Breast Cancer Analysis

Utsav Patel

28 April 2019

Abstract

About 1 in 8 U.S. women (about 12%) will develop invasive breast cancer over the course of her lifetime. There are two types of tumors: Benign and Malignant. Malignant tumor is cancerous and is dangerous if it is not treated in the early stages. It is also very expensive in the United States for the cancer diagnosis and treatment. Hence, in the search to find a cheaper diagnosis technique, we propose a few models like logistic regression, linear classification on PCA(Principal Component Analysis) and Random Forest, which can help diagnose breast cancer as benign or malignant.

Introduction

Breast cancer is a type of cancer that develops from breast tissue and is often associated by a lump in the breast, change in breast shape, development of red and patchy skin, or fluid emanating from the nipple. The causes for breast cancer have not been fully understood till date. There are some genetic factors, and some environmental factors associated with its development. Breast cancer is preliminarily detected by a mammogram exam and confirmed by a biopsy.

There is no single measurement that can be used to determine whether a given sample is benign or malignant. In 2019, an estimated 268,600 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 62,930 new cases of non-invasive (in situ) breast cancer. About 2,670 new cases of invasive breast cancer are expected to be diagnosed in men in 2019. A lifetime risk of breast cancer for man is about 1 in 883. Breast cancer incidence rates in the U.S. began decreasing in the year 2000, after increasing for the previous two decades. They dropped by 7% from 2002 to 2003 alone. One theory is that this decrease was partially due to the reduced use of hormone replacement therapy (HRT) by women after the results of a large study called the Women's Health Initiative were published in 2002. These results suggested a connection between HRT and increased breast cancer risk.

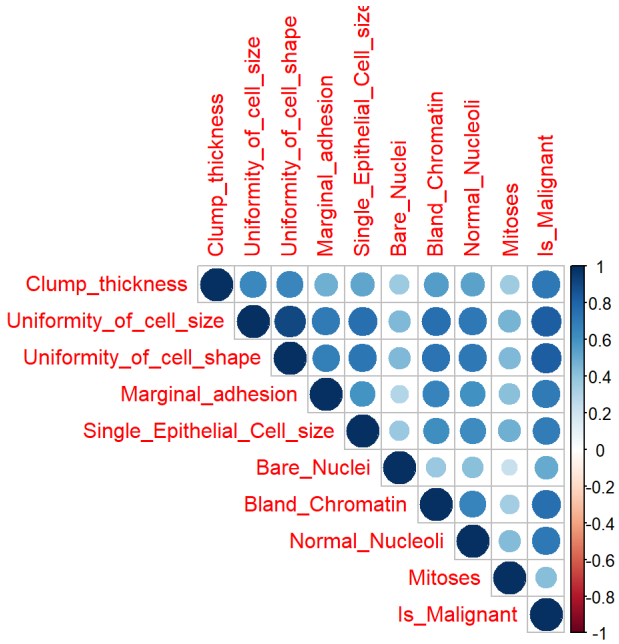
There can be cancer because of 2 types of tumor: Benign and Malignant. - Benign tumor are non-malignant/non-cancerous tumor. A benign tumor is usually localized, and does not spread to other parts of the body. Most benign tumor respond well to treatment. However, if left untreated, some benign tumor can grow large and lead to serious disease because of their size. Benign tumor can also mimic malignant tumor, and so for this reason are sometimes treated. - Malignant tumor are cancerous growths. They are much dangerous than the benign tumor. They usually grow very rapidly. They are often resistant to treatment, may spread to other parts of the body and they sometimes recur after removal.

Dataset Characteristics

We have 9 attributes which can help us detect whether the tumor is benign or malignant. Let's see what are those.

- **Clump Thickness**: This is used to assess if cells are mono-layered or multi-layered. Benign cells tend to be grouped in mono-layers, while cancerous cells are often grouped in multi-layer.
- **Uniformity of Cell Size**: It is used to evaluate the consistency in the size of cells in the sample. Cancer cells tend to vary in size. That is why this parameter is very valuable in determining whether the cells are cancerous or not.
- **Uniformity of Cell Shape**: It is used to estimate the equality of cell shapes and identifies marginal variances because cancer cells tend to vary in shape.
- **Marginal Adhesion**: Normal cells tend to stick together. Cancer cells tend to loose this ability. So loss of adhesion is a sign of malignancy.
- **Single Epithelial Cell Size**: It is related to the uniformity. Epithelial cells that are significantly enlarged may be a malignant cell.
- **Bare Nuclei**: This is a term used for nuclei that is not surrounded by cytoplasm. Those are typically seen in benign tumor.
- **Bland Chromatin**: Describes a uniform texture of the nucleus seen in benign cell. In cancer cell, the chromatin tends to be coarser.
- **Normal Nucleoli**: Nucleoli are small structures seen in the nucleus. In normal cell the nucleolus is usually very small if visible at all. In cancer cell the nucleoli become much more prominent, and sometimes there are more of them.
- **Mitoses**: It is an estimate of the number of mitosis that has taken place. Larger the value, greater is the chance of malignancy

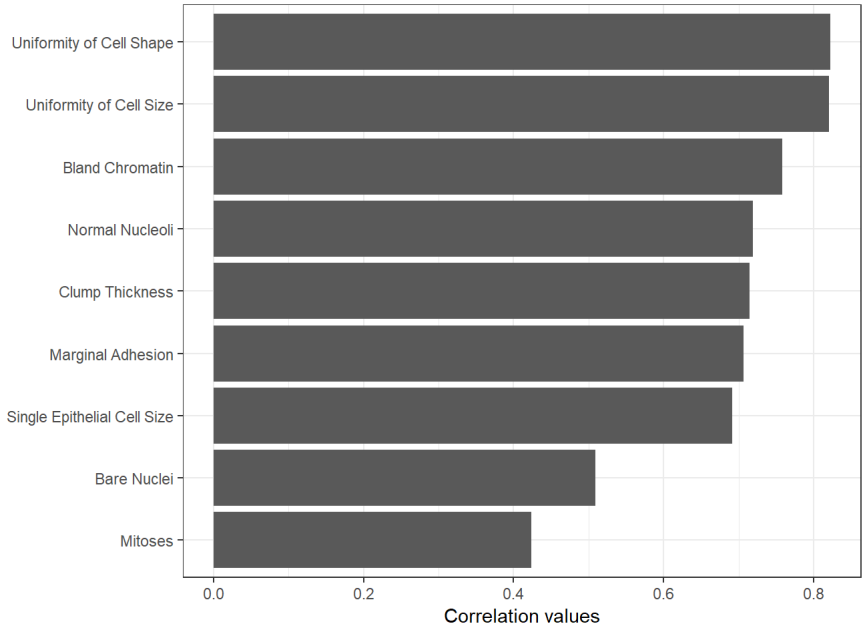
Correlation heatmap of our data



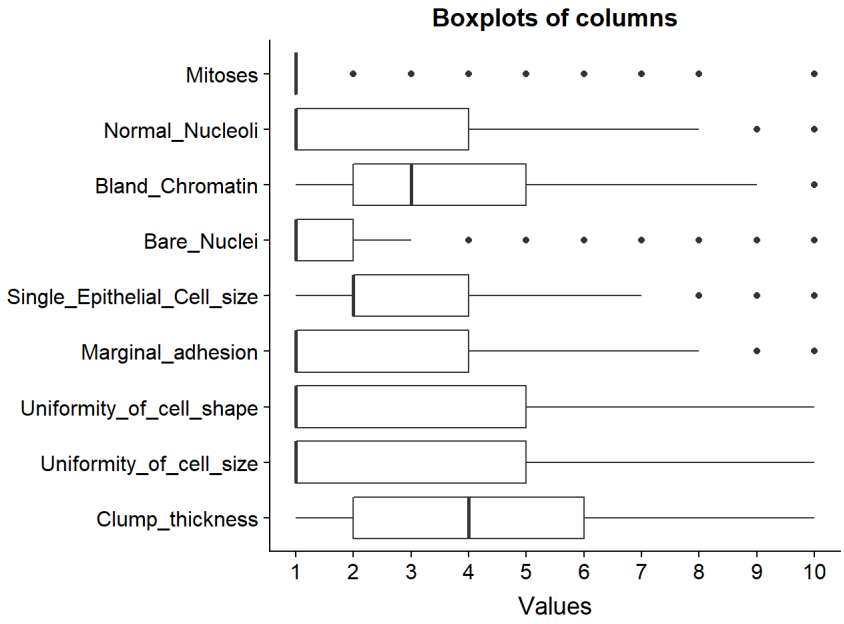
The above heatmap shows the correlation between variables. Darker the blue color higher positive correlation, darker the red color, higher negative correaltion. As we can see that all the variables are postively corelated with each other and with target variable. Size of the dot is proportional to absolute value of correlation.

Correlation of each variable with respect to target variable:

To see how each variable is correlated with target variable, we have provided a clearer picture of the correlation values in a sorted manner.



Boxplot of our data

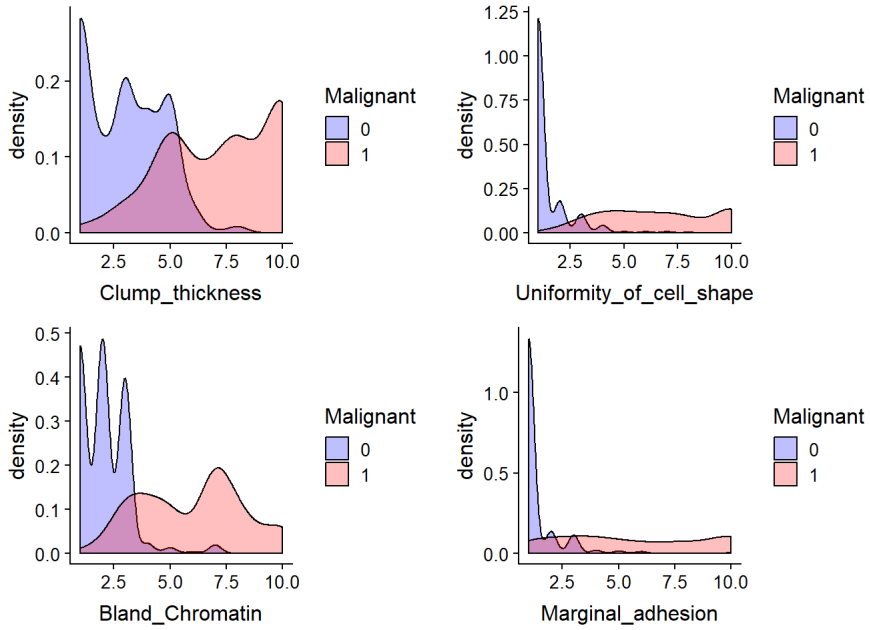


Just to get a range of the variable values, boxplot has been plotted. For example, clump thickness ranges from 2 to 6 and has a median close to 4. In case of uniformity of cell size, uniformity of cell shape value starts from 1 and that itself is median that means at least 50% of the data has value for these variable 1. In case of mitoses, it is clear that data is congested at value 1 and there is not much variation except few outliers. So, how mitoses affect breast cancer is difficult to analyze from this dataset.

Logistic Regression with Backward elimination technique

We have used Logistic Regression with backward elimination technique to reduce the number of variables. The process we followed was to check what impacts the residual change and eliminated the variables accordingly. We removed one variable at a time. Each time we removed the variable which had highest p value till we found all the variable with p value less than 0.01 as we had set significance level as 0.01. For example, when we did not eliminate any variable the model summary showed coefficient of p value for Uniformity_of_cell_size was highest with a value of 0.773024. So we eliminated Uniformity_of_cell_size. Similarly, one by one we removed 3 more variables. Finally, we were down to 4 predictors from 9. Clump thickness, Uniformity of cell shape, Bland Chromatin and Marginal Adhesion are our final 4 variables for glm model. Let us analyze their individual impact on target variable.

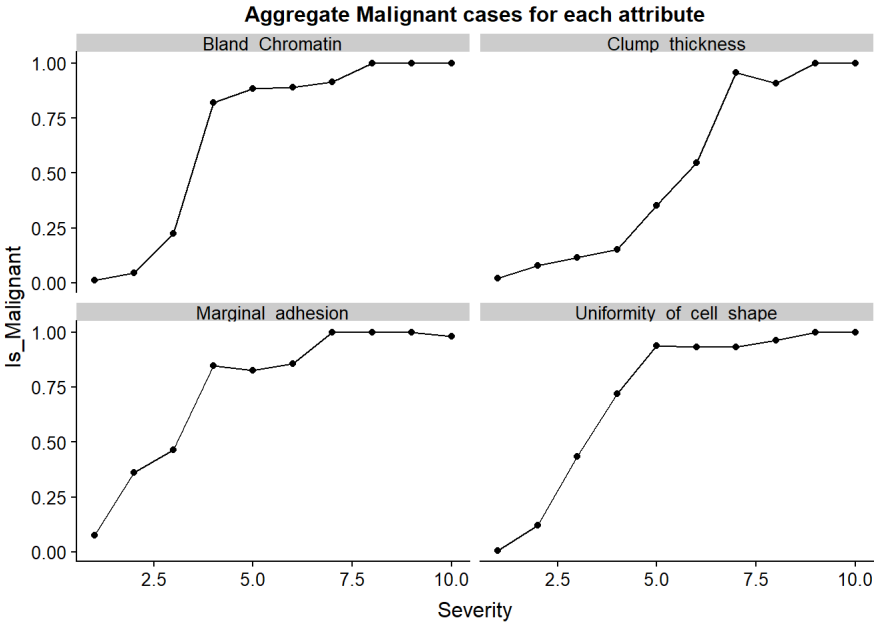
Density Plot



Clump thickness: Values for benign cells tends to be on lower end and for malignant cells values tends to higher in general

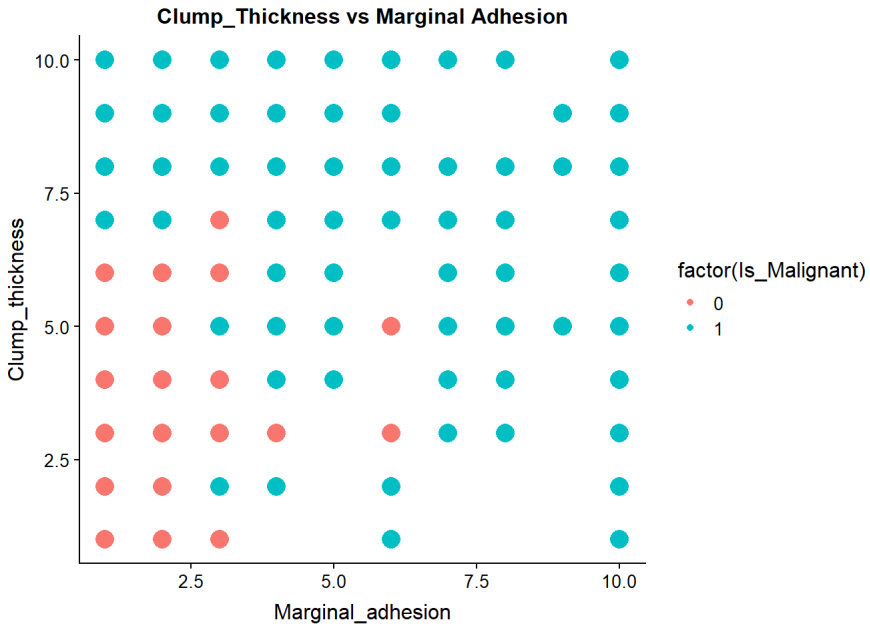
Bland chromatin, Uniformity of cell shape, Marginal adhesion: Values are highly dense at the lower end of the range and does not vary much for benign cells where in case of malignant cells, values have higher variance and are roughly spread across the entire range.

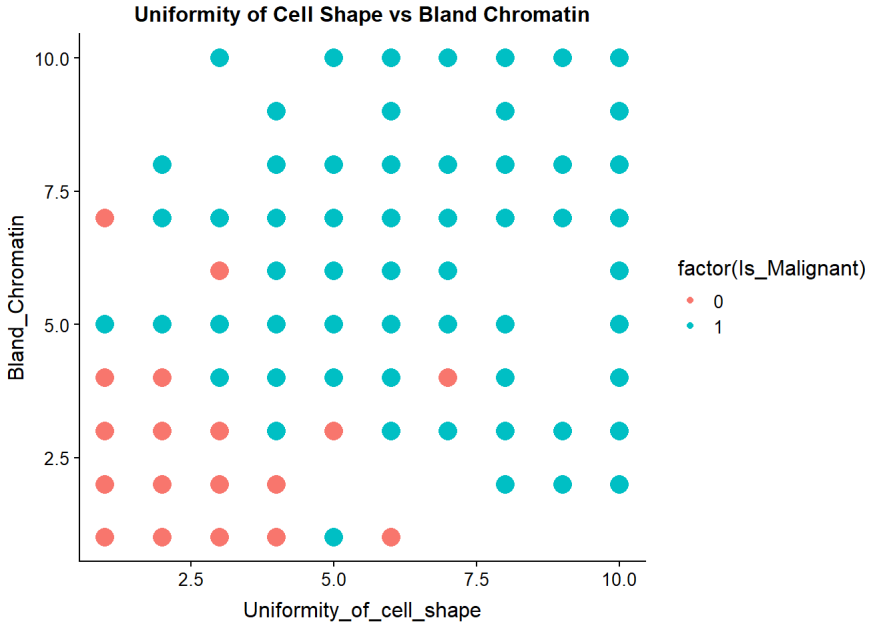
How probability of cancer vary with each variable:



The graph represents how the probability of cell being malignant varies as the value for each of the variable increases. In all the 4 variables, as the values increases probability of cell being malignant increases. This corroborate the insight provided by density plot above where we saw that benignant cells value tend to be on the lower end.

Grid based classification using 2 variables





To see how the class distribution varies with 2 variables we have plotted these graphs. We see that as the value for any variable increases from somewhere around 6.5 to 7.5, it is certain that cells are malignant. For values lower than 3 it is safe as they are benignant cells but for values between 3 and 6.5 it is uncertain.

Fitting the glm model:

To increase the complexity of the model and eventually accuracy of the prediction, we have decided to include the combinations of two variable as well. We have 4 variables so to choose 2 from 4 we had total 6 options and out of these 6 we have picked 3 combination based on their interaction plots and later through hit and trial, it was verified to perform the best.

```
##
## Call:
## glm(formula = Is_Malignant ~ Clump_thickness + Marginal_adhesion +
##     Uniformity_of_cell_shape + Bland_Chromatin + Clump_thickness:Marginal_adhesion +
##     Clump_thickness:Uniformity_of_cell_shape + Clump_thickness:Bland_Chromatin,
##     family = "binomial", data = bc)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.81582  -0.08713  -0.02247   0.06079   3.00260
##
## Coefficients:
##                                       Estimate Std. Error z value
## (Intercept)                          -14.92552    2.40593  -6.204
## Clump_thickness                        1.58569    0.38883   4.078
## Marginal_adhesion                      0.94342    0.31700   2.976
## Uniformity_of_cell_shape               1.25862    0.43419   2.899
## Bland_Chromatin                        1.08953    0.38048   2.864
## Clump_thickness:Marginal_adhesion     -0.10144    0.04981  -2.037
## Clump_thickness:Uniformity_of_cell_shape  -0.09581    0.06822  -1.404
## Clump_thickness:Bland_Chromatin       -0.07267    0.06364  -1.142
##                                      Pr(>|z|)
## (Intercept)                          5.52e-10 ***
## Clump_thickness                      4.54e-05 ***
## Marginal_adhesion                     0.00292 **
## Uniformity_of_cell_shape              0.00375 **
## Bland_Chromatin                       0.00419 **
## Clump_thickness:Marginal_adhesion     0.04168 *
## Clump_thickness:Uniformity_of_cell_shape  0.16020
## Clump_thickness:Bland_Chromatin       0.25353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 121.97  on 675  degrees of freedom
## AIC: 137.97
##
## Number of Fisher Scoring iterations: 8
```

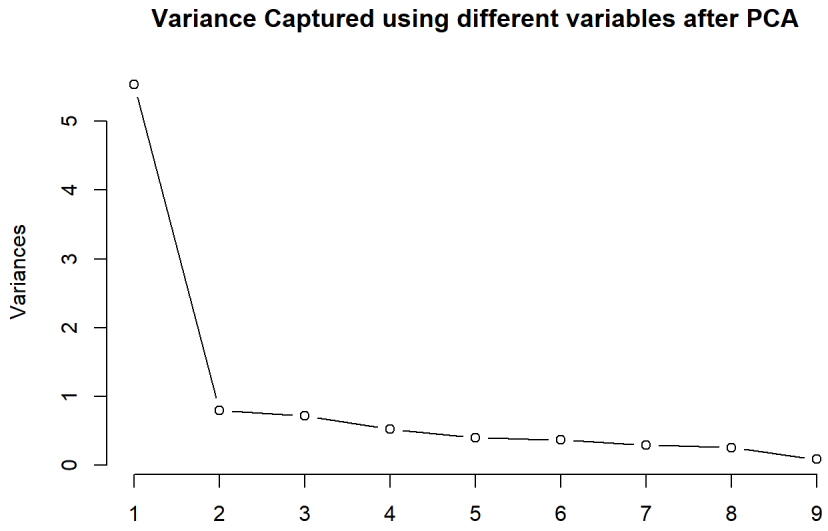
Accuracy on test set:

```
## [1] "Accuracy = 99.2481203007519 %"
```

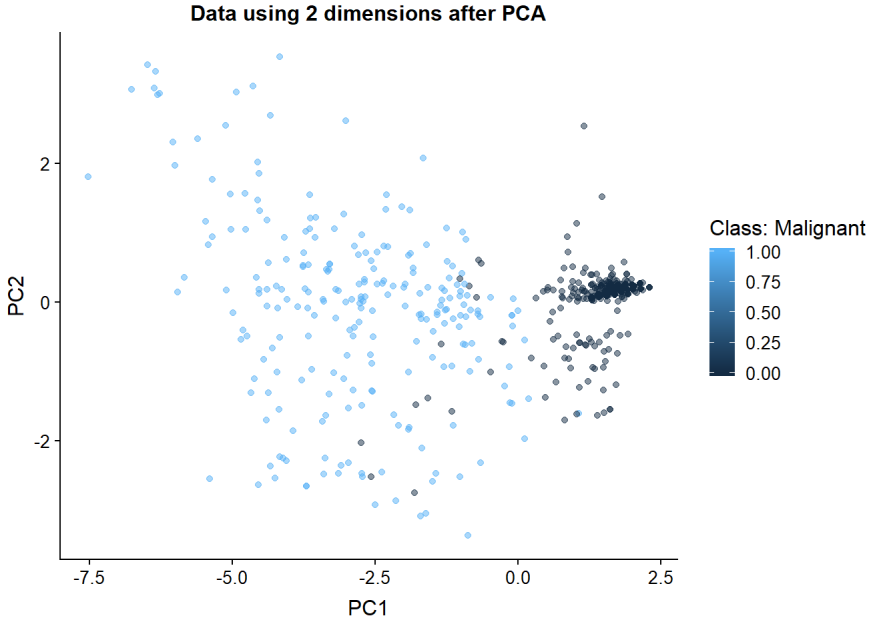
Dimensionality Reduction: PCA

- If we have a large amount of data, we would like to avoid the curse of dimensionality and would like to reduce the time and space required.
- We tried to transform the dimensions of the data which captures the maximum variance using PCA and observed how much variance we can capture using PCA.
- This would let us easily visualize the data.

- It would also help us remove the multi-collinearity, which would help us ignore redundant features.



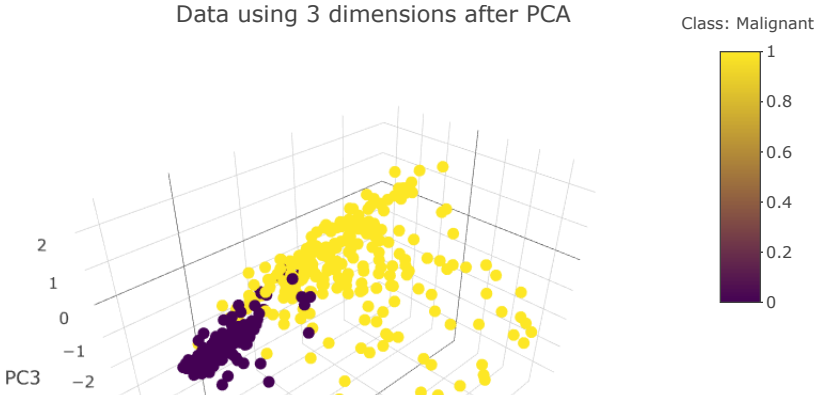
- We can see that, after applying PCA, the maximum variance is captured by the first variable. The second and the third variable also contributes to the variance a little. So now we can check how the data looks using the most important 2 and 3 dimensions.

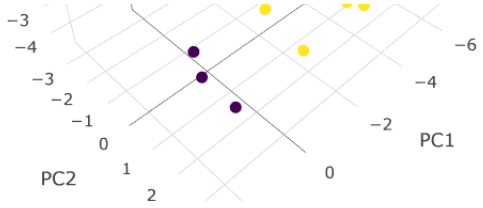


- We can see that the data seems quite seperable using just the 2 dimensions. We will just check how it looks using 3 dimensions.

```
## Warning: package 'plotly' was built under R version 3.5.2
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

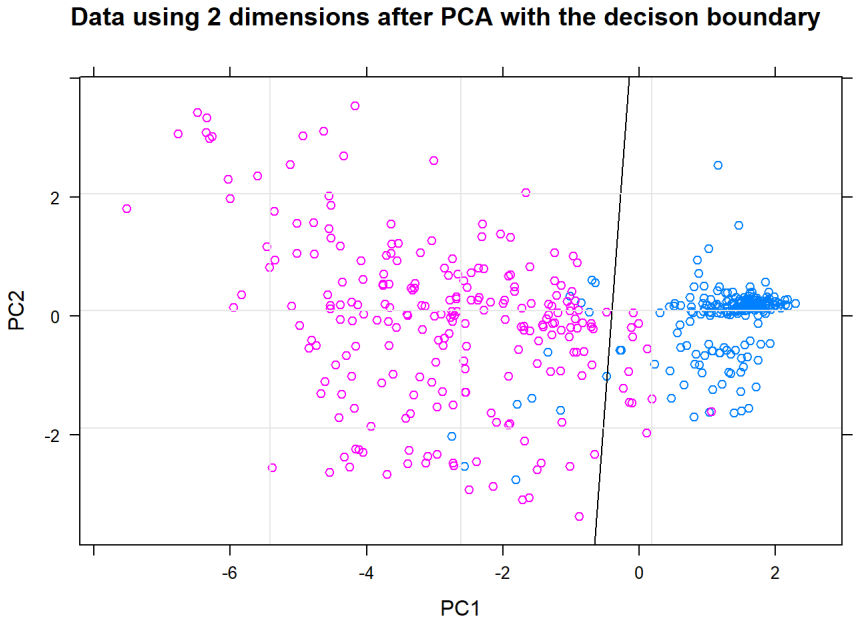




- Again, we see that the data looks pretty seperable using 3 dimensions as well.
- We can observe that using just the 2 dimensions, we get very seperated results. Hence, we proceed with applying a linear model to classify the malign and benign tumors on the 2 PCA reduced dimensions.

Decision boundary

Let us draw a decision boundary on the data from the 2 dimensional PCA.



- The decision boundary which is drawn separated the data pretty well enough. Now, we calculate the accuracy of the data separated by the decision boundary.

```
## [1] "Accuracy 0.970717423133236"
```

Results

- We observe that the accuracy of around 97.07% is a good prediction using data with such low dimensions.
- But, here the problem is we cannot interpret the dimensions which we get from PCA. Hence, we need something that can give us some high accuracy with interpretable features.
- Therefore, we next tried random forest, whose result features would be interpretable and more practical.

Random Forest

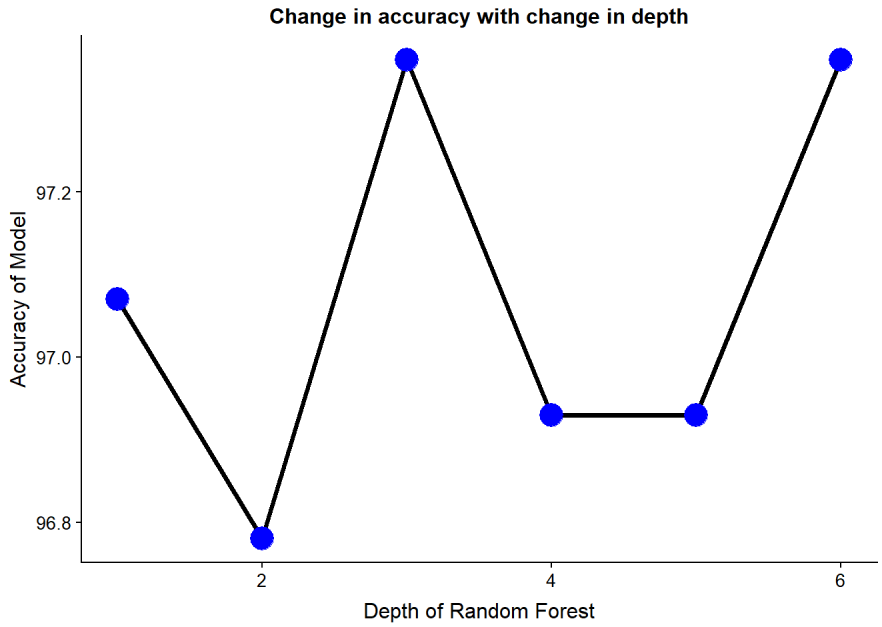
- As discussed above PCA gives us good results but is less interpretable. Hence, we will use a random forest model which is developed by aggregating trees.
- The advantage of random forest is that its results are very interpretable and it also avoids overfitting.
- We can perform feature selection based on their importance.
- As an extension, there are many other health test which can be done which can be useful for breast cancer diagnosis, due to which we can have more number of features that can be handled using random forest.

We experiment the model with different depths and check which one performs the best.

Accuracy for models with depths: 1,2,3,4,5,6 =

```
## [1] 97.07 96.78 97.36 96.93 96.93 97.36
```

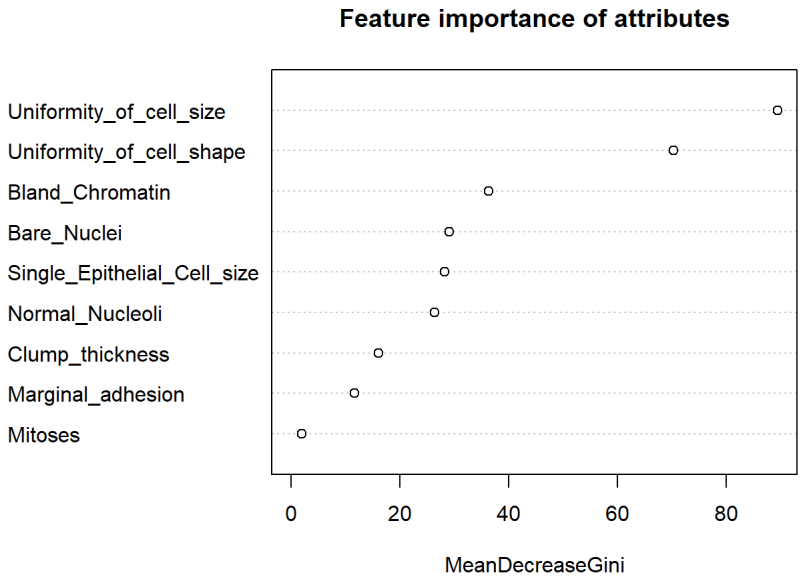
- Now lets visualise the accuracy values.



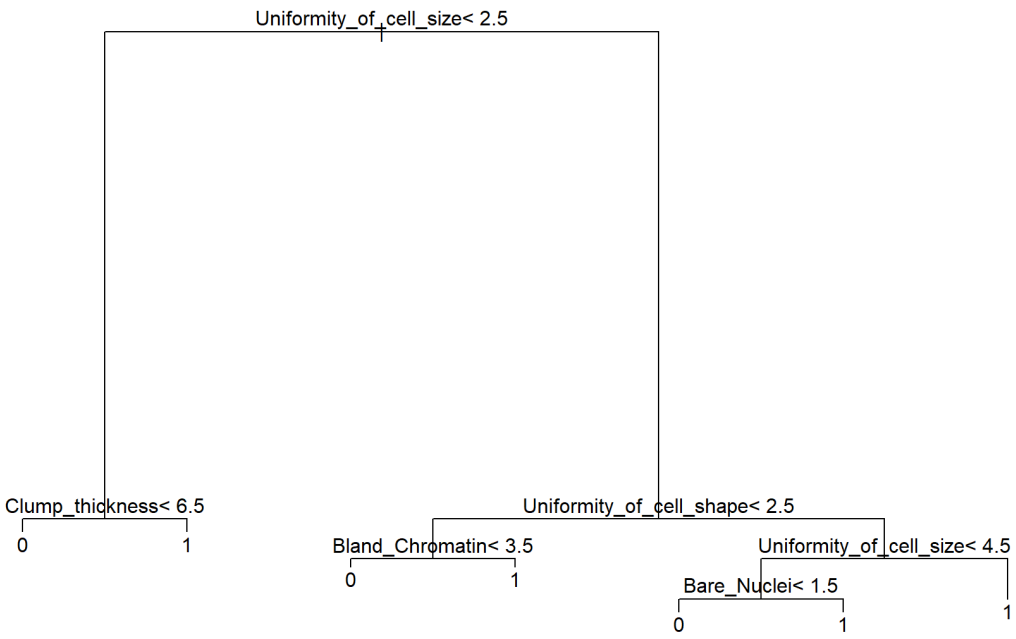
For depth = 1, it gives the accuracy of 97.07 and after increasing the depth it does not help much with the accuracy. At max we reached to 97.37% of accuracy at depth =3 and depth =6 and this change is not sufficient enough to go for higher depth and unnecessarily overfit our model.

Gini importance of features in Random Forest:

Gini Importance or Mean Decrease in Impurity (MDI) calculates each feature importance as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits. We have plotted relative feature importance of the model that we have created.



- We can observe that the uniformity of cell size and cell shape has the higher importance compared to remaining features. Further, lets create a tree from which we can interpret our results very easily by just looking at the symptoms of a patient.



- As we can see, now the results are very interpretable as we can just measure each of the attributes and we can decide with good confidence, whether a person has a malignant tumor or a benign one.
- For example, if a patient has the uniformity of cell size less than 2.5, uniformity of cell shape greater than 2.5 and Bland chromatin greater than 3.5, then the person is most likely to have a malignant tumor.

Accuracy:

We tried 3 different models: GLM, GLM with PCA and Random Forest. Accuracy of those models are 99.4% , 97% and 97.07% respectively. Random Forest was chosen to be the final model because it gave us pretty high accuracy just using DEPTH=1 and estimators=100.

Conclusion:

We have done detailed analysis on which features are useful in predicting malignancy of breast cancer. We also gave a simple dendogram chart prepared from Random Forest that can help anyone to predict whether a tumor is benign or malignant from 9 given features. From our derived conclusions we can conclude that there are mainly 1 or 2 features of tumor cell which are most important in order to predict malignancy of tumor. Also, it is much important to identify cancerous nature of tumor cells in early stages as nearly 86% of patients could be cured if tumor is treated early. As a future work, we can test on more practical data to get conclusions on robustness of our model and improve our findings.

Acknowledgement:

We thank Prof. Brad Luen and Seiji Sloan for guiding us throughout the project. The inspiration to try Random Forest Model and make simple dendogram like chart came from suggestions by Prof. Brad Luen during the presentations and we really thank him to guide us throughout the semester.

References

- https://www.researchgate.net/profile/Akash_Nag2/publication/325868350_Identifying_Patients_at_Risk_of_Breast_Cancer_through_Decision_Trees/links/5b2
Patients-at-Risk-of-Breast-Cancer-through-Decision-Trees.pdf?origin=publication_detail
(https://www.researchgate.net/profile/Akash_Nag2/publication/325868350_Identifying_Patients_at_Risk_of_Breast_Cancer_through_Decision_Trees/links/5b
Patients-at-Risk-of-Breast-Cancer-through-Decision-Trees.pdf?origin=publication_detail)
- https://www.breastcancer.org/symptoms/understand_bc/statistics (https://www.breastcancer.org/symptoms/understand_bc/statistics)
- http://pathology.jhu.edu/pc/BasicTypes1.php (http://pathology.jhu.edu/pc/BasicTypes1.php)