

# Analysis of Breast Cancer

Prepared by

**Team: Section B**

Yeji (Amy) Lee	63843416
Teng-Yun (Jacob) Chung	92443989
Zhuoer (Stella) Yang	71165437
Kuang (Steven) Li	46448655
Tanvir Brar	36241934

December 2019

## **Introduction**

According to the Susan G. Komen for the Cure, breast cancer is the most common cancer among women in the U.S, with 3.5 million women in the U.S. having a history of breast cancer. Women in the U.S. have a 1 in 8 lifetime risk of being diagnosed with breast cancer and every two minutes, one case of breast cancer is diagnosed in women. Breast cancer is a complex and heterogeneous disease due to its diverse morphological features, as well as different clinical outcomes. As a result, breast cancer patients may respond to different therapeutic options. Currently, difficulties in recognizing breast cancer types lead to inefficient treatments. Generally, there are two types of breast cancer types, known as malignant and benign. Therefore it is necessary to devise a clinically meaningful classification of the disease that can accurately classify breast cancer tissues into relevant classes.

We are proposing this topic because it is a very common cancer among women and it can happen to anyone close to us. According to peer-reviewed healthcare journal “Health Affairs”, national expenditure for false-positive mammograms and breast cancer overdiagnosis is estimated at \$4 billion a year. In addition to this, from the Breast Cancer Detection Demonstration Project (BCDDP), the false-negative rate of mammography is approximately 8-10% in the U.S.

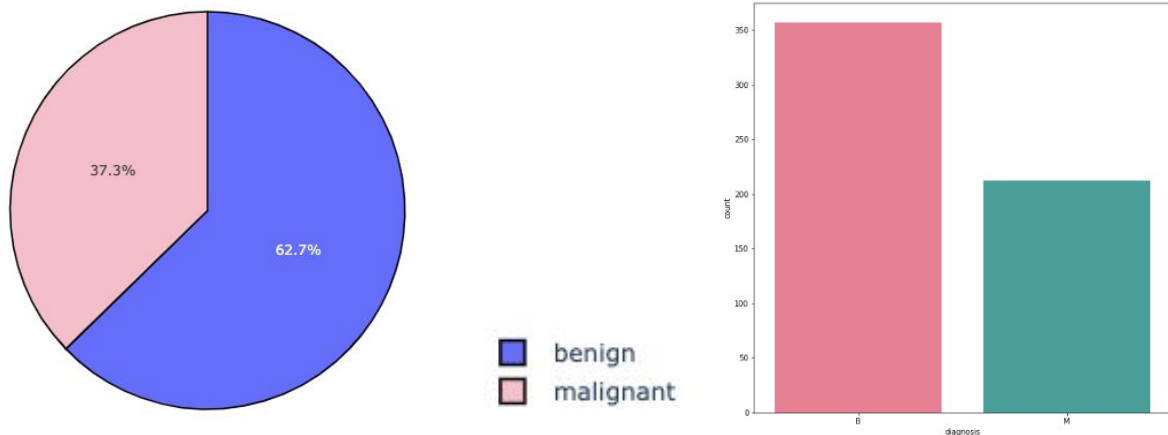
That being said, early diagnosis significantly increases the chance of survival and our goal of this project is finding a suitable machine learning algorithm to help make a better prediction to detect breast cancer based on our pre recorded data set.

## **Data information**

The dataset we used, often called “The Breast Cancer Wisconsin (Diagnostic) DataSet” , was obtained from UCI Machine Learning Repository and was originally supplied by University of Wisconsin . The dataset contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. FNA methodology is a biopsy procedure that uses a thin, hollow needle to remove a sample of cells from the abnormal area of the breast. All attributes in the dataset describe characteristics of the cell nuclei present in the image. As feature 1 shows below, the dataset has over 569 observations, of which 357 are malignant and 212 are benign.

There are a total of 30 attributes in the dataset and three main values associated with 10 attributes ; “Mean”, “Standard Error”, and “Worst”. “Mean” value of each attribute defined as ‘mean distance from the center to points on the perimeter’, “Standard error” of each attribute indicates ‘standard error for the mean of distances from the center to points on the perimeter’ and “worst” of each attributes means ‘the largest value for the mean of distances from the center to points on the perimeter’ according to UCI Machine Learning Repository.

**(357 Malignant , 212 Benign)**



**Feature1**

## Data cleaning

There are no missing values found in our dataset (Appendix 1). We decided to drop column “ID” and “Unnamed: 32” since they are not significant to our research and “Unnamed: 32” is a blank column.

## Literature Review

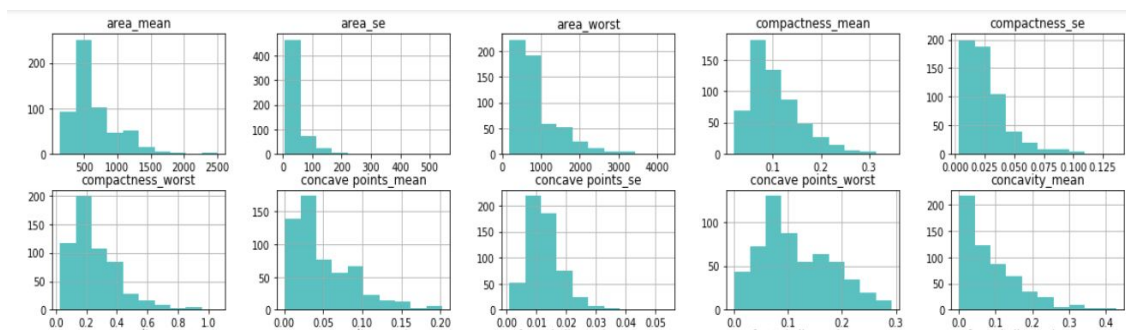
Before we studied this dataset, we wanted to make sure to check on previous research or experiments related to breast cancer to gain a baseline model so we could evaluate our performance. According to “Computerized Breast Cancer Diagnosis And Prognosis From Fine Needle Aspirates” from William H. Wolberg, M.D. et al, shape, size, texture of the tumors are the most significant features they used in predicting the prognosis or outlook of a woman with the disease.

# Exploratory analysis

## Descriptive Statistics

- **Distribution**

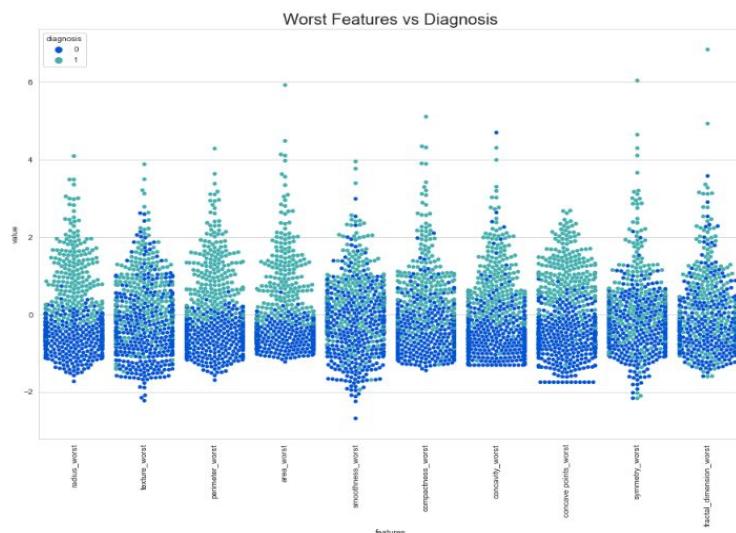
From Appendix 3, when skewness = 0 : normally distributed., when skewness > 0 : more weight in the left tail of the distribution and when, skewness < 0 : more weight in the right tail of the distribution. We can tell some of the columns are right-skewed (Appendix 4 for more), however, most of the features are normally distributed



Feature2

- **Outliers**

From the swarm plots and other box plots from Appendix 5, we found out there are some outliers in the dataset. We used normalization to ensure we can properly utilize that data for further queries and analysis.



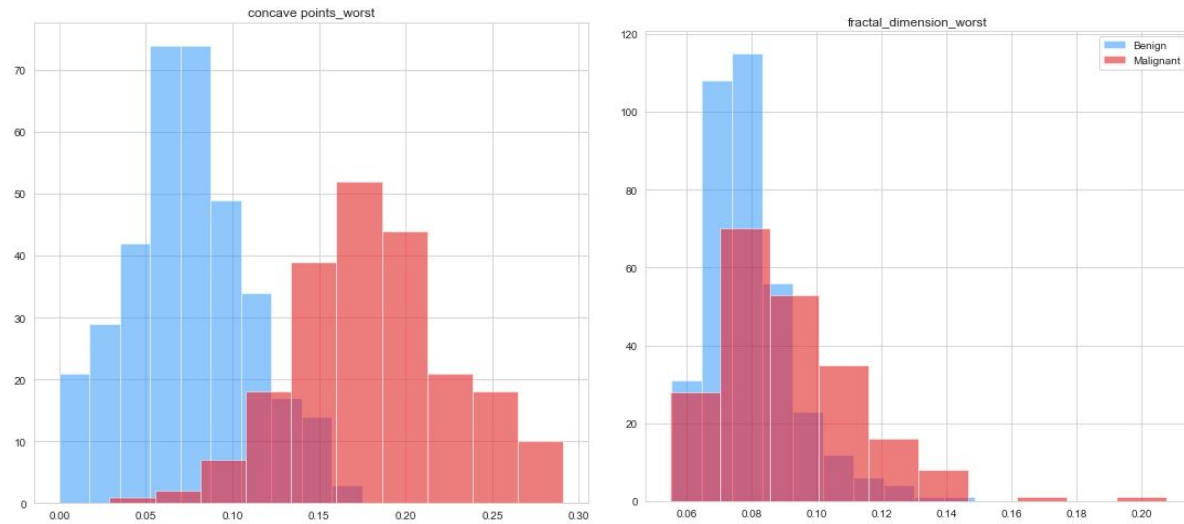
Feature3

## **Malignant vs Benign**

Before we explored the data and built models classifying tumors into malignant or benign, we wanted to know more about benign and malignant tumor cells to better understand the project. Benign tumor cells grow only locally and cannot spread by invasion or metastasis. On the other hand, malignant cells invade neighboring tissues, enter blood vessels, and metastasize to different sites. In our data, as you can see from the feature1 above, we have 37.3% of patients with malignant tumors and 62.7% of patients with benign tumors. We can confidently say that the dataset is not imbalanced, therefore, we are not going to use 'smote' or any other techniques that help make the dataset balanced.

## **Relationship between each attribute and Diagnosis**

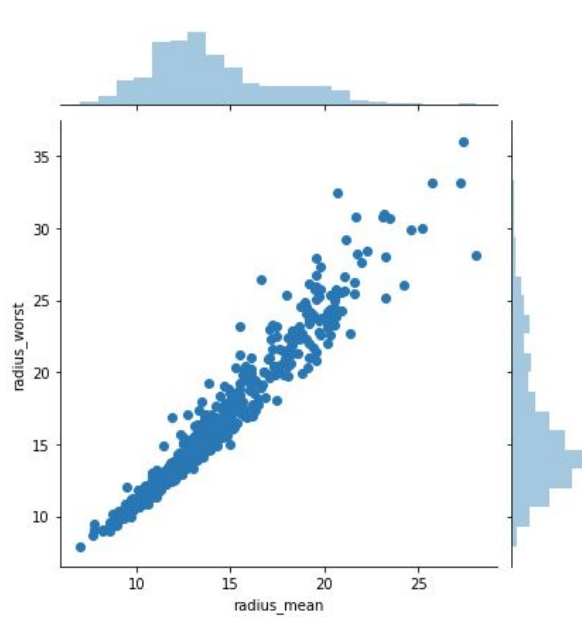
Moreover, we investigated the relationship between each attribute and diagnosis by using a histogram (Appendix 6-1). We believed that if there is a significant difference in malignant and benign diagnosis for each attribute, then we can confidently distinguish the tumors and it would assist doctors in finding where the threshold should be to differentiate whether the diagnosis is benign or malignant. For example, as feature4 shows, "concave\_point mean" is highly distinguishable between malignant and benign, which indicates that different cell status could have a different value of concave point, so it will be easier for the model to predict the correct result. On the other hand, the second graph of the feature 4 shows fractal\_dimension's distribution. The graph indicates that no matter whether the cell is malignant or benign, the fractal dimension values seem to be similar, and the most frequent value is 0.08. We are going to use this for feature selection for our baseline model.



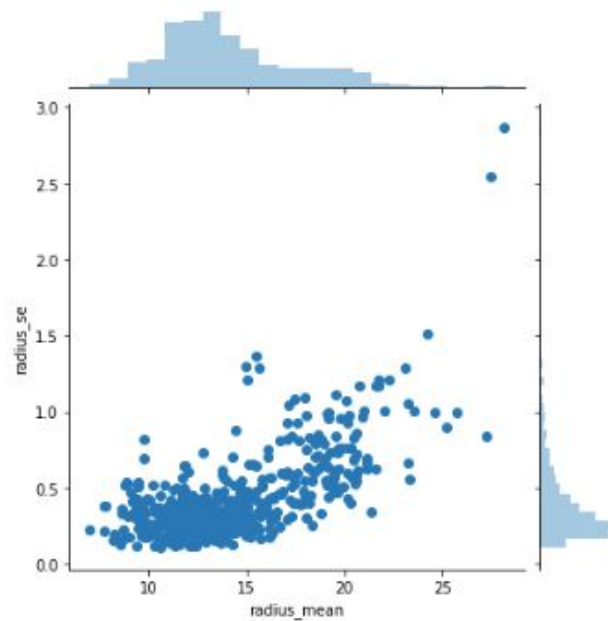
## Feature4

### Correlation matrix

We used correlation matrix to see the relationship between independent variables and found that there are many features which are intercorrelated especially between “mean” attributes and “worst” attributes. For example, as feature 5 shows below, “radius mean” and “radius worst” are highly correlated. As we mentioned above, the dataset records ten different attributes, such as “area”, “radius”, “symmetry”, etc, and for every attribute, there are three different values associated with the attributes. Again, these are the “mean”, the average value of the attribute, “worst”, largest mean value of the attribute, and “standard error”, standard error of the attribute. Therefore, by definition, we believe “worst” and “mean” are intrinsically correlated. In order to get better performance of our models and feature selections, we created three different dataframes by “mean”, “standard error” and “worst”.



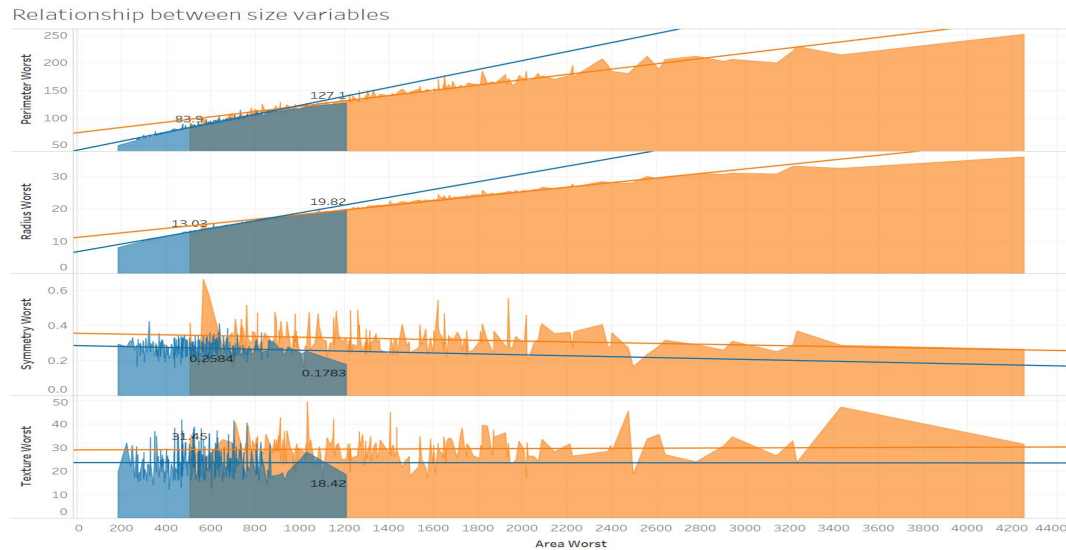
**Feature5** (radius mean vs radius worst)



(radius mean vs radius standard error)

## Linear relationship

In order to further validate the features we selected, we also conducted tableau analysis (feature 6) and found out that perimeter, radius and area form a positive linear relationship, which means there is a positive correlation between those variables and the dependent variables('diagnosis'). On the other hand, both "symmetry" and "texture" do not have significant relationship with "area". Instead, throughout all the observations that include benign and malignant, "symmetry" and "texture" values are widely distributed, which means that linear models would not work well with these two attributes included. Moreover, we could see a clear threshold of "area\_worst", meaning if a cell had grown to over 1200 units of area, then the cell had no chance to be benign. The same thing also happened to the low-end threshold, if a cell did not exceed 500 units of area, then the patient's cell is highly likely to be benign.



**Feature 6**

## Feature selection for hypothesis model and the baseline model

So far we have explored the data and had baseline knowledge on how to use the data for our model. Based on the data form (binary), we decided to use classification models. In order to build models, a critical step for our quantitative analysis is feature extraction. In order to gain higher classification performance, we need to extract the appropriate features and ensure we are getting rid of redundancy that could affect model accuracy later.

- **Attributes selection for hypothesis model**

Based on this research, we decided to test this statement by extracting attributes from our dataset that were related to shape, size, and texture to compare with our baseline model to see if our models built by these attributes bring higher accuracy than our baseline model. Among the 30 attributes of our data, we found that 21 attributes are related to shape, size, and texture. Since each attribute might use different measurements, we normalized the data and used correlation matrix to identify the correlation between independent variables and then dropped those features that are highly intercorrelated. We ended up having 9 attributes ready for the hypothesis model. ('radius\_mean','texture\_mean','radius\_se','texture\_se','radius\_worst','concavity\_se','symmetry\_worst','concave points\_se','symmetry\_worst').



- **Attributes selection for our baseline model**

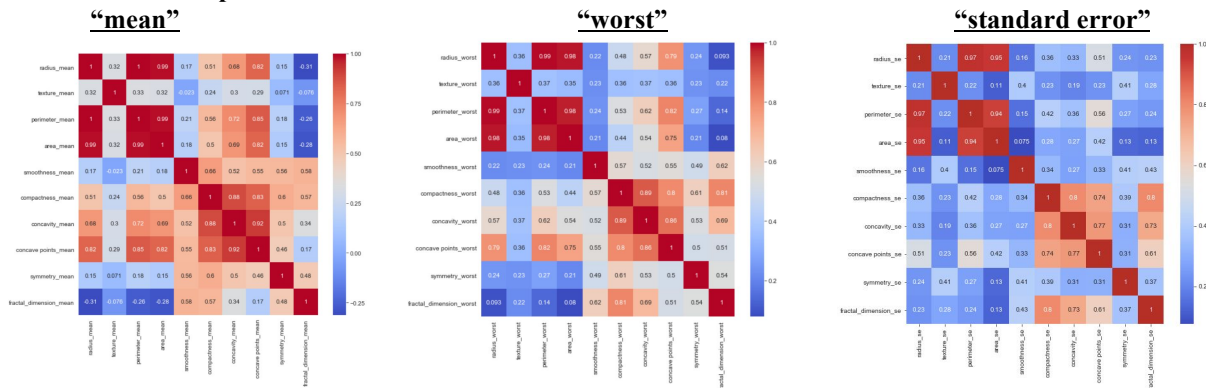
In order to decide whether the hypothesis model is relatively good or not, we need a comparison. Therefore, we need to select attributes for our baseline model. Instead of selecting attributes related to size, shape, and texture, we considered all 30 attributes for the baseline model. Once we normalized the data, we ran correlation for each data frame ('mean', 'standard error', 'worst' feature 7) and dropped attributes that are highly intercorrelated.

Then, we used a histogram (Appendix 6-1) to identify features that are enable to differentiate between whether the diagnosis is benign or malignant, such as “concave\_point mean” from the feature 4. On the other hand, we dropped attributes such as “fractal\_dimension\_worst” that the threshold for malignant and benign is not very clear. We also cross-checked through the frequency table (Appendix 6-2) and finally narrowed down to 7 attributes for our baseline model.

('radius\_mean', 'perimeter\_mean', 'concavepoints\_mean', 'concavity\_mean', 'radius\_worst', 'perimeter\_worst', 'concave points\_worst')

## Feature 7

### Correlation heatmap for



## Model Building and Results

Since it is a binary case, we used four classification algorithms; Naive Bayes, Decision Tree, Random Forest, and Logistic Regression. We splitted the dataset into 70% for the train and 30% for the test. Overall, as you can see from the table below, the hypothesis model had a higher accuracy than the Baseline model for each algorithm.

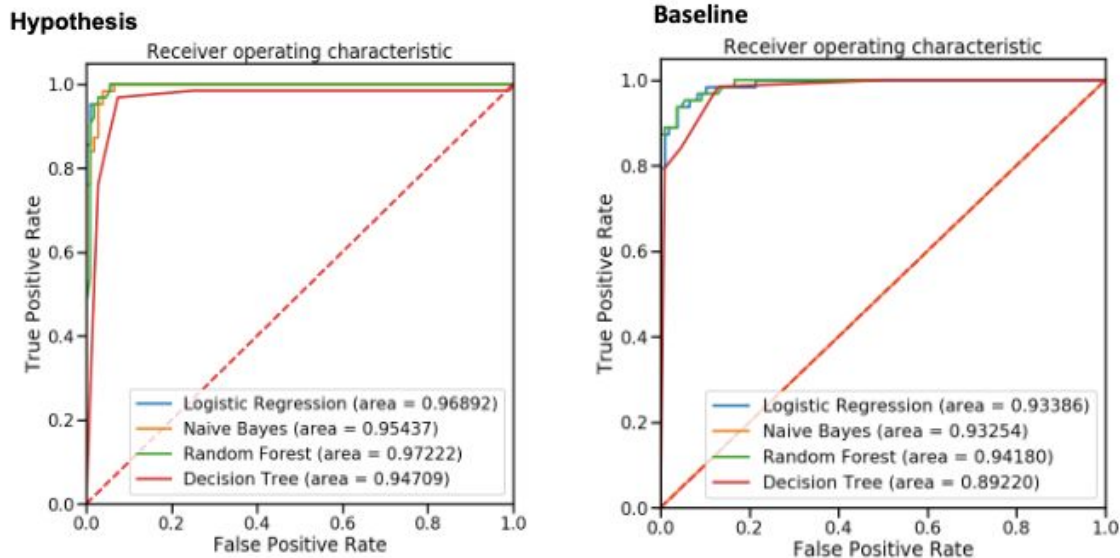
	Naive Bayes	Decision Tree	Random Forest	Logistic Regression
Hypothesis	95.9%	94.2%	96.5%	96.5%
Baseline	93.6%	91.8%	94.7%	94.2%

Logistic Regression and Random Forest algorithm had the highest accuracy for the both hypothesis and baseline model. Since accuracy is the same, we wanted to look more into recall. Recall is the number of true positives divided by the number of true positives plus the number of false negatives. The reason why we want to focus more on recall rather than precision is that for this specific topic, it is very important to reduce the number of false-negatives (patients actually have malignant tumors but the model diagnoses them as benign). The higher recall percentage, the lower number of false-negatives in the model. Feature 8 shows that Random Forests recall is 100%, which means 0 false-negatives. **Therefore, we decided to choose the “Hypothesis” model using the Random Forest algorithm.**



Feature 8

## AUC - ROC Curve Analysis



### Feature 9

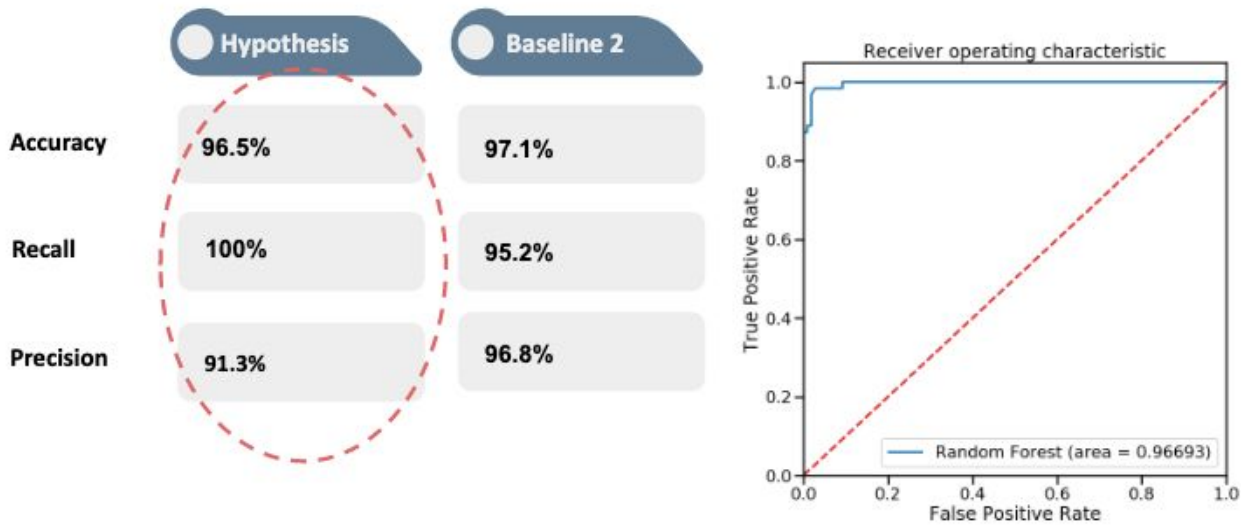
Accuracy may or may not be a good metric to judge the models. In many problems, and particularly in the medical field, incidences of positive and negative examples are not evenly distributed. Although our dataset is somewhat balanced, we still want to use the AUC- ROC curve metric to find an algorithm in which we can achieve a high TPR while minimizing the FPR. The higher AUC, the better the model is at predicting TPR (Benign as Benign, Malignant as Malignant). Feature shows that for both models, Random Forest algorithm is the best algorithm that is capable of distinguishing between Malignant and Benign.

Random Forest with Hypothesis model has the highest AUC rate, meaning there is 97.22% chance that the hypothesis model will be able to distinguish between positive class and negative class.

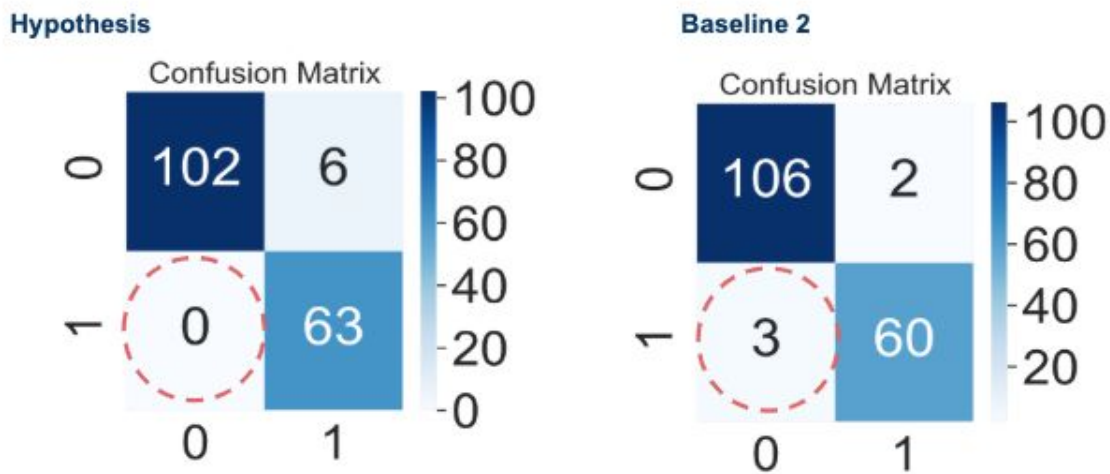
### Baseline 2

Although our baseline model did not work better than the hypothesis, we wanted to see if we can increase the accuracy yet maintain low false-negatives by adding other important features that we may have ignored using the Random Forest algorithm. We added “area\_worst” and “texture\_worst” to our baseline model, which is now called “baseline 2”. The Feature 10 shows us we are able to increase the accuracy from 94.7% to 97.1%. However, the accuracy rate is still

lower than the hypothesis model, which means the false-negative numbers are higher than the hypothesis.



#### Feature 10



#### Feature 11

## Conclusion

Despite the fact that we are able to have a higher accuracy than the hypothesis model by adding two features, the hypothesis model recall rate is still higher than the baseline 2. As we mentioned before, a decrease in false-negative numbers is very important for this topic, and we concluded that **hypothesis model using random forest** is the best model that can be used for helping to make a better prediction for the diagnosis. Also, we can prove that size, texture, and shape are the most important features to predict breast cancer diagnosis.




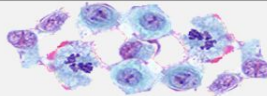

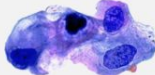


## Insights

### 1. Why does baseline2 have a higher accuracy than the hypothesis by adding “texture\_worst” and “areat\_worst”?

**Area\_worst:** We did not include this feature to our initial baseline model because it was highly intercorrelated with other independent variables. However, according to Canadian Cancer Society, “as a tumour gets bigger, cancer cells can spread to surrounding tissues and structures by pushing on normal tissue beside the tumour. Cancer cells also make enzymes that break down normal cells and tissues as they grow.” Thus, we started looking into the area variable. We found out that the bigger the area is, the more likely the cell could be malignant. Despite the fact that “area\_worst” is highly intercorrelated with other worst variables, we decided to include this variable because past research stated it is important(Canadian Cancer Society). Additionally, as our given dataset is relatively small, we need to consider there might be other factors that can cause collinearity for this variable.

**Texture\_Worst:** As Feature 12 shows below, malignant and benign cancer cells are visually different, and most statistical data can be visualized and recorded from the mammogram. The most highly used attributes are the three that we mentioned through our modeling and feature selection segments, which are “size”, “shape”, and “texture” related attributes. Since we already

proved how impactful “shape” and “size” variables are, the first takeaway will be why “texture” improved our model accuracy by almost 3%. In fact, in research conducted previously to 2008, researchers did not have a clear understanding of “texture”, which is why most research at that time did not build models around utilizing the “texture” variable. However, in nascent research, people have started to utilize this piece of information, and our hypothesis model achieved a high accuracy by adding “texture” to the model. In fact, according to research done by Angkoon, “texture” is described as the surface of an object as being fine, coarse, or smooth. It is calculated based on fractal dimension, which is one of the attributes included in our dataset. However, fractal dimension itself does not improve our model, even though it is part of texture’s function. In fact, texture is a function of the distribution of pixel intensity and can be calculated from the gray values at each point in the image. Therefore, we conclude that, even though texture itself does not correlate with neither with “shape” or “size”, we should combine the three to observe the impact on classification.

Normal	Cancer	
		Large, variably shaped nuclei
		Many dividing cells; Disorganized arrangement
		Variation in size and shape
		Loss of normal features

**Feature 12**

## 2. Cases where Baseline 2 needed to be used

The medical field is a case-by-case field and different methodology can be used for different patients. With that being said, Baseline 2 model (accuracy: 97.1% , 3 false-negative) still can be used for specific patients when needed by adjusting threshold. We adjusted threshold to 0.3 and 0.7 given default with 0.5. As feature 13 shows, on threshold 0.3 the baseline 2 model can

achieve 0 false-negatives yet still have the highest accuracy.

	Threshold		
	0.3	0.5	0.7
True Positive	63	60	55
False Positive	10	2	1
False Negative	0	3	8
True Negative	98	106	107

#### Feature 13

### Algorithms Analysis

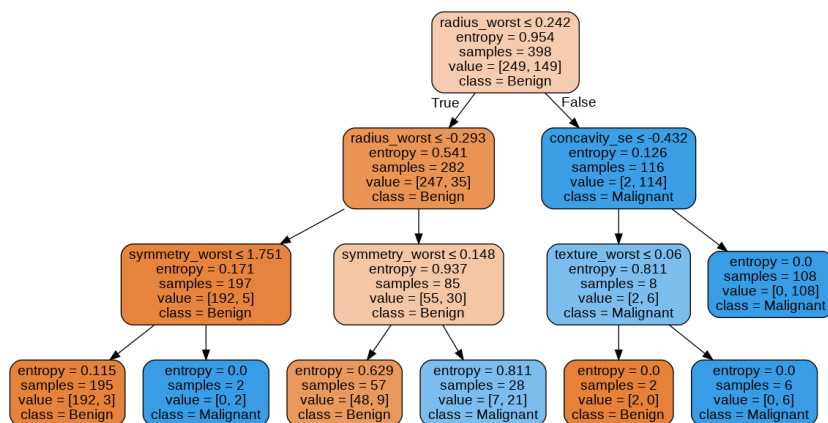
- **Naive Bayes : [accuracy] hypothesis : 95.9 %, Baseline: 93.6%**

Naive Bayes is a very popular classification algorithm that is mostly used to get the base accuracy of the dataset, because Naive Bayes deems every variable independent with no correlation. This is why it performs relatively well when we drop attributes that are highly correlated. We used Naive Bayes to do real-time prediction, wanting to predict whether the breast cell is malignant or benign, based on variables like “radius”, “area”, “symmetry“, “texture”, and etc. So the logic behind the algorithm is that, first, the model computes all the input attributes into a frequency table. For instance, how many times a certain value of “radius” appears when the diagnosis is malignant or benign. From there the model calculates the possibility of each class given each attribute. Finally, the model uses the Bayesian equation to calculate possibility for malignant and benign according to the variables, and the final prediction is the highest among all classes. The accuracy of Naive Bayes is somewhat lower than both

logistic and random forest, and we believe the reason being is our data is numeric. Therefore, even though we drop the highly correlated variables, we do not achieve as high as an accuracy as other models because of the inherent nature of Naive Bayes working best with categorical variables.

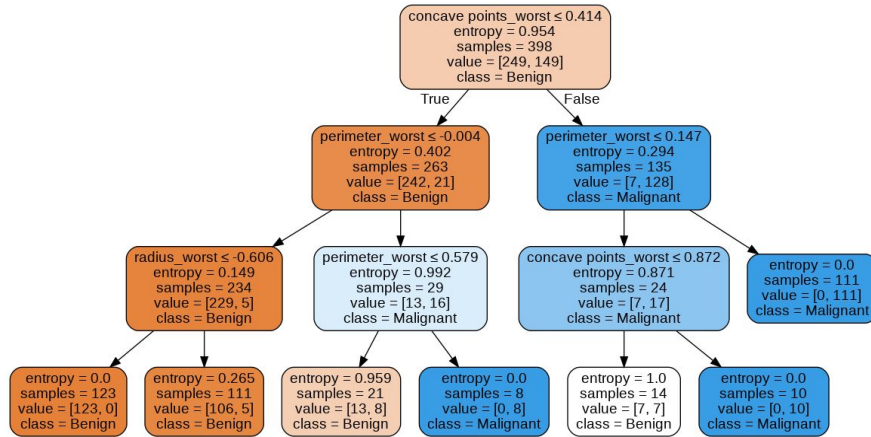
- **Decision Tree: [accuracy] hypothesis : 94.2 %, Baseline: 91.8%**

We used categorical Variable Decision Tree, which has categorical target variable, Malignant and Benign. As feature 14 shows, for our hypothesis variables, “radius”, “texture”, “symmetry” and “concavity” are ordered first in the decision tree, meaning these four variables have the highest information gain. Therefore, this helped to narrow down features that we need to select while building our baseline 2 model. However, as Appendix 15 and 16 show, for our baseline model 1 and 2, they replaced “radius” with “perimeter” and the root node, which is the variable that has the highest information gain, became “concave point”, instead of “radius”. Therefore, we believe “radius”, “symmetry” and “texture” are the right attributes to choose, and we decided to move forward with random forest to achieve a better accuracy.

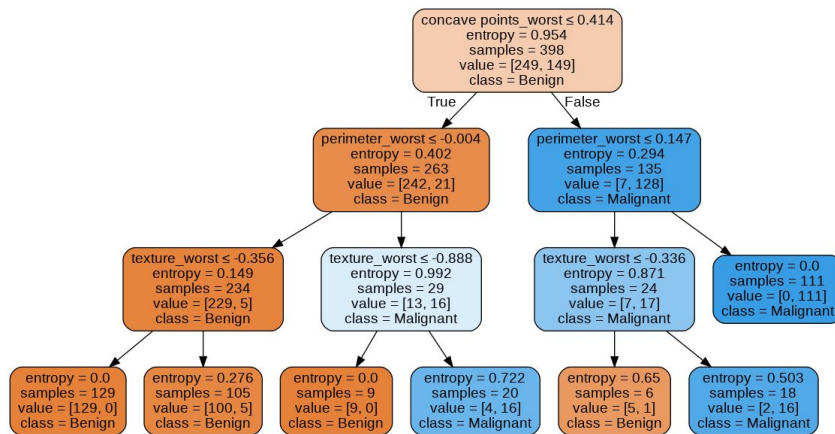


Feature 14





**Feature 15**



**Feature 16**

- **Random Forest: [accuracy] hypothesis : 96.5 %, Baseline: 94.7%**

Random forest will generate results by accumulating N decision trees, and each tree "votes" or chooses the class, and the class receiving the most votes by a simple majority is the predicted class. As a result our random forest model achieves a high accuracy with a low false negative result (shown in Appendix 4,5). The accuracy for random forest is 96.5%, which aligns with our logistic model and which almost all past research prefers to use. The recall achieved 100% for random forest model as well which solidified it as our best model.

- **Logistic regression : [accuracy] hypothesis : 96.5 %, Baseline: 94.2%**

Logistic regression is a regression model. Just like linear regression assumes that the data follows a linear function, logistic regression models the data using the sigmoid function. In our case, “diagnosis” is a target variable which can take only two possible types, “malignant” or

“benign”. Then we generate a feature matrix with the selected feature variables and 568 observations. We put the vectors in the matrix into sigmoid function to get the probability of classification. At last, we set a threshold to determine the category of each observation. As a result, we built a logistic regression model with a default threshold with 0.5 and we got 96.5% accuracy for the hypothesis model and 94.2% for the baseline 1.

## Appendix

### Appendix 1 - Missing variables check

```

]: #checking missing values by each coulumn
breastcancer.isnull().sum()

# result: non

]: id                                0
   diagnosis                        0
   radius_mean                      0
   texture_mean                     0
   perimeter_mean                   0
   area_mean                        0
   smoothness_mean                  0
   compactness_mean                 0
   concavity_mean                   0
   concave points_mean              0
   symmetry_mean                    0
   fractal_dimension_mean           0
   radius_se                        0
   texture_se                       0
   perimeter_se                     0
   area_se                          0
   smoothness_se                    0
   compactness_se                   0
   concavity_se                     0
   concave points_se                0
   symmetry_se                      0
   fractal_dimension_se             0
   radius_worst                     0
   texture_worst                    0
   perimeter_worst                  0
   area_worst                       0
   smoothness_worst                 0
   compactness_worst                0
   concavity_worst                  0
   concave points_worst             0
   symmetry_worst                   0
   fractal_dimension_worst          0
   Unnamed: 32                      569
   dtype: int64

```

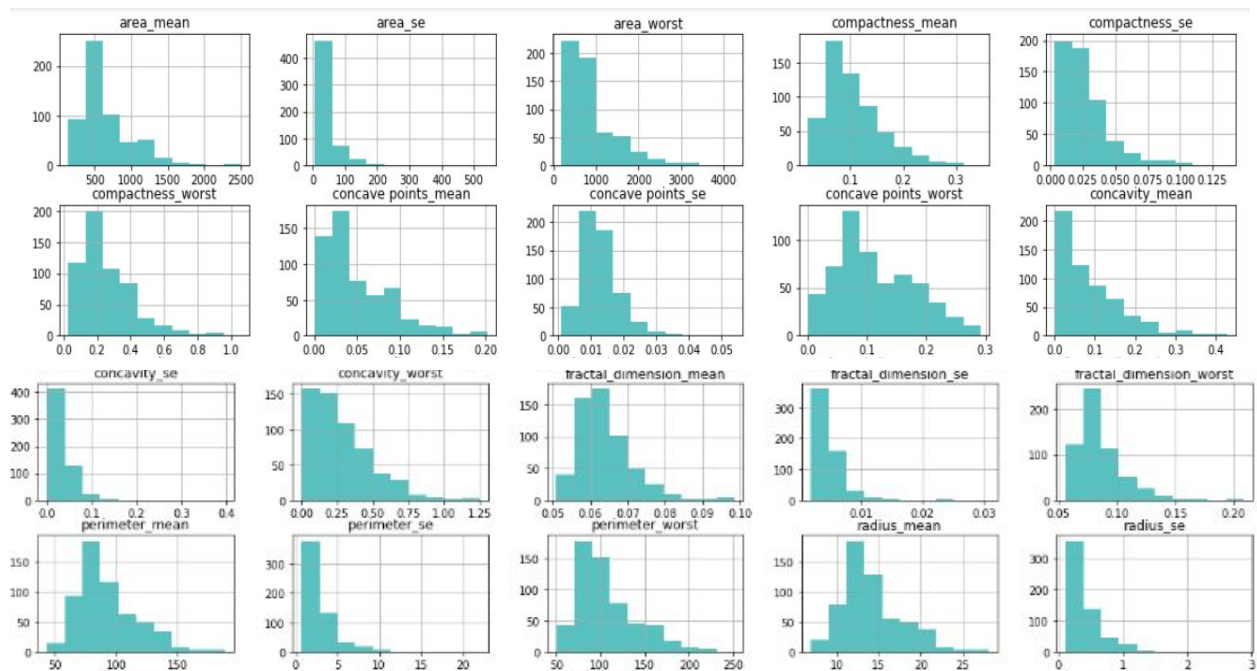
## Appendix 3 Descriptive statistics (Skewness)

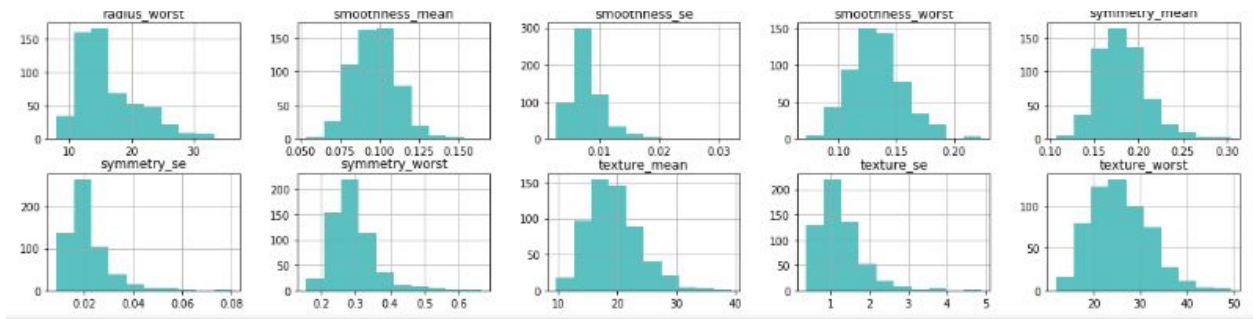
```

radius_mean      0.942380
texture_mean     0.650450
perimeter_mean   0.990650
area_mean        1.645732
smoothness_mean  0.456324
compactness_mean 1.190123
concavity_mean   1.401180
concave points_mean 1.171180
symmetry_mean    0.725609
fractal_dimension_mean 1.304489
radius_se        3.088612
texture_se       1.646444
perimeter_se     3.443615
area_se          5.447186
smoothness_se    2.314450
compactness_se   1.902221
concavity_se     5.110463
concave points_se 1.444678
symmetry_se      2.195133
fractal_dimension_se 3.923969
radius_worst     1.103115
texture_worst    0.498321
perimeter_worst  1.128164
area_worst       1.859373
smoothness_worst 0.415426
compactness_worst 1.473555
concavity_worst  1.150237
concave points_worst 0.492616
symmetry_worst   1.433928
fractal_dimension_worst 1.662579
dtype: float64

```

## Appendix 4 Descriptive statistics- distribution of each variable





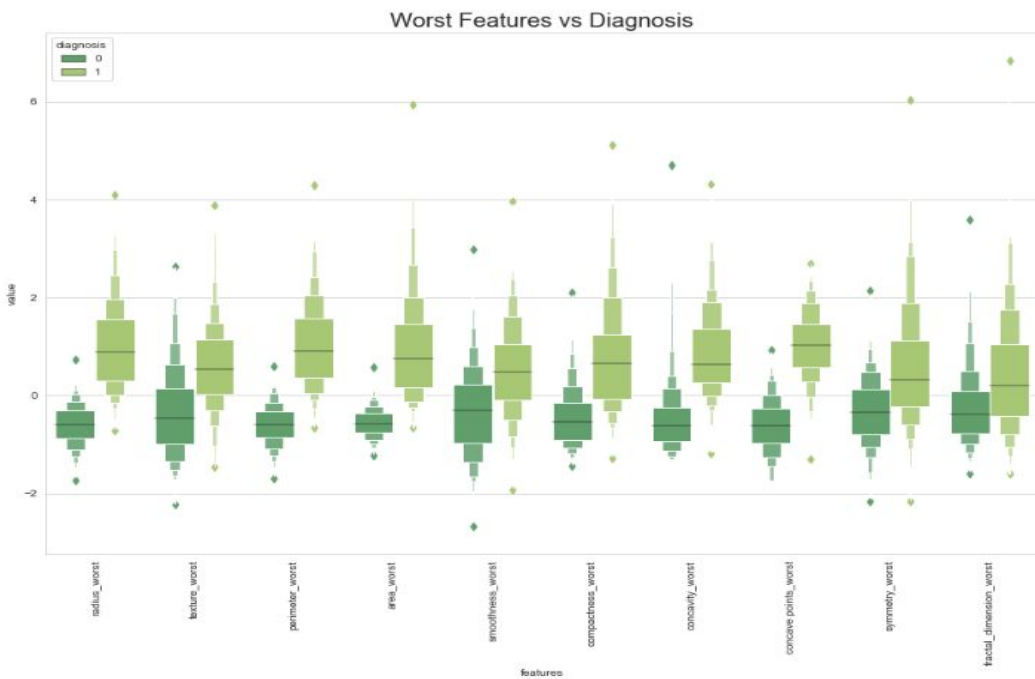
## Appendix 4-1 Descriptive statistics

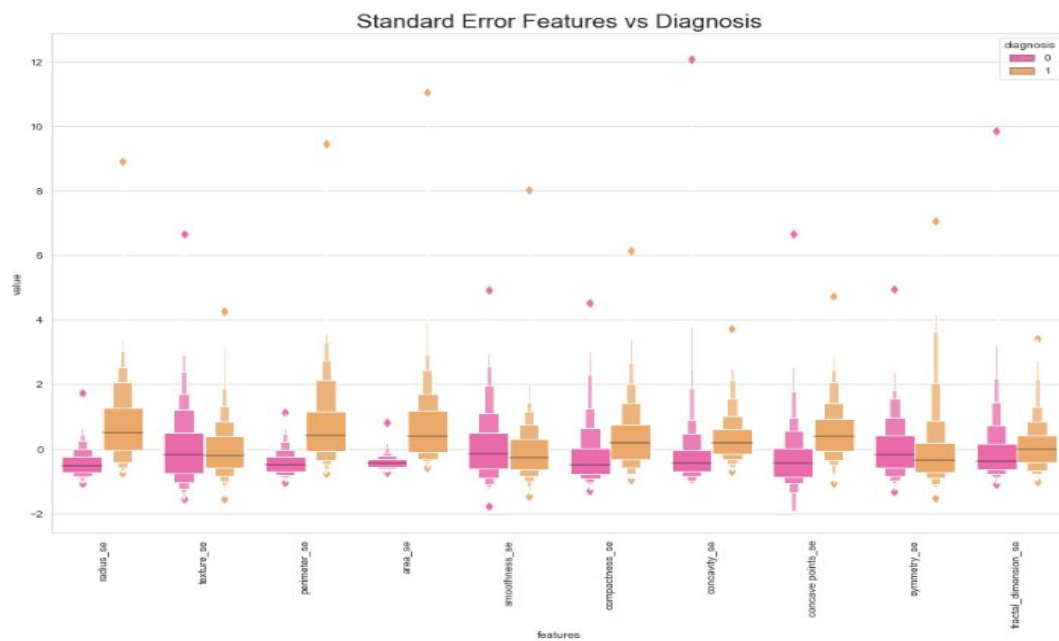
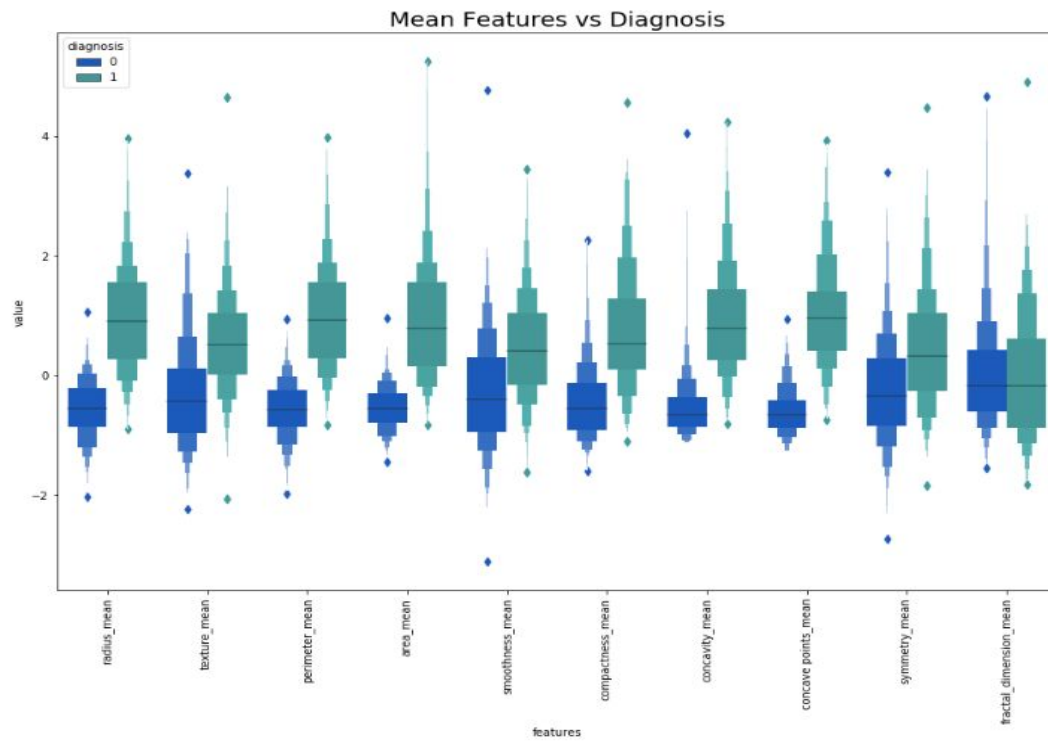
Out[9]:

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fract
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	

8 rows x 30 columns

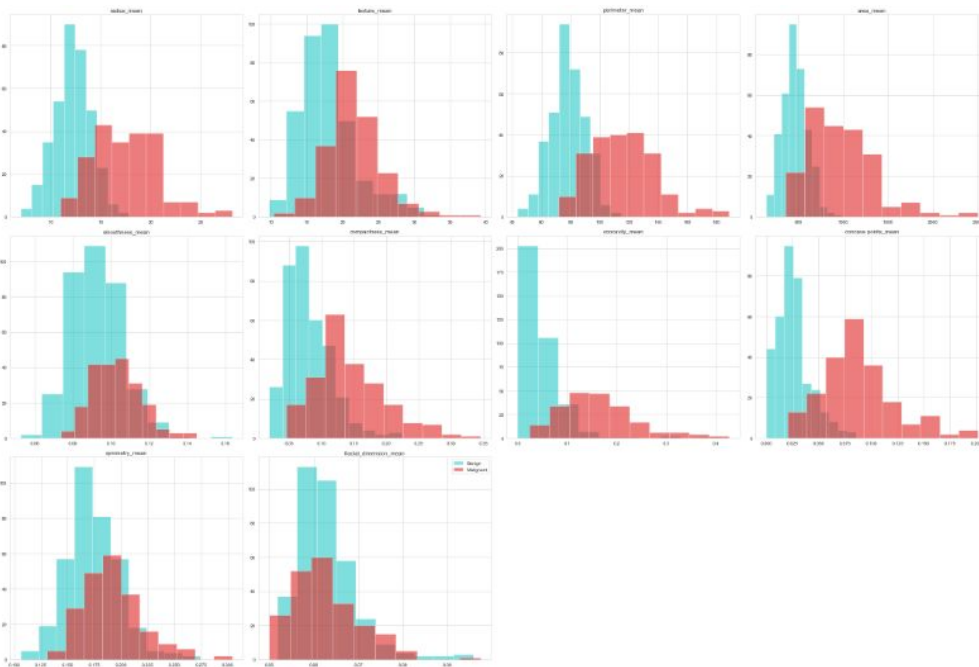
## Appendix 5 - Box plots



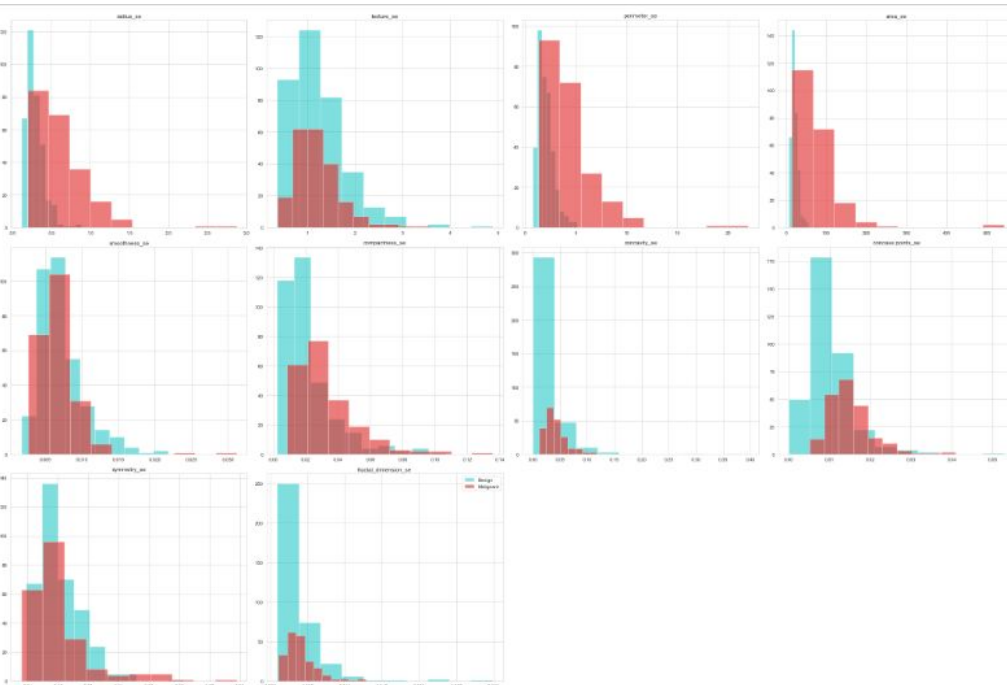


## Appendix 6-1 - Relationship between “Diagnosis” and each variable

“Mean”

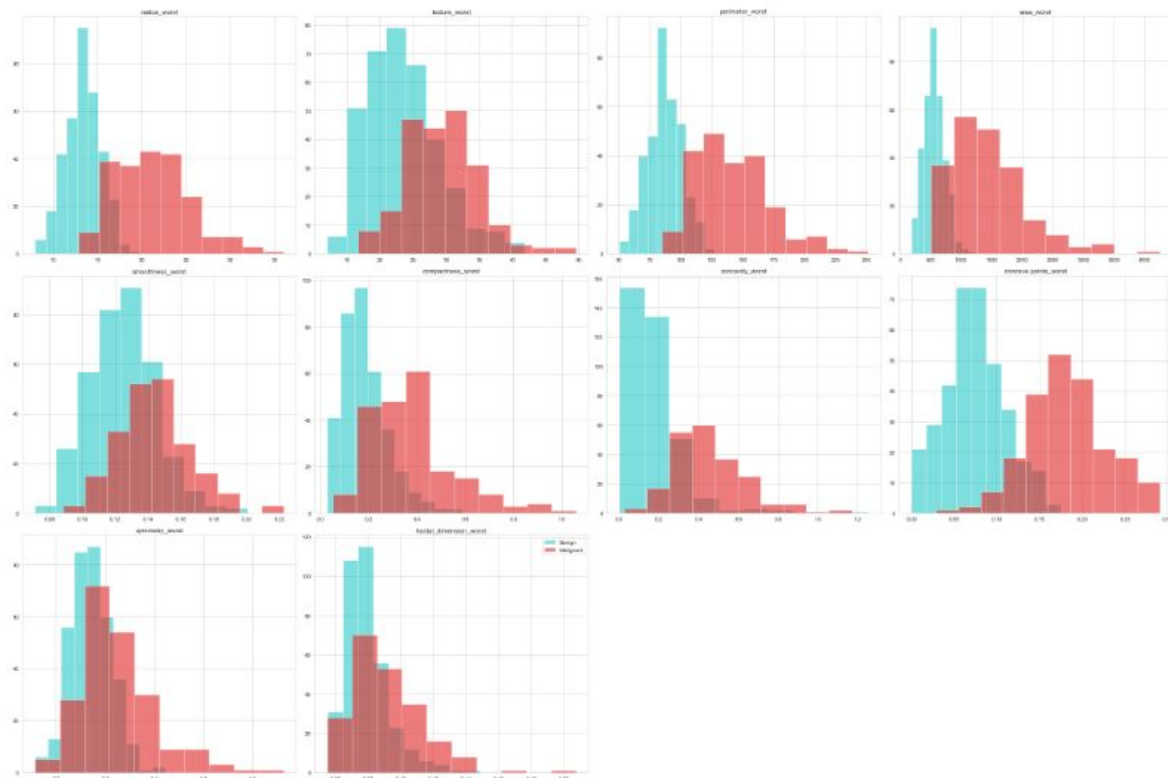


“Standard Error”



“Worst”

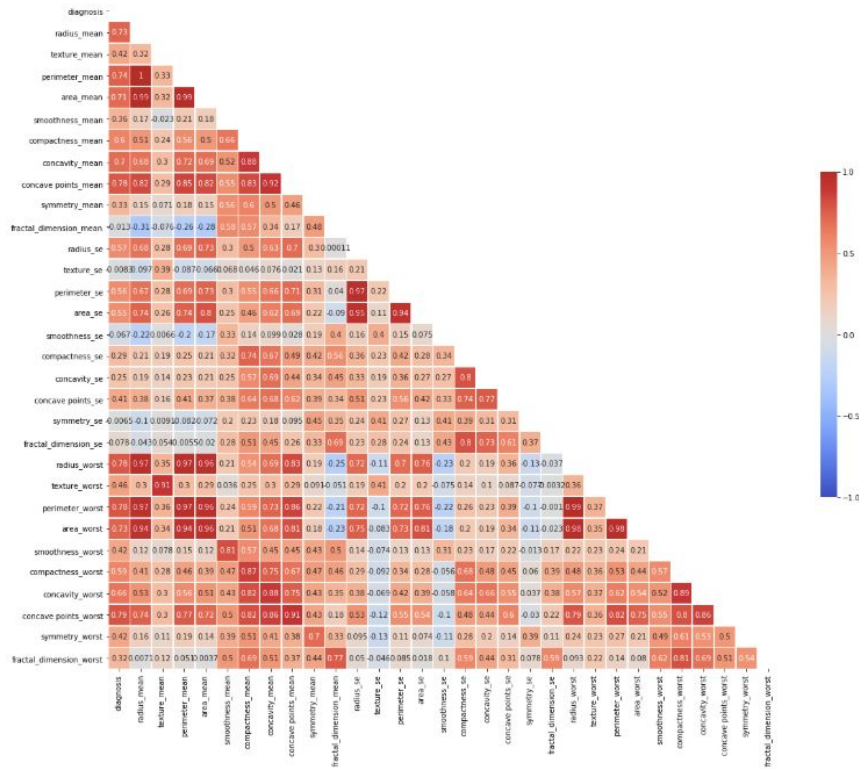




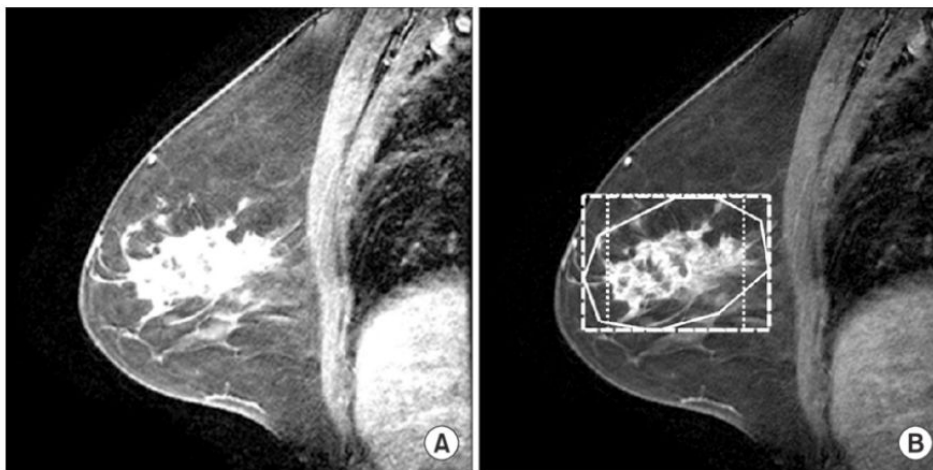
## Appendix 6 -2 Frequency table

	diagnosis	radius_se_bins	radius_se
0	B	<0.25	162.0
1	B	0.25-0.50	176.0
2	B	0.50-1.25	19.0
3	B	1.25-1.5	NaN
4	B	1.5-1.75	NaN
5	B	1.75-2.0	NaN
6	B	2.0-2.25	NaN
7	B	2.25-2.5	NaN
8	B	2.5-2.75	NaN
9	B	2.75-3.0	NaN
10	M	<0.25	15.0
11	M	0.25-0.50	80.0
12	M	0.50-1.25	110.0
13	M	1.25-1.5	4.0
14	M	1.5-1.75	1.0
15	M	1.75-2.0	NaN
16	M	2.0-2.25	NaN
17	M	2.25-2.5	NaN
18	M	2.5-2.75	1.0
19	M	2.75-3.0	1.0

## Appendix 7 Correlation heat map



## Appendix 8 how “Texture” are imaged digitally and calculated.



## Appendix 9 - how malignant and benign are different in texture and size <https://ww5.komen.org/>



Types of invasive breast cancer	Proportion of all invasive breast cancers	Tumor characteristics	Prognosis
Invasive ductal carcinoma (IDC)	70-80%	<ul style="list-style-type: none"> <li>• Hard tumor texture</li> <li>• Tumor is irregular, star-shaped</li> <li>• Cell features vary</li> <li>• DCIS often present</li> </ul>	<ul style="list-style-type: none"> <li>• Prognosis varies with stage and grade of tumor</li> </ul>

# Bibliography and Reference

## Past Research & Experiments

Wolberg WH, Street WN, Heisey DM, Mangasarian OL. Computerized Breast Cancer Diagnosis and Prognosis From Fine-Needle Aspirates. *Arch Surg*. 1995;130(5):511–516.

doi:<https://doi.org/10.1001/archsurg.1995.01430050061010>

Jitaree, S., Phinyomark, A., Boonyaphiphat, P., & Phukpattaranont, P. (2015). Cell type classifiers for breast cancer microscopic images based on fractal dimension texture analysis of image color layers. *Scanning*, 37 2, 145-51 .

Cancer Research UK. (2014, October 29). How Cancer Can Spread. Retrieved from: <http://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-can-spread>.

Canadian Cancer Research. How cancer starts, grows and spreads.

<https://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/how-cancer-starts-grows-and-spreads/?region=en>

## Variable descriptions, Breast Cancer Stat

<https://gis.cdc.gov/Cancer/USCS/DataViz.html>

<https://ww5.komen.org/AboutBreastCancer/FactsandStatistics/WhatIsBreastCancer/InvasiveBreastCancers.html>

<https://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/how-cancer-starts-grows-and-spreads/?region=en>

## Model explanation

<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

[https://medium.com/@rishabhjain\\_22692/decision-trees-it-begins-here-93ff54ef134](https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134)