**Group 7 BA305 Progress Report: Breast Cancer Classification**

**Our Idea & Dataset:**

Our team is working with a breast cancer classification dataset, classifying tumors into malignant or benign along with 10 different characteristics/parameters (radius, texture, area, etc.). The y variable is the Diagnosis (M or B) while the x variables are the 10 different characteristics provided in the dataset. Our goal is to run multiple analyses on the dataset to explore the relationship between different variables and the diagnosis of the tumor.

[Dataset Link](https://www.kaggle.com/yasserh/breast-cancer-dataset)

**Preliminary Stories**

We will conduct a PCA analysis to explore which principal components/variables provided in the dataset are most significant in classifying a breast cancer tumor into benign or malignant.

We will first create a heatmap between the variables to have an overview of how correlated variables are. Since there are several linearly dependent variables, we choose a relatively large number of principal components for PCA analysis, which is 12, in order to avoid missing important features. Then we will scale our data frame to zero mean for preprocessing of PCA analysis. Then we print the eigenvalues in descending order, which indicates how important the corresponding features are. With the help of the eigenvalues, we plot them with a threshold and find the number of principal components we need for further analysis.

**Findings**

Through our PCA analysis, we found 6 components that are significant for prediction. Those 6 components explain 91% of the variance in the data. Therefore, we will plan our focus on these components. Through our heatmap, we found that the perimeter, radius, and area are highly correlated with each other. Our next step will be to understand what these components represent by creating a correlation matrix between the components and variables.

**Week 3/1-3/4**

During this week, we worked on completing any preliminary steps before running any models. Specifically, we researched what each x variable (column) in the dataset meant in more general terms in order to better understand the data we are working with. The variables that were researched are as followed:

[Research Link](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.707&rep=rep1&type=pdf)

Additionally, we created a sample preliminary story and ran a PCA analysis to explore the idea. The story and findings can be found in the paragraphs above.

Y-variables:

1) ID number: Different data points (patients)

2) Diagnosis (M = malignant, B = benign): Dependent Variable

X-Variables:

Ten real-valued features are computed for each cell nucleus:

1. Nucleus 2. Radius 3. Texture 4.Perimeter 5.Area 6.Smoothness 6.Compactness 7. Concavity 8. Concave Points 9. Symmetry 10.Fractal dimension