**Classifying Breast Cancer Tumor Categories**

**Business Understanding**

* What problem are you trying to solve, or what question are you trying to answer?

*Breast cancer is the most common cancer, and it is a highly reported cancer type among women worldwide, making it a significant health problem. Tumors can be i.e., Benign (non-cancerous) and Malignant (cancerous). The goal is to build a model that could accurately classify the tumor type.*

* What industry/realm/domain does this apply to?

*Healthcare*

* What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)

*The early diagnosis of Breast cancer can improve the prognosis and chance of survival and accurate classification of benign tumor type can prevent patients undergoing from unnecessary treatments.*

**Data Understanding**

* What data will you collect?

*The dataset consists of several human cell sample records, each of which contains the values of a set of characteristics of the nucleus.*

* Is there a plan for how to get the data (API request, direct download, etc.)?

*Direct Download from Kaggle*

* What are the features you’ll be using in your model?

*There are 10 columns that could be used in the model development. However, I would like to do a feature selection on the dataset and then select the features that will be useful for model building.*

**Data Preparation**

* What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?

1. *Check for missing values and duplicates,*
2. *Encoding categorical variables present in the dataset*
3. *Check for Multicollinearity*
4. *Feature Engineering – Remove highly correlated variables, check if feature scaling of the features is required.*

* What are some of the cleaning/pre-processing challenges for this data?

*Feature selection*

**Modeling**

* What modeling techniques are most appropriate for your problem?

*Supervised/Classification models : Random Forest, KNN, AdaBoost, SVM, Logistic Regression*

* What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)

*Diagnosis (Benign/ Malignant)*

* Is this a regression or classification problem?

*Classification*

**Evaluation**

* What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)?

*Classification report: F1, Accuracy, precision, Recall*

**Tools/Methodologies**

* What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?

*Random Forest, KNN, AdaBoost, SVM, Logistic Regression*