% versus the 95% accuracy rate of the benign class.

Table 2: K-means Results

	B	M	Total
Cluster 1	339	36	375
Cluster 2	18	176	194
Total	357	212	569
Percent Correct	0.95	0.83	0.91

Mixture Clustering

Now, we will test the accuracy of the mixture clustering method. We see that the algorithm suggests 4 mixtures in the data using the BIC plot method.



However, after running the algorithm with 4 clusters, it seems that 4 clusters would not be ideal for diagnosis analysis. The benign cases are almost evenly split between the 3rd and 4th cluster while most of the malignant cases are in cluster 2. Cluster 1 is the smallest group, but is not clearly either benign or malignant cases.

Table 3: Mixture Clustering k= 4

	B	M
Cluster 1	15	23
Cluster 2	17	178

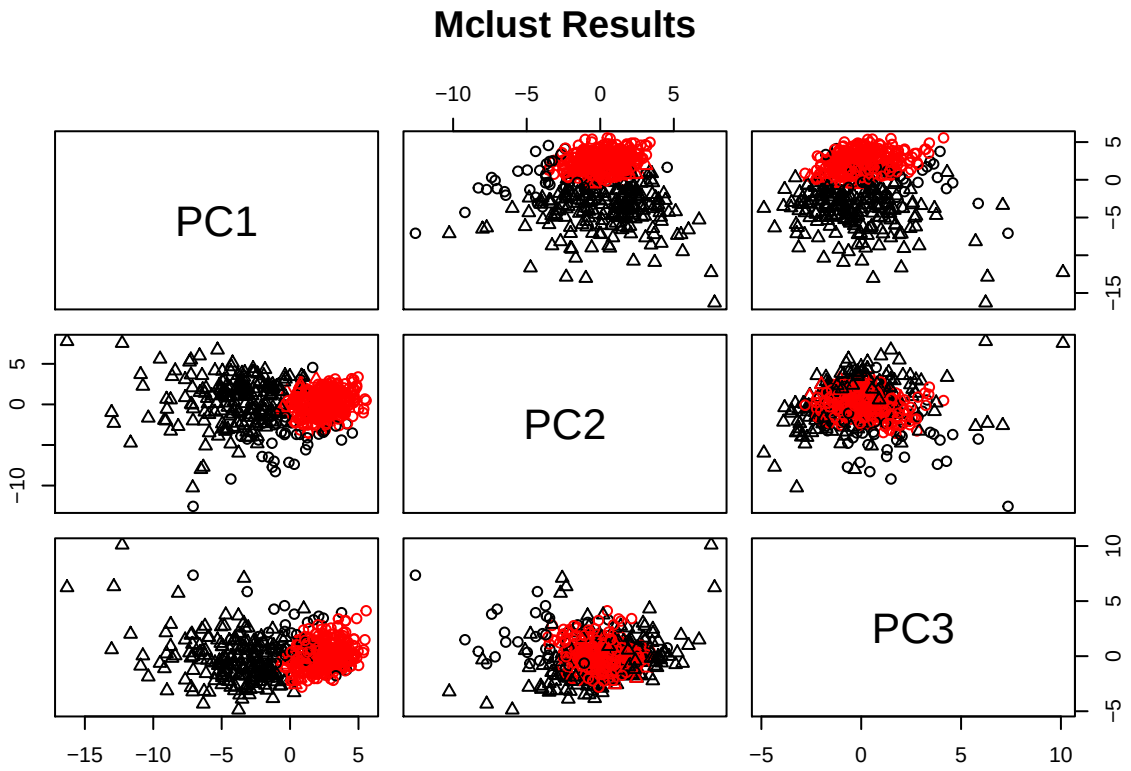
Cluster 3	132	1
Cluster 4	193	10

Using the mixture models algorithm with 2 clusters seems to be more helpful in tumor diagnosis analysis; an accuracy of 89% is achieved. Surprisingly, the malignant cases have a higher accuracy of 91% versus the 88% accuracy rate of the benign cases.

Table 4: Mixture Clustering k=2

	B	M	Total
Cluster 1	44	193	237
Cluster 2	313	19	332
Total	357	212	569
Percent Correct	0.88	0.91	0.89

Similarly to the K-means algorithm, the third principical component does not seem to be significantly helpful in the classification of the tumors.



Cases of Concern

By design, higher parameter values are tied to malignancy. Some malignant tumors are more easily identifiable since they have high or extreme values across the board. However, other malignant tumors may have parameter values that resemble that of benign tumors or a mixture of extreme and less extreme values; it

may be harder to identify these tumors as malignant. Identifying the malignant cases that most resemble benign would be helpful as physicians could proceed with much more caution when diagnosing these cases.

We calculated a total score based on the summation of the mean feature values in the raw data. Three classes will be defined as follows:
Benign
Malignant (Low Score)
Malignant (High Score)

The malignant cases with low scores are the cases that in theory would be more challenging to identify as malignant. Malignant with high scores are those cases that would be easier to identify as malignant. The threshold that divides the malignant cases into the low score and high score classes is the maximum total mean score of the benign cases.

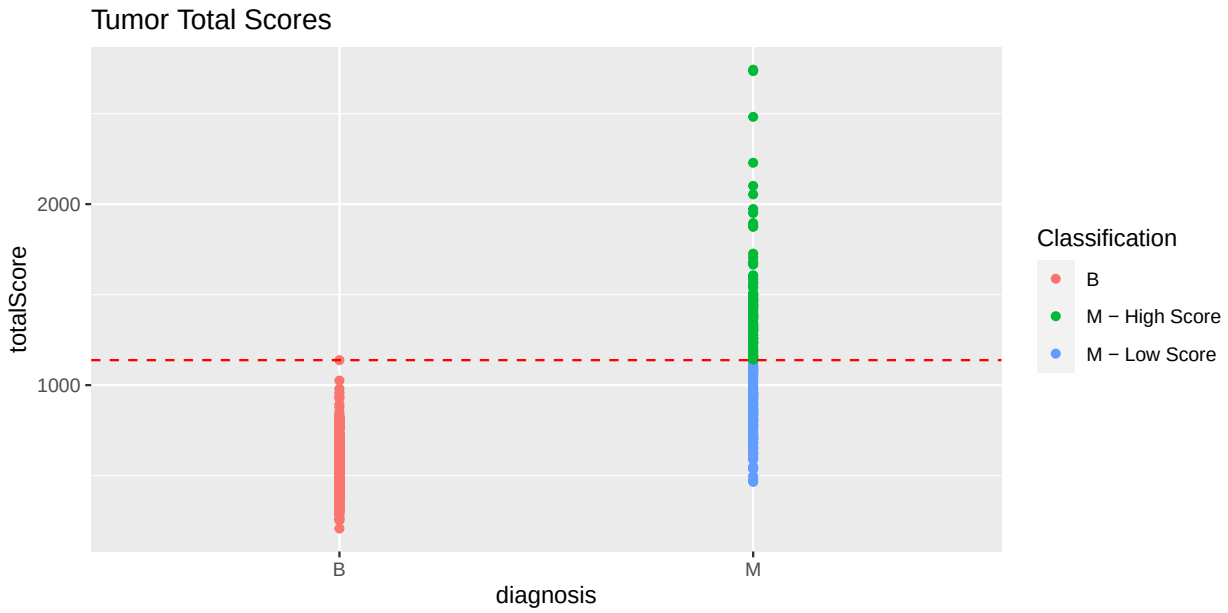


Table 5: Group Counts

Group	Count
1	357
2	113
3	99

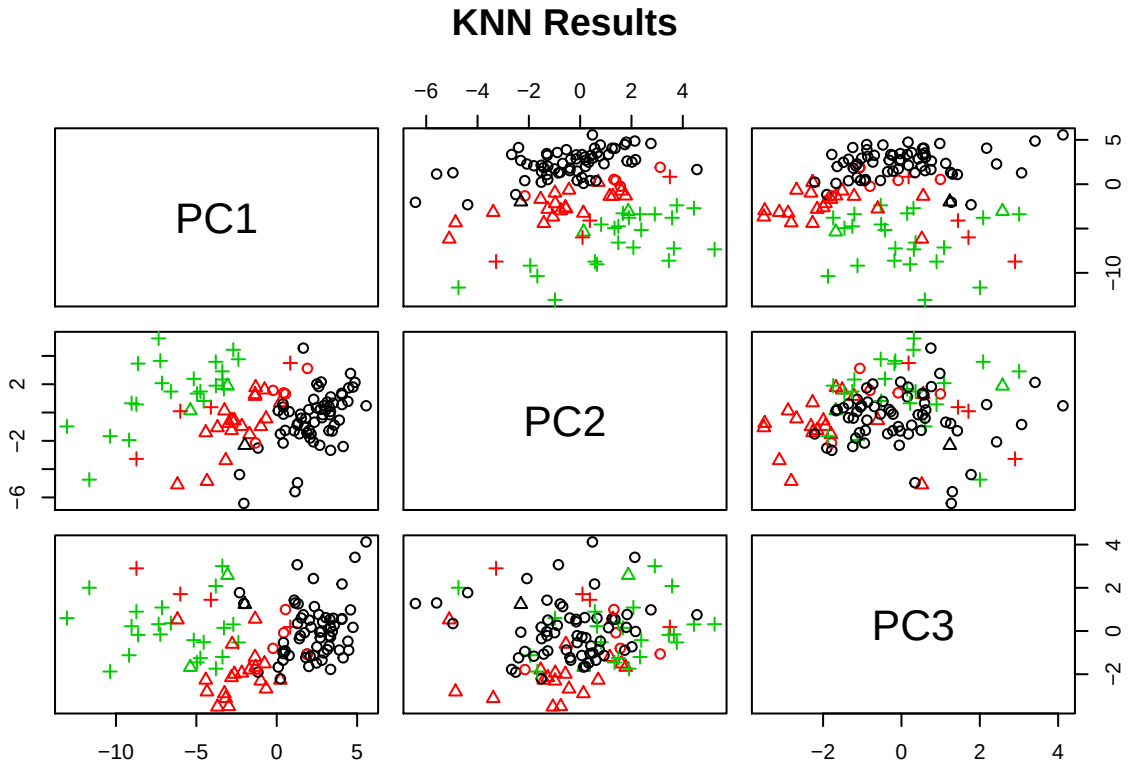
KNN (K-Nearest Neighbors)

We will test the KNN method to determine how well it can identify the 3 classes. In order to train and test the K-NN model, 80% of the data will be defined as the train set and the remaining 20% will be defined as the test set. We will use k=3 and euclidean distances.

Table 6: KNN Results

	B	M - (LS)	M - (HS)	Total
Group 1	61	5	0	66
Group 2	1	19	2	22

Group 3	0	4	22	26
Total	62	28	24	114
Percent Correct	0.98	0.68	0.92	0.89



The K-NN method has an 89% accuracy rate. Unsurprisingly, the malignant cases with lower scores had the lowest accuracy of 68%; this aligns with the idea that these cases are at a higher risk of being misdiagnosed as benign and these cases should be proceeded with extreme caution during the diagnosis process. The malignant cases that seem to be more easily identifible had an accuracy of 92% while the benign cases had the highest accuracy at 98%. We also see the same theme that has appeared in the other methods; the third principal component might not introduce new useful information.

Conclusions

The K-means and Mclust methods have high accuracies when classifying benign and malignant cases. When the malignant cases are broken down into cases that are easier and harder to identify, the accuracy unsuprisingly decreased; however the k-nn method still had a high overall accuracy. The high accuracies achieved by these methods should be met with much caution and skepticism as the data only contains 569 data points. The next steps should include the testing of these methods on a larger set of data to eliminate the possibility of overfitting. A further modification to these methods would be to eliminate the third principal component as the models suggested that it was unnecessary. Additionally, it would be interesting to run the same clustering models on new data as image processing techniques have had over 27 years to advance since 1992.

References

1.Street, N., Wolberg, W. H., & Mangasarian, O. L. (1992). Nuclear feature extraction for breast tumor diagnosis. Madison, WI: University of Wisconsin-Madison, Computer Sciences Dept.Uci. Retrieved from https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis.
2. Breast Cancer Wisconsin (Diagnostic) Data Set. Retrieved from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.