**Introduction**

According to cancer.org, breast cancer (BC) is common cancer among women worldwide. There is a 1 in 8 chance that a woman will develop breast cancer, representing most new cancer cases and cancer-related deaths, making it a significant public health problem in today’s society.

The early diagnosis of BC can significantly improve the prognosis and chance of survival, as it can promote timely clinical treatment. The accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classifying patients into benign or malignant groups is the subjected to extensive research. Data mining and machine learning (ML) is widely recognized as the methodology of choice in breast cancer pattern classification and forecast modeling because of its unique advantages in critical features detection from complex BC datasets.

Classification and data mining methods are effective ways to classify data, especially in the field of clinical analysis, these methods are widely used in diagnosis and analysis to make decisions.

My solution to this problem involves building a model that accurately classifies tumors as Benign or Malignant based on the tumor shape and geometry.

**About Data:**

This is Dataset classifies the Benign and Malignant cells using the description about the cells in the form of columnar attributes. This data was donated by researchers of the University of Wisconsin and includes the measurements from digitized images of fine-needle aspirate of a breast mass.

The data sets is provided with 569 examples of cancer biopsies, each with 32 features. One feature is an identification number, another is the cancer diagnosis and 30 are numeric-valued laboratory measurements.

The diagnosis is coded as

* "M" to indicate malignant
* "B" to indicate benign.

The other 30 numeric measurements comprise the mean, standard error and worst (i.e. largest) value for 10 different characteristics of the digitized cell nuclei

Ten real-valued features are computed for each cell nucleus

* radius
* texture
* perimeter
* area
* smoothness
* compactness
* concavity
* concave points
* symmetry
* fractal dimension

For each image - The mean, standard error and "worst" of these features were computed and this results in 30 features.

For example; the real-vales feature radius is computed as

* radius\_mean
* radius\_se
* radius\_worst

There are no missing values identified in dataset.

Class distribution: 357 benign, 212 malignant

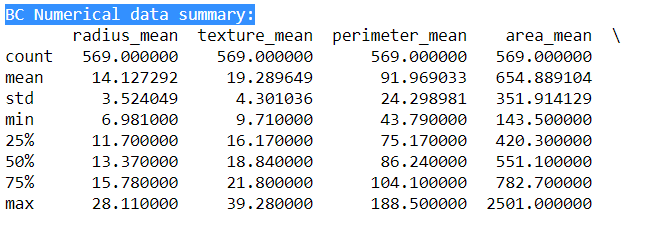
**Part 1 - Graph Analysis**

**Data cleansing steps:**

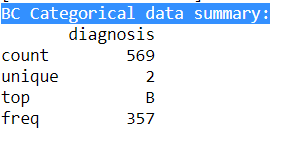
* Dropped ID column – this is not required for analysis.
* Checked not null count. Not null count 569 is matching with RangeIndex 0 to 568. This proves that data does not contain nulls.
* The dimension of the table is: (569, 31)

**Summary of numerical data and categorical data**

**BC Numerical data summary:**

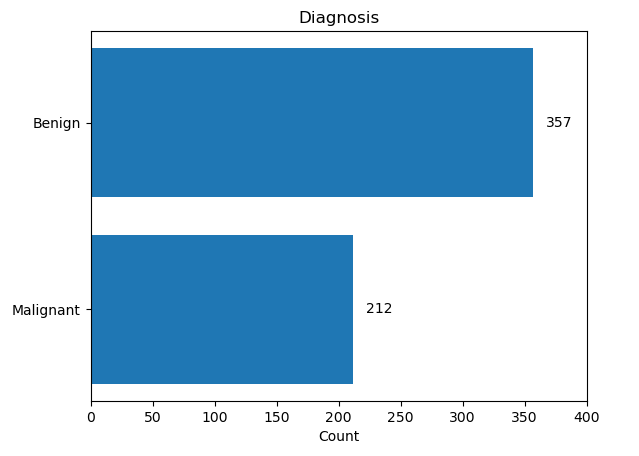


**BC Categorical data summary:**



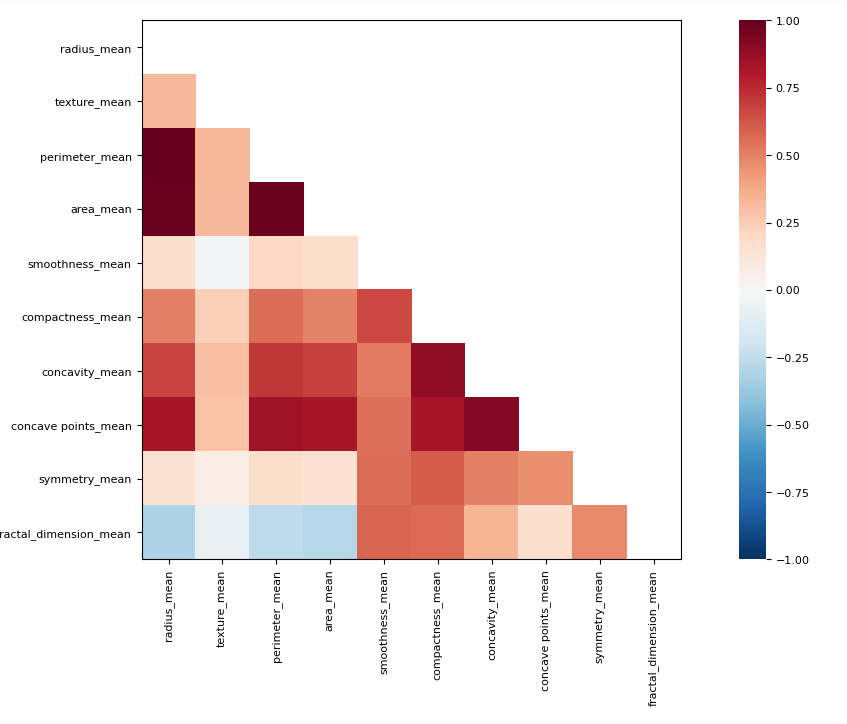
**Bar Chart:**

Below chart represents the break up count of malignant and benign



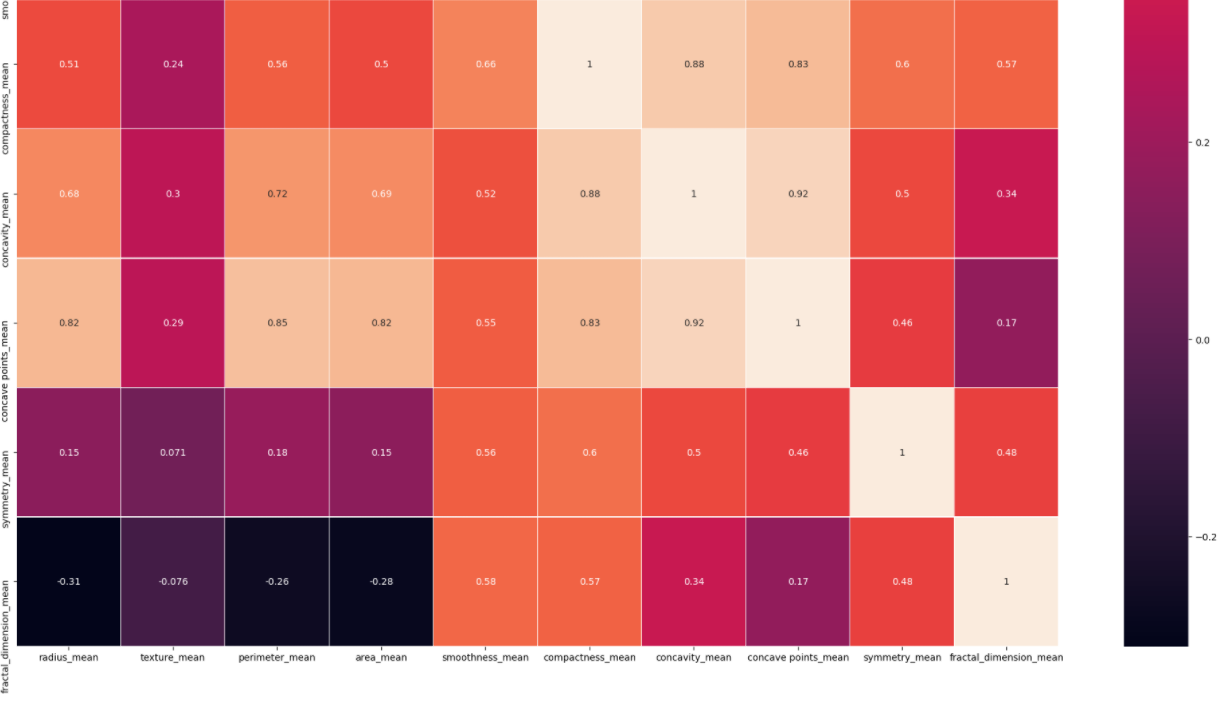
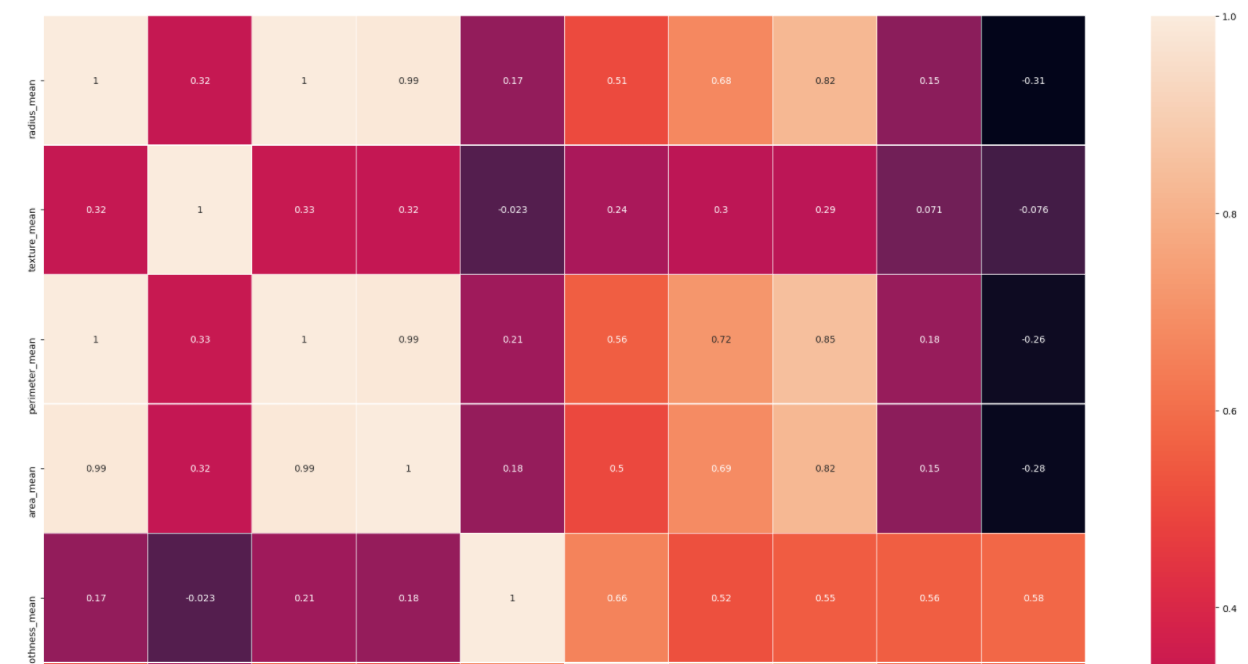
The dataset contains a class imbalance between the malignant and benign cases. Out of the 569 cases in the dataset, 62%(count: 357) cases are benign cases of cancer whereas 37% (count 212)cases are malignant.

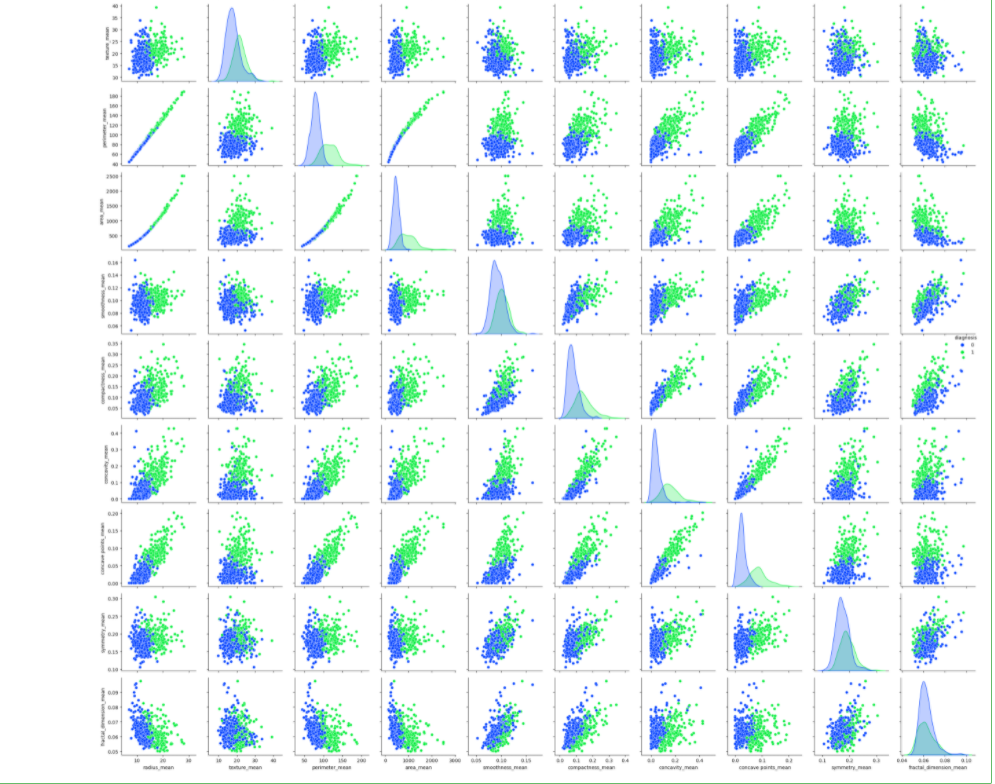
**Correlation plot** is demonstrated using Pearson Ranking, Heat Map and pair plot:

An important step in the graph data analysis step is to identify if at all there is any correlation between any of the 32 variables. First, I have considered 8 mean variables; by using Pearson’s correlation, I have identified that the positive correlation between two variables is demonstrated through the darkness of the blue color, i.e., the darker the blue-colored box, the stronger is the positive correlation between respective variables. Similarly, a negative correlation between two variables is demonstrated through the darkness of the orange color, i.e., the darker the orange-colored box, the stronger is the negative correlation between respective variables.

**Heatmap:**

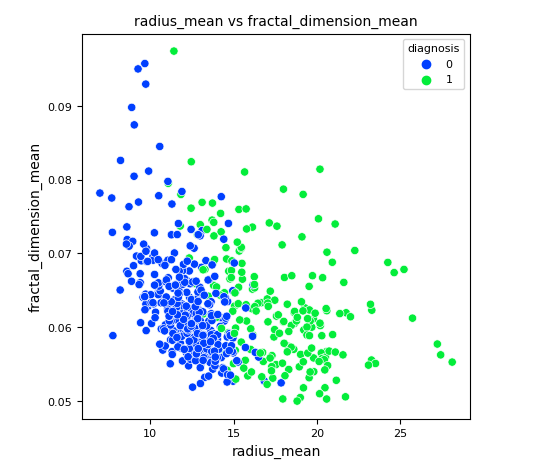
The positive correlation between two variables is demonstrated through the darker black shade color, i.e., the darker the black shade, the more substantial the positive correlation between respective variables. For example, radius\_mean vs. fractal\_dimension\_mean have a positive correlation. Similarly, the negative correlation between two variables is demonstrated through the lighter orange, i.e., lighter the colored box. More substantial is the negative correlation between respective variables, such as radius\_mean vs. perimeter\_mean.



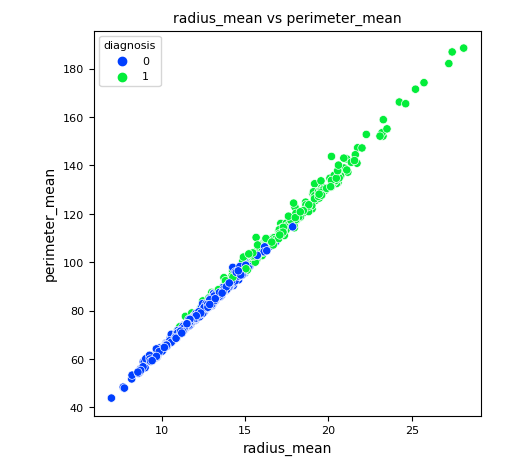
**Pair plot**: with the help of below plot; I tried to visualize the relationships between variables wr.t Class distribution malignant and benign 

**Scatter plot**: From the pair plot, heatmap and Pearson Ranking it’s been observed that the highly correlated pairs are:

Positive correlation: radius\_mean vs fractal\_dimension\_mean

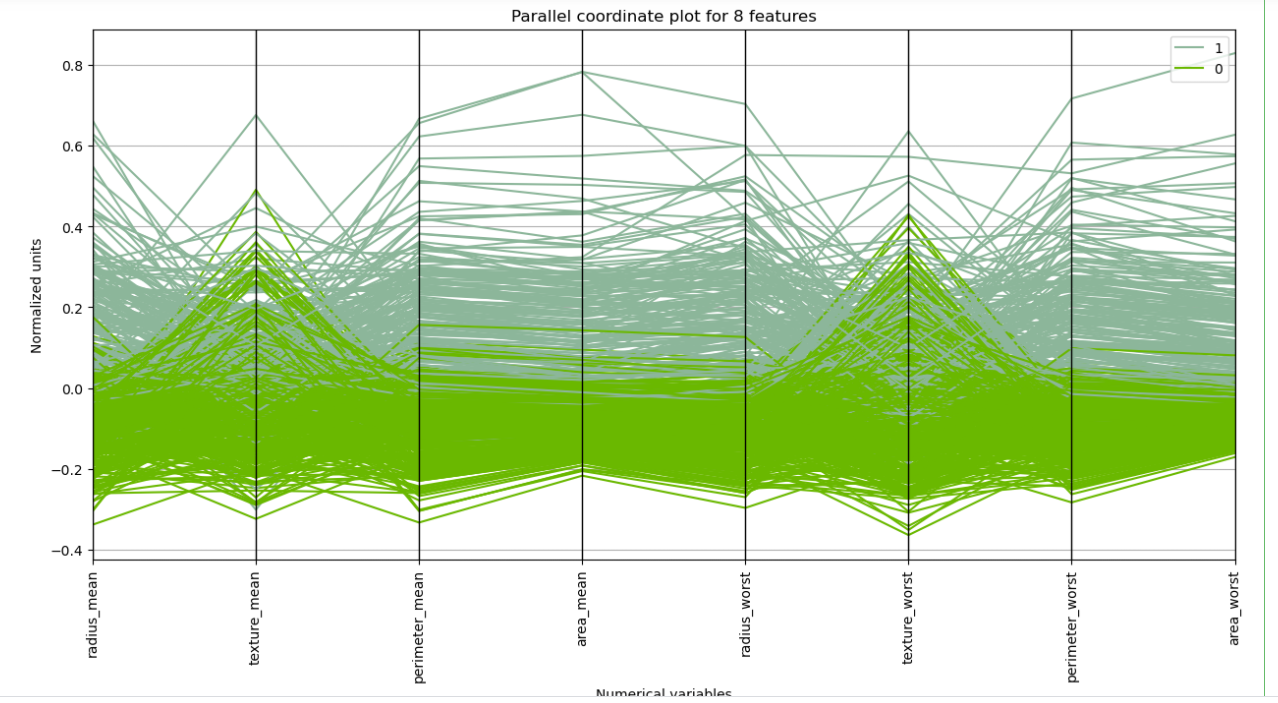


Negative correlation: radius\_mean vs perimeter\_mean



**Parallel coordinates:**

I tried to plot multivariate numerical data using Parallel coordinates. I tried comparing the variables mean vs. worst in plotting, but the downside of this graph is that the visualization seems to be a bit cluttered. Suppose we observe the pattern at texture\_mean and texture\_worst. But this plot does support what was already observed in the above plots.



**References:**

The Data Visualisation Catalogue, . (2020). *Parallel Coordinates Plot*. Parallel Coordinates Plot - Learn about this chart and tools. <https://datavizcatalogue.com/methods/parallel_coordinates.html>.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.