Breast Cancer Risk Prediction
Using Predictive Analysis To Predict Diagnosis of a Tumor

- **Identify the problem**

Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a results of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor.
A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed.

- **Expected outcome**

Given breast cancer results from breast fine needle aspiration (FNA) test (is a quick and simple procedure to perform. Our goal is to build a model that can classify a breast cancer tumor using two training classification:

- 1= Malignant (Cancerous)
- 0= Benign (Not Cancerous)

- **Data Source and Objectives**

The data is taken from the Breast Cancer Wisconsin Center(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). The dataset contains 569 samples of malignant and benign tumor cells. Since the labels in the data are discrete, the predication falls into two categories, (Malignant or benign). In machine learning this is a classification problem.

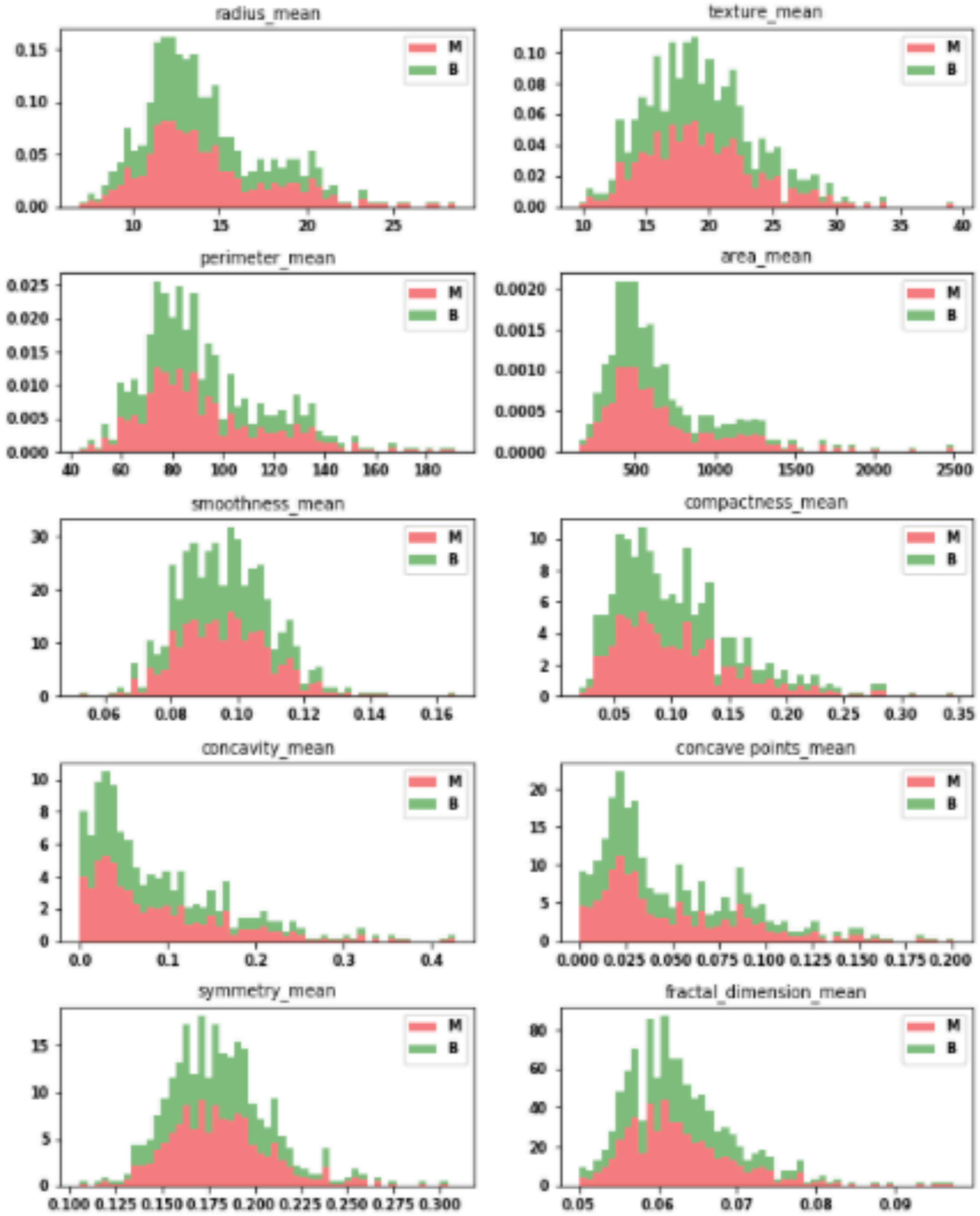
- The first two columns in the dataset store the unique ID numbers of the samples and the corresponding diagnosis (M=malignant, B=benign), respectively.
- The columns 3-32 contain 30 real-value features that have been computed from digitized images of the cell nuclei, which can be used to build a model to predict whether a tumor is benign or malignant.

- **Exploratory Data Analysis**

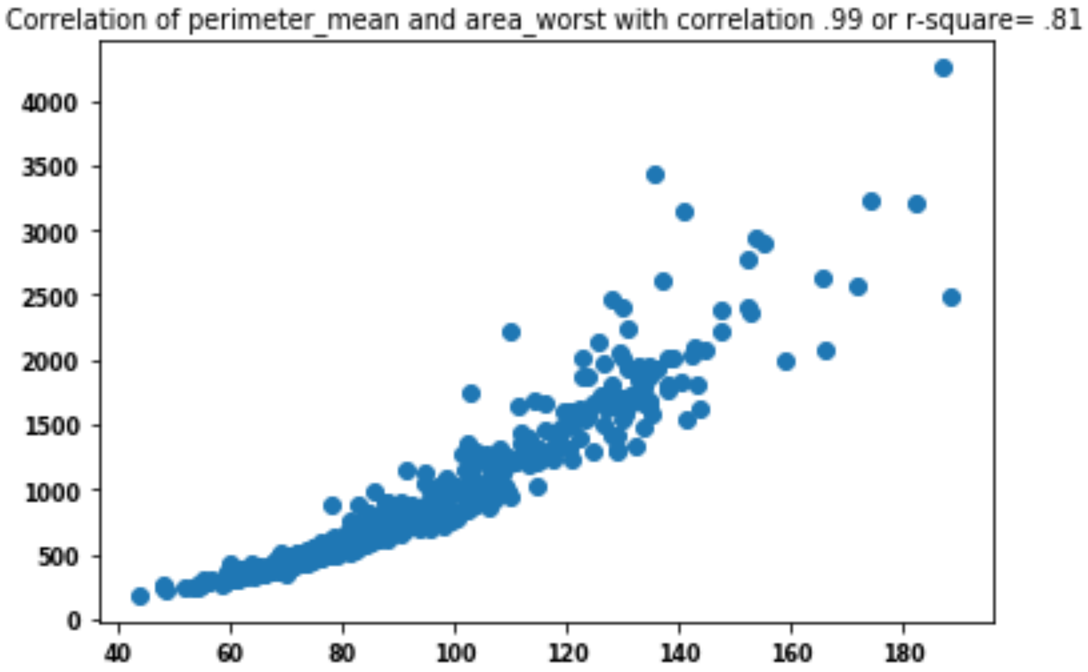
Exploratory data analysis (EDA) is a very important step which takes place after feature engineering and acquiring data and it should be done before any modeling. This is because it is very important for a data scientist to be able to understand the nature of the data without making assumptions. The results of data exploration can be extremely useful in grasping the structure of the data, the distribution of the values, and the presence of extreme values and interrelationships within the data set.

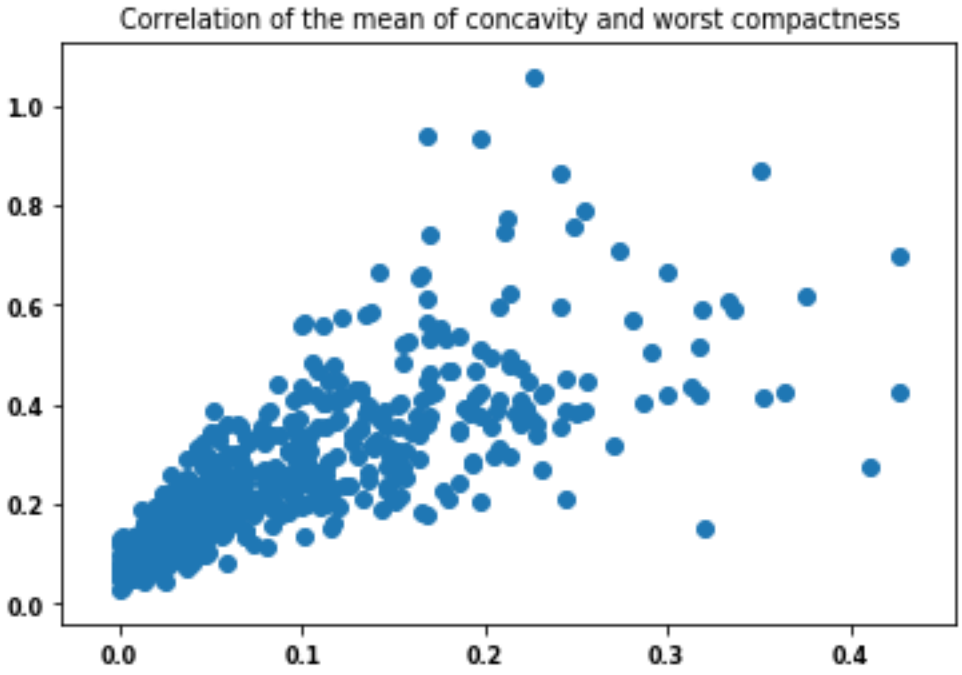
1. **Descriptive statistics** is the process of condensing key characteristics of the data set into simple numeric metrics. Some of the common metrics used are mean, standard deviation, and correlation. In this project, There are 10 main variables;
    1. Radius
    2. Texture
    3. Perimeter
    4. Area
    5. Smoothness
    6. Compactness
    7. Concavity
    8. Symmetry
    9. Concave points
    10. Fractional dimension

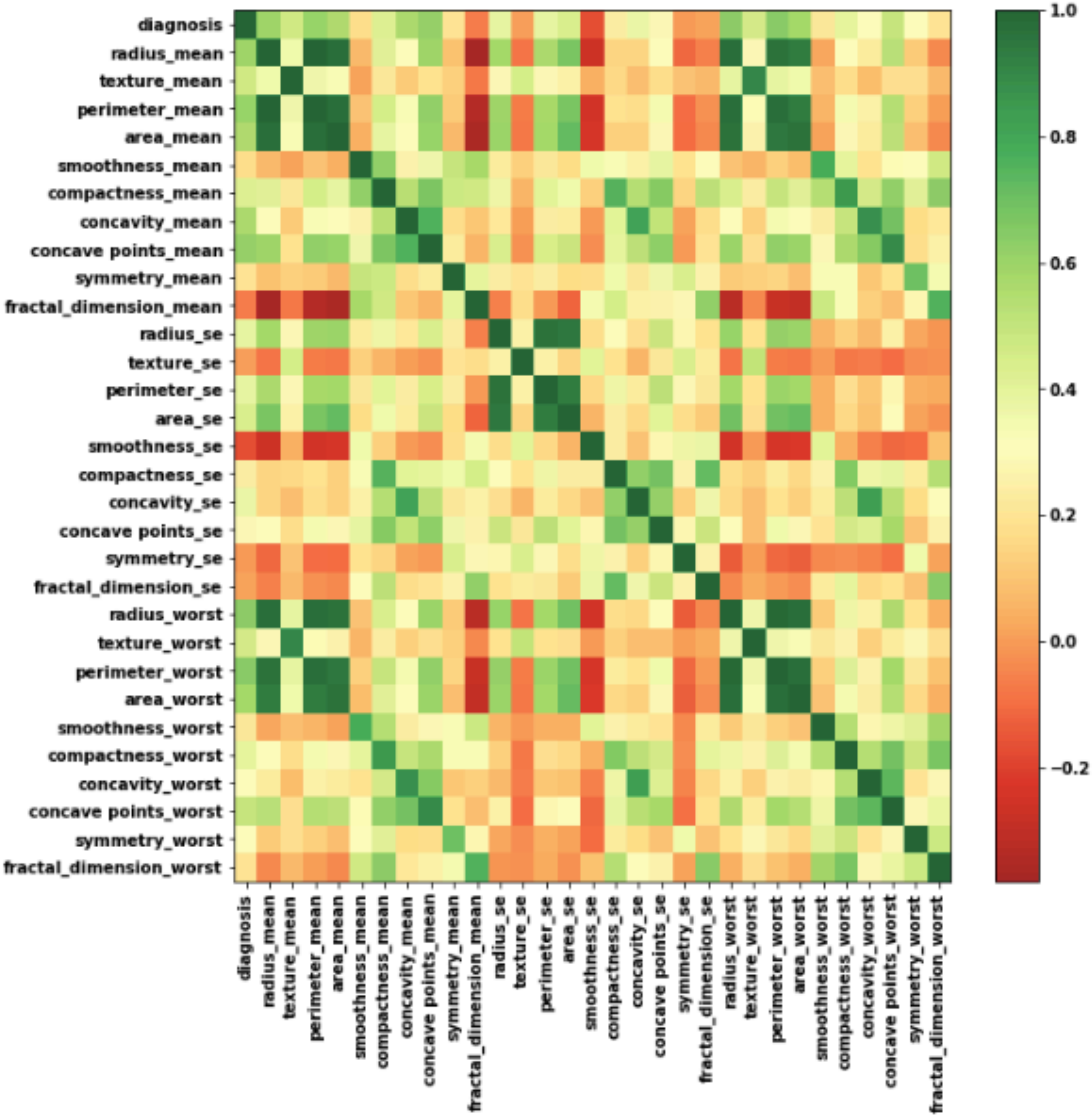
2. **Visualization** is the process of projecting the data, or parts of it into abstract images. In the data mining process, data exploration is leveraged in many different steps including preprocessing, modeling, and interpretation of results.



The graph shows, Malignant cases have higher values in almost every attribute compared to Benign cases. In the next graphs, I chose 2 variable to evaluate the correlation.


Correlation of perimeter_mean and area_worst with correlation .99 or r-square= .81


Correlation of the mean of concavity and worst compactness



- The mean area of the tissue nucleus has a strong positive correlation with mean values of radius and parameter;
- Some parameters are moderately positive correlated, like concavity and area, concavity and perimeter etc
- Likewise, we see some strong negative correlation between fractal_dimension with radius, texture, parameter mean values.

Based on the heatmap, we can going to choose highly correlated features and predict the result using classification model. Although, I've decided not to eliminate any information from the data. I'm going to train the data set into 5 different model and measure time and accuracy, choose the best one. Also, precision and recall are two extremely important model evaluation metrics. While precision refers to the percentage of your results which are relevant, recall refers to the percentage of total relevant results correctly classified by your algorithm. At the end, by creating a pipeline, i'm trying to improve my models and get a better result out of it.

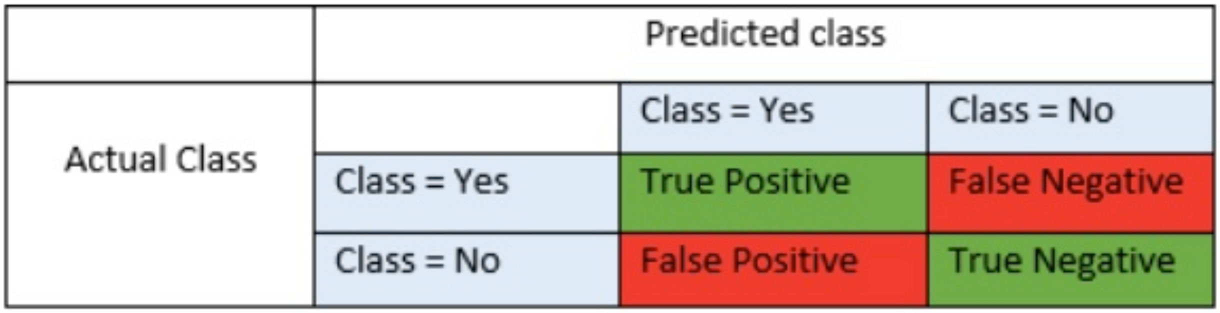
- **Classification**
  - Assigning numerical values to categorical data;
  - Handling missing values; and
  - Normalizing the features (so that features on small scales do not dominate when fitting a model to the data).

- **Classification with cross-validation**

Splitting the data into test and training sets is crucial to avoid overfitting. Cross-validation extends this idea further. Instead of having a single train/test split, we specify K-folds so that the data is divided into similarly folds by similar size.

The results show a similar tight distribution for all classifiers except SVM which is encouraging, suggesting low variance. The poor results for SVM are surprising. It is possible the varied distribution of the attributes may have an effect on the accuracy of algorithms such as SVM. In the next section we will repeat this spot-check with a standardized copy of the training dataset using pipeline.

- **Models**



		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

Recall = TP/TP+FN

In statistical modeling and machine learning, a commonly-reported performance measure of model accuracy for binary classification problems. For this purpose, i've choose, KNN neighbor classifier, Decision tree, Random forest, Gaussian NB, SVM and Ada boost. After the initial training, the most accurate model is Ada boost and second is Random forest. AdaBoost, short for "Adaptive Boosting", is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. As we can see, is takes more time to train the model but the has better accuracy compare to the others. First, I recognized this problem as classification model, and as we went through the process, the salt shows the best models are Ada boost and random forest which are the known for classification problems. Accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally. However, it does not give detailed information regarding its application to the problem. Precision helps when the costs of false positives are high, and Recall helps when the cost of false negatives is high. In this project, we are working on cancer detection problem and a false negative has devastating consequences. Statistics provides us with the formal definitions and the equations to calculate these measures. Data science is about knowing the right tools to use for a job, and often we need to go beyond accuracy when developing classification models. Also, I took another step to get a result of precision and recall from test and train dataset separately to make sure the model is not causing overfitting or under-fitting.