**Breast Cancer Classification – Binomial Classification**

1. **Problem Statement:**

Classify the sample into malignant tumor (0) or Benign tumor (1) based on the various features of the cell.

1. **Import Data & Review**

* **The data has 569 samples & 31 features along with target**
* **There are no missing values & No categorical values**
* **Among 31 feaures , 10 feaures are Highly correlated with other. so they were removed from analysis**

1. **EDA**

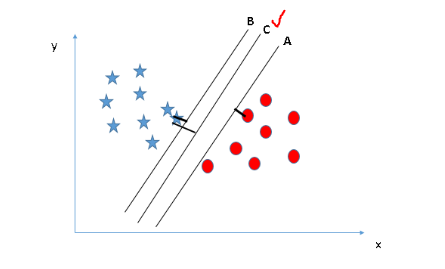
* **From pairplot it is clearly visible thay malignent cells have higher "mean\_area, mean\_compactness, mean\_concave\_points" compared to Benign tumors\***
* **Nearly all the mean features shows good correlation among each other and analysis w.r.t to target**

1. **Machine Learning model**

* **Split the train\_test with stratify with target**
* **Categorical Encoding is not required as all the independent variables are numerical**
* **Both Malgnent & Benign tumors are correctly distributed across train & test datasets**

**Support Vector Machine (SVM)**

**SVM divides the Malgnent & Benign with a hyperplane(Line) with max margin ( max distance from the closest point to hyperplane)**



**Accuracy of the raw model comes up to 64%**

1. **Improving the Model - Feature scaling**

* **As all the parameters are of different scale . we need to convert them to single scale. Apply MinMaxscaler to train & test.**
* **Fit the scaled parameters vs target & evaluate**
* **Feature scaling the accuracy of the model was increased from 64% to 94%.**

1. **Further Improving the Model - Parameter tuning**

* **Above models were trained with default parameters . But we train it with best parameters. And this the best parameters are identified by grid search.**
* **Important parameters of SVM are kernel, gamma, C**

**kernel:  “linear” - linear hyper-plane, “rbf” and “poly” are useful for non-linear hyper-plane . (default value is “rbf”).**

**gamma: Used only when kernel is ‘rbf’, ‘poly’ and ‘sigmoid’.**

**Higher the value of gamma, will try to exact fit the as per training data set**

**This causes over-fitting problem so, optimal values to be provided.**

**C: Penalty parameter C of the error term. It also controls the trade off between smooth decision boundary and classifying the training points correctly.**

**GridSearchCV (SVC(), param\_grid, refit=True, verbose=4)**

* **# refit helps to refit the model with best paramets**
* **Results : With the Parameter Tuning the accuracy of the model was futher increased from 97%. But Type II error increased compared to earlier. So will check with all features**

1. **Summary**

* The accuracy of the model was increased to 97% with Feature scaling and Paramter Tuning
* There are some Type errors but it constitutes of less of 2% of the total records