Capstone project

Dataset I selected in breast cancer dataset. Because 1 of 8 women is diagnosed with breast cancer in their lifetime. So I wished to create an AI model to predict that.

**Domain** : Machine Learning

**Learning** : Supervised Learning

**Regressor or Classification** : Classification

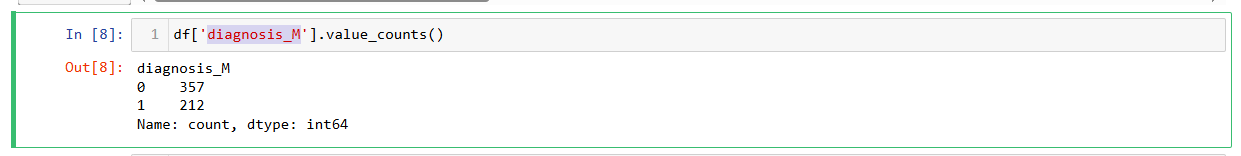


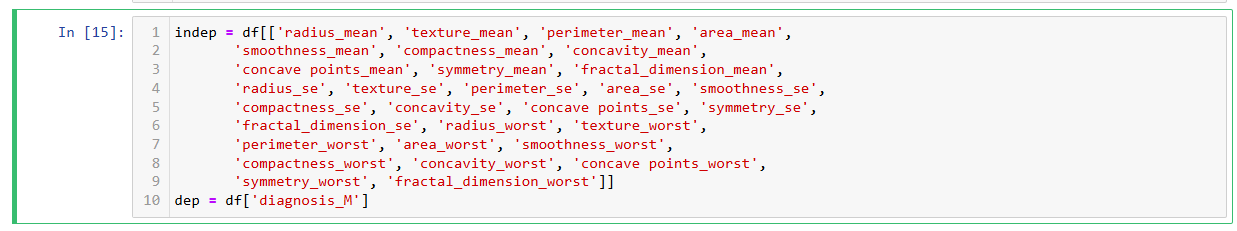
Dataset has 569 rows and 32 columns. Id column is not required for us so I removed it from the Dataset.



So my column count now becomes 31.

Our output column is ‘diagnosis\_M’ and all other columns are our input columns.

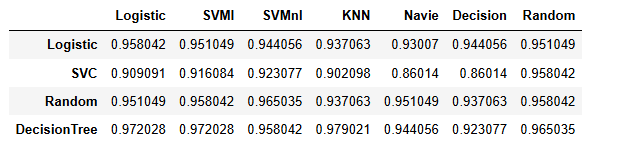




* After this I applied univariate analysis to this dataset for identifying any outliers is present or not.
* Since our input columns are 30, it is little difficult for the user to provide all the values for the input variables. So we applied **feature selection** and **dimentionality reduction** to get the lesser variable count required to create a model to predict breast cancer.

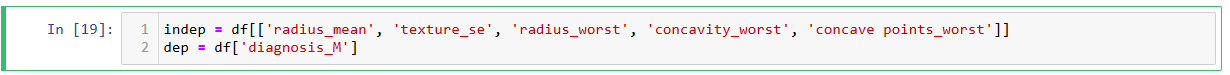




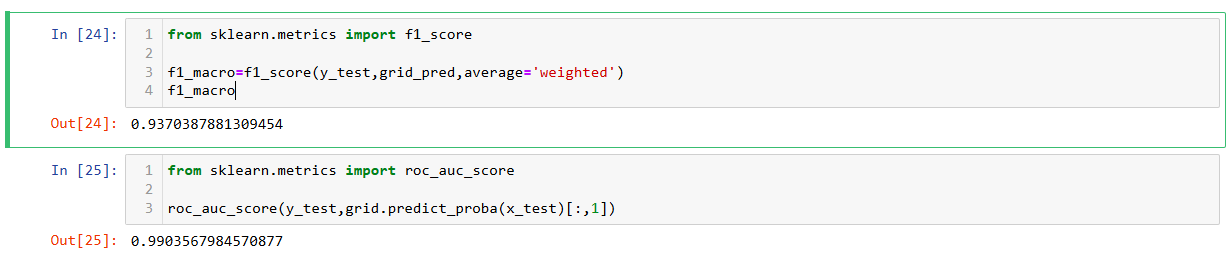


We are getting LogisticRegression model with **95.8% with 5 input columns**.

So now our input and output variable becomes,



We are splitting our df into train and test set and the applied Standard Scaler to it and we created Logistic Regressor model and we got 93.7% **f1\_score** and 99.0% **roc\_auc score**.



* Since we applied Standard Scaler preprocessing to our input variables, we saved our final model and standard scaler also to load these two in our deployment phase.
* We loaded the saved model and saved preprocessing technique in our deployment phase, got the inputs and applied preprocessing to the input and passed those to the model for prediction of Breast cancer Malignant tumor or not.

