**CAPSTONE FINAL REPORT**

**Introduction**

Breast cancer is not only the most common type of cancer, but also one the most dangerous types of disease for women. It has the second highest mortality rate. If breast cancer is not detected early, it can and may result in the death of the individual. Approximately 12% of women are affected by breast cancer, and the number is steadily increasing. Unfortunately, doctors do not have the ability to readily detect early stages of breast cancer due to the symptoms of the disease being absent. When breast cancer is diagnosed, it is due to an abnormal lump found in the breast via self-examination or x-ray(when a tiny speck of calcium is seen). Once lump is found, doctors then determine whether it is cancerous.

Due to the inability to detect breast cancer early, it is necessary to discover and interpret new knowledge through Machine Learning. Machine Learning is both known and used in Bioinformatics for breast cancer diagnosis.

The goal of this project is to determine how one predicts early stages of breast cancer given the breast cancer features. What we will be doing is looking at the sizes of: the mean radius, and mean area to see if the size (the bigger the mean radius or area) is an indication of breast cancer

**Data & Data Cleaning**

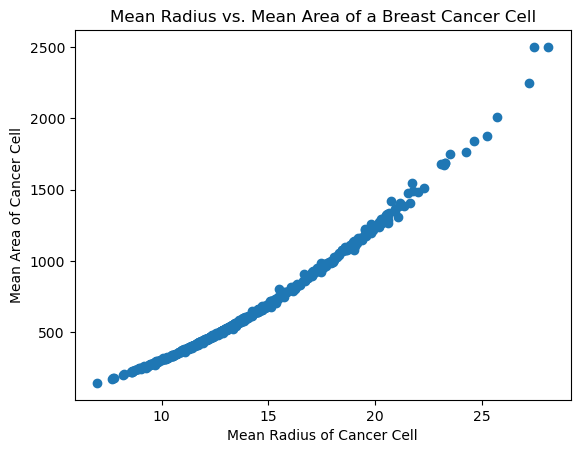
The data used in this project was collected from Kaggle, which is a huge website for large amounts of data is collected. This project had enough data obtained from Kaggle in order to complete this project. The data from Kaggle initially had three different csv files, but it was narrowed to one to have a more focused topic, and obtain a good predictive model. Below are the links for the data:

<https://www.kaggle.com/datasets/merishnasuwal/breast-cancer-prediction-dataset>

After obtaining a good dataset from Kaggle, the next step was to upload the data into jupyter notebook to see the format, shape and column names. Next, we figured out what the data types were, and the summary statistics. After that, we checked if there were any missing values or duplicates. If there were any, they would be deleted. In our case, there weren’t any missing values, but on the other hand, there were 563 duplicate values in the dataset, which were dropped. Once the data was cleaned, it was then ready for Exploratory Data Analysis (EDA)

**Exploratory Data Analysis (EDA)**

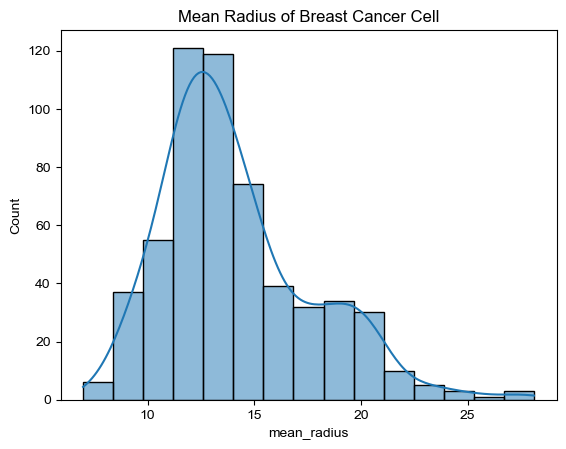
In this section, we further explore the data. First step was to obtain the information for the data, then further narrow it down by obtaining information for both the mean radius and the mean area. This was to see the comparison between both, and see if there are any trends. The mean radius was plotted on the x-axis, while the mean area was plotted on the y-axis. Below is the graph:



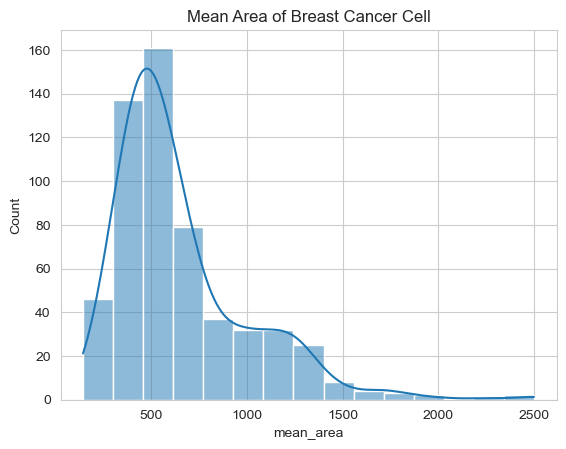
**Figure 1.** Mean Radius vs. Mean Area of cancer cell

Looking at figure 1, one can see that there is a relation between mean radius and mean area of cancer cells. The higher the mean radius, the higher the mean area of the cancer cells. This is slowly and indication of our hypothesis.

We went further to plot an histogram of just the mean radius of breast cancer cells to see what the approximate average is, and how many were found based on the given data. Below is the result.



**Figure 2.** Mean radius of breast cancer cells (no units provided in data).

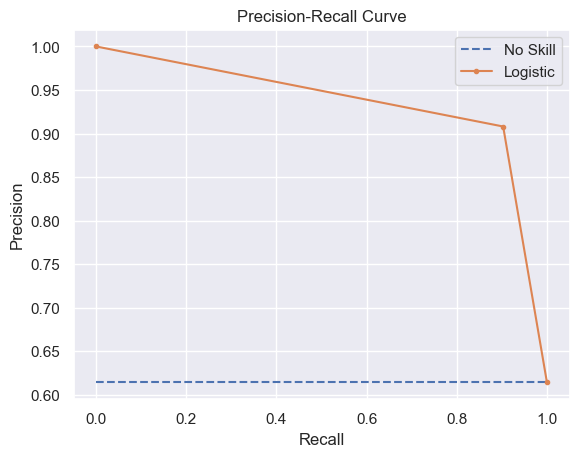


**Figure 3.** Mean area of breast cancer cell (no units provided in data)

Looking at figure 2, one can see that the average mean radius with the most count is between 10 and 15, this indicates that this could be the average mean radius for detection of breast cancer cells. Proceeding to look at figure 3, one can see that the mean area’s count and mean radius’ count are approximately the same count, therefore indicating they are directly correlated.

**Algorithms and Machine Learning**

Prior to choosing a model, the data was split into training set and test set. Once data was split, there were two models that were used, that would show the best prediction: Random Forest and Logistic Regression. For Random Forest, the score model was 99% for the test and as for Logistic Regression, the accuracy was approximately 93% precision call for the training set. With the results, Logistic Regression was the best option. Below is the graph for the Logistic Regression precision curve



**Figure 4.** Logistic Regression precision curve.

**Predictions & Future Improvements**

Based on the results of all of the figures above, one can determine a way to detect breast cancer by looking at both the mean radius and mean area. The larger the radius, the larger the mean area, and more likely that breast cancer can be detected. With that being said, it is best that Logistic regression be used, because the precision score on the training set is close to that of the test set.

Although the data and logistic regression and random forest produced great results, in the future, I would like to explore other ML algorithms such as: Linear Regression, Gradient Boosting, and K Nearest Neighbor. Trying these three different models in addition to the ones tried in this project, to have both a variety and accurate models in detecting breast cancer.