Breast Cancer Capstone Project

I. Project Proposal and Problem Statement

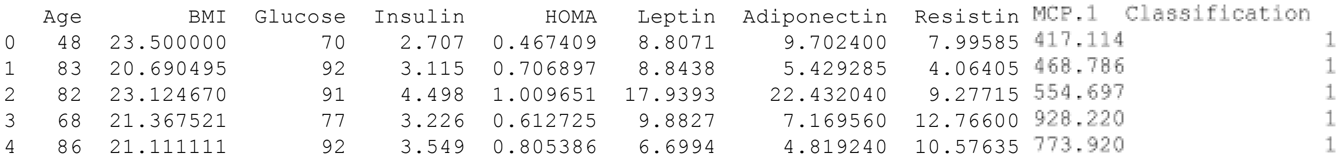
Breast cancer is one of the most common cancers in the United States with invasive breast cancer affecting an estimated 12% of women and has the second-highest mortality of any cancer that affects women (Breastcancer.org, 2019). However, death rates vary widely based on time of detection with 90% of patients surviving if it is detected early and 15% for those diagnosed in late stages (cancerresearch.uk, 2018). Unfortunately, early detection can be difficult as breast cancer is often symptomless in its earliest stages (Webmd, 2019). Regular physician check-ups and radiology scans for all women is not a viable solution as average costs for diagnostics are $400 for mammograms, $130 for ultrasounds, and over $2000 for biopsies. Thus, 53% of surveyed women do not have regular screenings due to financial barriers (McAlearney, 2007).

Research to determine the most common risk factors for breast cancer can alleviate some of these concerns as healthcare and charitable organizations would be able to determine more high-risk patients. Charities such as the National Breast Cancer Foundation can use the information to understand which patients need more immediate or regular access to screenings. Research organizations, hospitals, and pharmaceutical companies may find the information useful for a better understanding of the cause of breast cancer and possible treatments. For example, if a higher level of leptin appears to correlate with a higher risk of cancer, researchers can focus on whether or not there is a potential relationship between the protein level and cancer. Also, if a certain trend such as higher BMIs is shown to correlate with a higher incidence of breast cancer, physicians may choose to have those patients do more regular screenings or can advise patients that the screening may be worthwhile because they are of higher risk.

In order to find these risk factors, I will be analyzing the Breast Cancer Coimbra Data Set. This set has data for both breast cancer and non-cancer patients and has a variety of characteristics that can be identified easily with a blood test. Factors include glucose, insulin, leptin, BMI, age, etc. The data currently do not have any missing values, so it is already relatively clean and easy to use. I will start by making plots between the measurements and the number of patients that have breast cancer. This will give me a rough outline of potential correlations and if a relationship is present, can help show whether it is linear, polynomial, etc. Once I can find the most appropriate model for the characteristics with the most promising correlation, I can determine which algorithm to use to better quantify this trend such as linear regression and may use factor analysis to see which attributes are more useful. Once, I understand the common risk factors that can lead to breast cancer, I will summarize my findings in a PowerPoint Presentation. The presentation will include more information on the dangers of breast cancer and the usefulness of the experiment. Plots such as 2d color plots or box plots will also be present in the presentation based on which will convey the statistical analysis more appropriately. This will be the first of many data science challenges I will undertake and will be an exciting opportunity to bridge my previous experiences in research, healthcare, and cancer with my new career in data science.

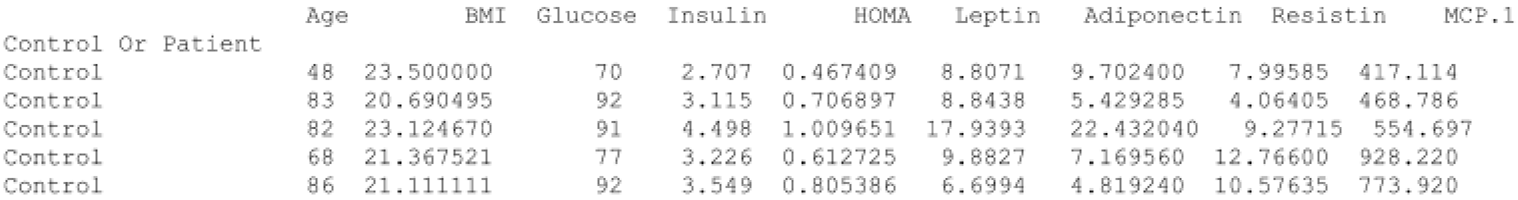
II. Data Cleaning and Wrangling

Raw Data Example- Below is a 5 row sample of the original, unmodified data from the original source. A 5 row sample will be used for every sample in this document.



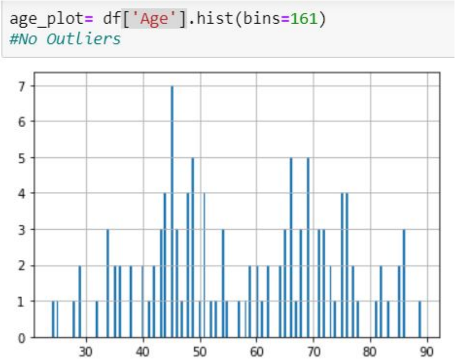
```
   Age        BMI  Glucose  Insulin      HOMA   Leptin  Adiponectin  Resistin MCP.1  Classification
0   48  23.500000       70    2.707  0.467409   8.8071     9.702400   7.99585 417.114              1
1   83  20.690495       92    3.115  0.706897   8.8438     5.429285   4.06405 468.786              1
2   82  23.124670       91    4.498  1.009651  17.9393    22.432040   9.27715 554.697              1
3   68  21.367521       77    3.226  0.612725   9.8827     7.169560  12.76600 928.220              1
4   86  21.111111       92    3.549  0.805386   6.6994     4.819240  10.57635 773.920              1
```

Data Indexed by Classification and Index renamed to Control/Patient

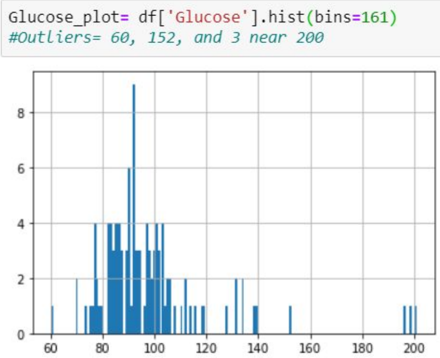


```
                    Age        BMI  Glucose  Insulin      HOMA   Leptin  Adiponectin  Resistin    MCP.1
Control Or Patient
Control              48  23.500000       70    2.707  0.467409   8.8071     9.702400   7.99585  417.114
Control              83  20.690495       92    3.115  0.706897   8.8438     5.429285   4.06405  468.786
Control              82  23.124670       91    4.498  1.009651  17.9393    22.432040   9.27715  554.697
Control              68  21.367521       77    3.226  0.612725   9.8827     7.169560  12.76600  928.220
Control              86  21.111111       92    3.549  0.805386   6.6994     4.819240  10.57635  773.920
```

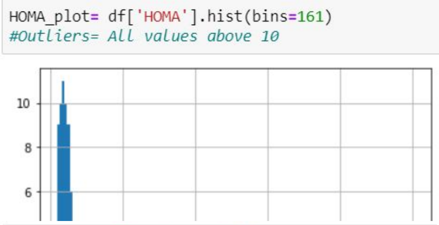
Outliers were determined by plotting histograms for each variable and are shown below. Outliers will not be used for calculations and will be omitted when looking for relationships.



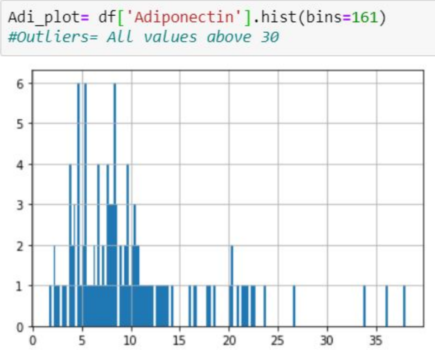
```
age_plot= df['Age'].hist(bins=161)
#No Outliers
```



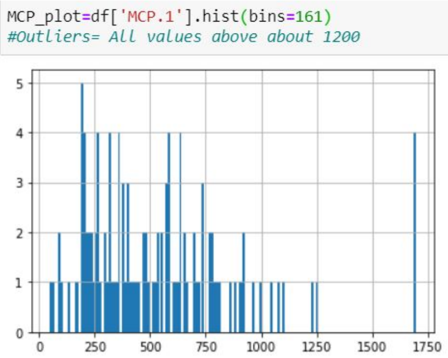
```
Glucose_plot= df['Glucose'].hist(bins=161)
#Outliers= 60, 152, and 3 near 200
```



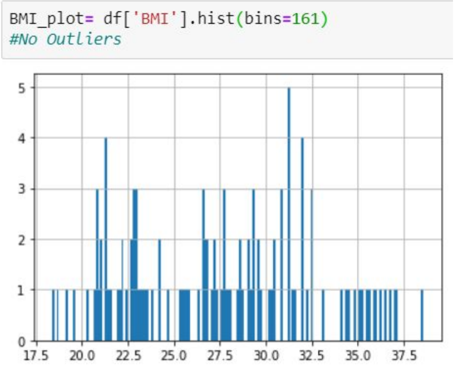
```
HOMA_plot= df['HOMA'].hist(bins=161)
#Outliers= All values above 10
```



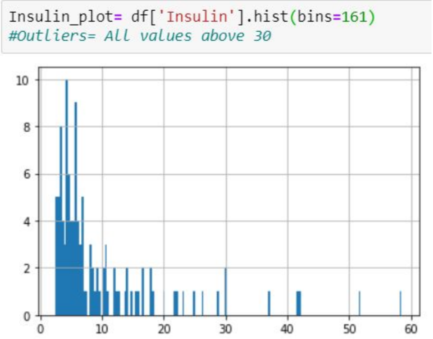
```
Adi_plot= df['Adiponectin'].hist(bins=161)
#Outliers= All values above 30
```



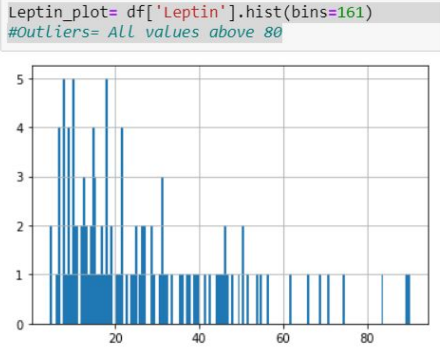
```
MCP_plot=df['MCP.1'].hist(bins=161)
#Outliers= All values above about 1200
```



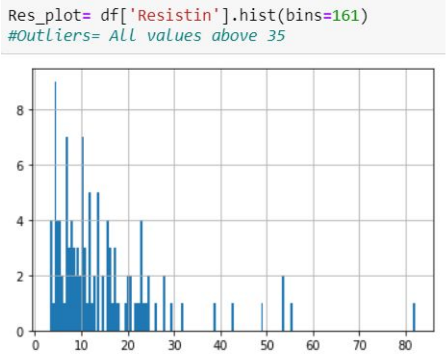
```
BMI_plot= df['BMI'].hist(bins=161)
#No Outliers
```



```
Insulin_plot= df['Insulin'].hist(bins=161)
#Outliers= All values above 30
```



```
Leptin_plot= df['Leptin'].hist(bins=161)
#Outliers= All values above 80
```



```
Res_plot= df['Resistin'].hist(bins=161)
#Outliers= All values above 35
```

Means of Variables Grouped by Control/Patient- This will gi

ve me a rough estimate of whether a relationship between variables is likely

```
                           Age        BMI     Glucose    Insulin      HOMA
Control Or Patient

Control              58.076923  28.317336   88.230769   6.933769  1.552398
Patient              56.671875  26.984740  105.562500  12.513219  3.623342

                      Leptin  Adiponectin   Resistin       MCP.1
   Control          26.637933    10.328205  11.614813  499.730692
   Patient          26.596512    10.061167  17.253777  563.016500
```

III. Data Storytelling:

In order to be able to determine what possible data stories I would be able to tell with the data, I wanted to limit my variables to those with the most promise of having a clear difference between the control and patient lab values. In order to do this, I grouped the data into the categories of patient and control and compared the averages of both groups for each variable. I then limited my research to 5 variables that had what looked like the most significant differences: glucose, insulin, HOMA, resistin, and MCP.1 protein levels. For each of the variables, I created box plots to eliminate outliers and made scatter and bar plots to illustrate the differences between the control and patient numbers.

For all of the variables, I noticed similar patterns for the controls and patients. The controls would have lower and less varied levels that are closer together. The values would be similar on the lower spectrum of the values, but the patients would also have an extremely high value with points in between. It is also significant to remember that insulin, glucose, HOMA, and resistin are closely related which would explain their similar trends. Insulin is a hormone released in order to help the body use glucose, HOMA (homeostatic model assessment) is an index used to measure insulin resistance and the effectiveness of β cells in the pancreas that release insulin, and resistin is a hormone derived from adipose that has been linked with causing insulin resistance.

Based on their functions, these four variables are directly related and if one of them is increased, it would make sense that the other three are increased as well. The relationship between glucose and cancer has been studied before and strong relationships between cancer and diabetes have been found. In fact, patients with type 2 diabetes are twice as likely to develop liver or pancreatic cancer. In fact, some researchers believe that the relationship may be strong enough to perhaps use glucose levels of tissues to find tumors in the body. Cancer patients tend to have higher glucose levels due to cancer cells needing more energy due to their rapid spread and reproduction rates. In fact, cancer cells can use up to 200 times more glucose than control cells (Rosenberg, 2019).

If cancer cells are using more glucose, then the glucose levels of patients will be higher, they will release more insulin to try to mobilize this glucose to cells, the HOMA index will measure the higher levels of insulin resistance as the body tries to prevent the insulin from causing organ failure, and resistin will also be released to increase insulin resistance.

MCP.1 protein is a protein that helps the immune system respond to tissue injury. The higher levels seen in cancer patients compared to the control makes sense since the body will detect the growing number of cancer cells and will attempt to make an immune response against them.

Currently, the plots have illustrated an interesting story that can be somewhat explained or reasoned with a simple understanding of biology and metabolism and the immune response. Now that the trends have been visualized, I look forward to being able to more precisely measure these differences using statistics and continuing to research these trends to better understand how breast cancer can be more easily detected.

IV. Statistical Analysis:

Based on the Data Storytelling assignment that I did for the Breast Cancer data, I was able to limit the variables down to 5. I selected those 5 variables based on their differences in averages and based on box plots and other visuals that displayed the data. I also was able to remove outliers from the data in order to further clean it. Next, I wanted to determine whether the differences were statistically significant.

Firstly, I converted the “Classification” data from 1s and 2s into 1s and 0s in order to more accurately reflect the difference between cancer and control patients numerically. The 1s represented cancer patients while the 0s represented controls. Next, I created a “pearson_r” function to calculate the Pearson correlation between increases in the lab values and cancer diagnosis (the 1’s in the Classification data). The Pearson correlation showed a small association between cancer diagnosis and resistin, HOMA, and insulin and showed a moderate association between cancer diagnosis and glucose. It found no relationship between MCP.1 and cancer.

The other statistical test I ran on the data was finding T-tests for each of the potentially significant values. It gave extremely low P values when comparing glucose, insulin, resistin, and HOMA lab values when comparing between data of control and cancer groups. This helps support the notion that since there is a statistically significant difference between these two groups, that the differences in lab values between cancer and control patients is very unlikely due to trial and error. Therefore, it is more likely to be the result of the cancer diagnosis. However, the P-value for comparing MCP.1 values between the control and cancer groups was much higher at around 0.72 which makes it appear that differences between the groups is likely due to trial and error.

Overall, the statistical analysis helped support the hypothesis that lab values for glucose, insulin, HOMA, and resistant are significantly higher for cancer patients than control. The T-test confirmed that there was a statistically significant difference between both groups and the Pearson correlation found a small to moderate association between cancer diagnosis and increases for the four lab values. However, both tests confirmed that MCP.1 and cancer diagnosis likely do not have an association or correlation. These relationships will be further explored using machine learning analysis as well.

V. Machine Learning

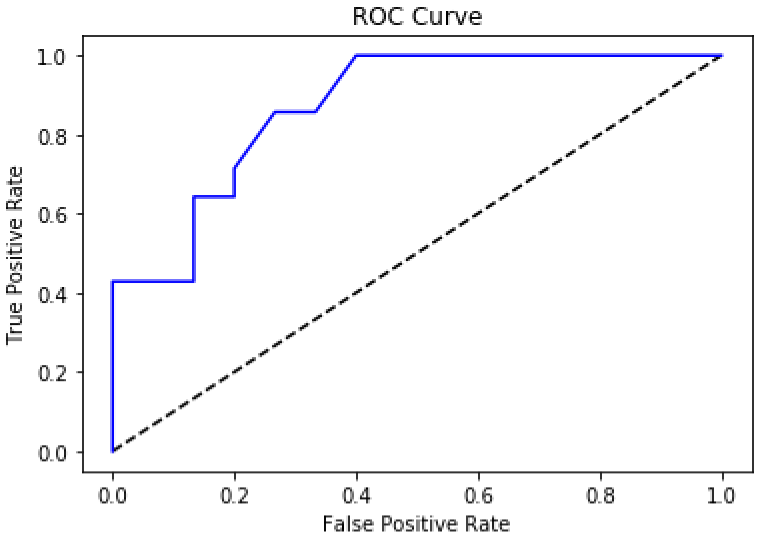
The final segment of my capstone project will be using machine learning to further analyze the trends in my data. It has so far been confirmed that there is a correlation between higher glucose, insulin, HOMA, and resistin values for cancer patients versus control patients both statistically and visually. However, machine learning will help us use this data to map out predicted outcomes for new data rather than only relying on the previous data we were given.

There are various different methods of machine learning that exist, however, our data is most suitable for the supervised machine learning method called classification. Classification is most appropriate because we have two different categories already specified before our analysis began: cancer patients and control patients.

The "y" was set equal to the Classification column of our data table so that the patient's status as a control or cancer patient could be the independent variable. Then, the columns that were found to be statistically and visually not related to cancer were dropped leaving only glucose, HOMA, insulin, and resistin blood values.

Then, the datasets that were left were split into training and testing subsets. The training subset is designed to create the machine learning model and fit it to the data while the testing dataset will provide predictive information. K nearest neighbor clustering was the method of supervised learning that was used to create our model. Then, the model was fit to the training data and then applied to the variable X-test. After, the model was fit to X-test, the predict function was used to produce test data based on the model.

Now that the model was created and applied to the data, I needed to test how accurate the model was at being able to determine whether the test data was from a cancer or control patient. First, an ROC curve produced below. It can be interpreted a few ways, but most importantly the curve is shown to the left of the dotted line. Anything to the right of the dotted line is considered insignificant or not related, however, since the curve's data points is to the left of the data, the model appears to be useful.



Also, a confusion matrix was produced as below. The top left is the True Positive, the top right is False Positive, bottom left is False Negative, and bottom right is True Negative. The matrix shows how accurate the model is based on how many times the model creates data that is correctly classified as one of the 4 categories.

```
[13  2]

 [ 5  9]
```

The confusion matrix can be quantified into a number that tells us how accurate the model is called ROC-AUC score. This number is based on the area under the curve and was calculated from our model as 0.75. Since it is above 0.5, it is considered fairly accurate. A score of 0.75 is considered an appropriate and dependable model in only certain cases depending on the subject being studied. The medical community usually strives for an accuracy of at least a 0.95 ROC AUC score meaning that our model is not accurate enough to be used for medicinal testing.

VI. Conclusion

The purpose of this experiment was to determine how useful certain lab results can be for cancer diagnosis. It was to see if physicians could look at certain lab values such as glucose and determine whether or not patients were more likely to have a breast cancer diagnosis. The project showed that there was an interesting promise to the idea that higher glucose, resistin, HOMA, and insulin levels may be indicators for higher likelihood of breast cancer.

This trend was seen visually through data tables and graphs and was proven to be statistically significant. A supervised machine learning model was also shown to be fairly accurate for categorizing patients as cancer or control based simply on these lab values. However, it was not considered a strong enough model for the medical community. Therefore, despite these promising trends, a larger dataset and additional research will be necessary to further explore this trend and relationship. Perhaps adding more subjects, more significant lab values, and/or fine-tuning the existing model can help lead to a method of accurate diagnosis of breast cancer based off of lab values.

Works Cited

Rosenberg, Abe. “Cancer and Diabetes: More Connections Than You Think.” City of Hope Comprehensive Cancer Center, City of Hope, 21 June 2019, www.cityofhope.org/cancer-and-diabetes-more-connections-than-you-think/.

“Breast Cancer Symptoms and Early Warning Signs.” *WebMD*, WebMD, www.webmd.com/breast-cancer/understanding-breast-cancer-symptoms.

McAlearney, Ann Scheck, et al. “Cost as a Barrier to Screening Mammography among Underserved Women.” *Ethnicity & Health*, U.S. National Library of Medicine, Apr. 2007, www.ncbi.nlm.nih.gov/pmc/articles/PMC4465254/.

Nbcf. “Information, Awareness & Donations :: The National Breast Cancer Foundation.” *Www.nationalbreastcancer.org*, www.nationalbreastcancer.org/.

“U.S. Breast Cancer Statistics.” *Breastcancer.org*, 13 Feb. 2019, www.breastcancer.org/symptoms/understand_bc/statistics.

“Why Is Early Diagnosis Important?” *Cancer Research UK*, 4 Mar. 2019, www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important.