**Risk Analysis on Genetic Diseases Using ML**

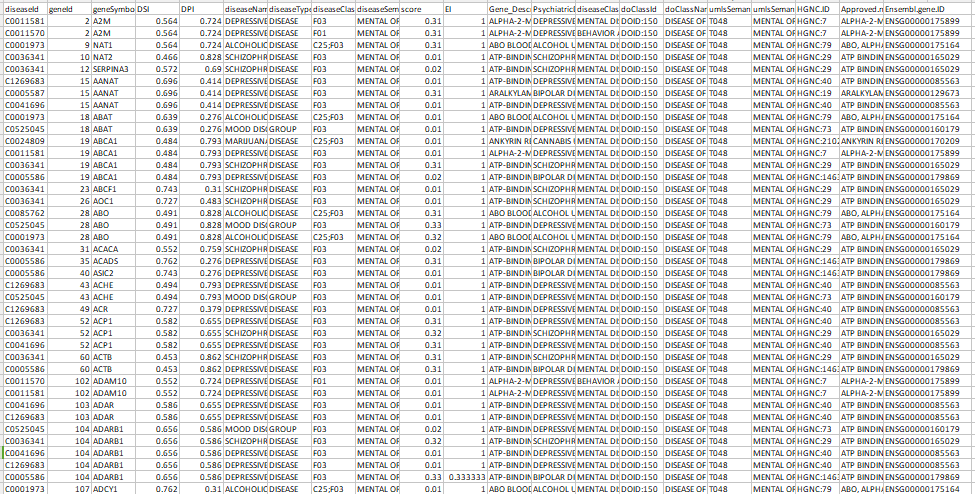
**Software requirements**

**• Programming Language : Python , Jupyter Notebook**

**• Packages : Numpy, Pandas, Matplotlib, Scikit-learn, seaborn.**

**• Tool : Python 3.7**

**• Dataset : Research on Dataset(www.kaggle.com)**



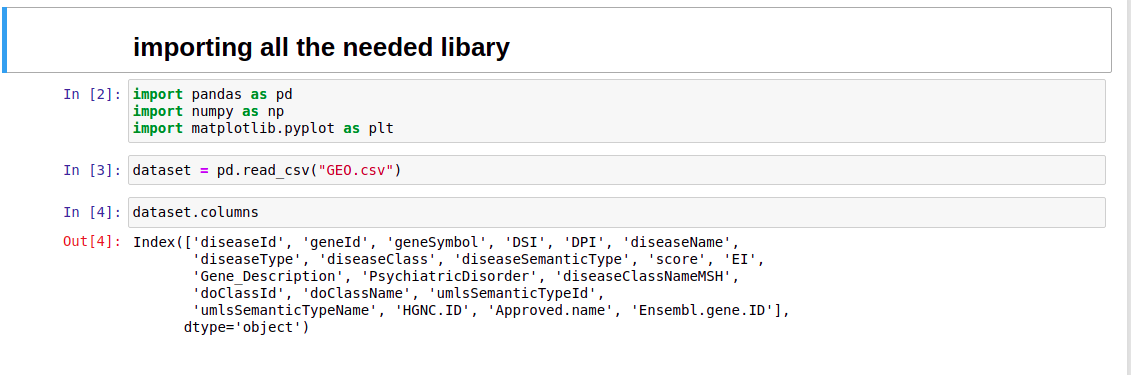
Dataset is collected from https://ghr.nlm.nih.gov/

We are data is collected for Risk Analysis on Genetic Diseases, in our dataset we taken 3766 records for risk analysis and 5397 records for disease prediction.

**Gene\_Id, Gene\_Symbol, Gene\_Description, Disease\_Id, DiseaseName, PsychiatricDisorder, diseaseId, name, type, diseaseClassMSH, diseaseClassNameMSH, hpoClassId, hpoClassName, doClassId, doClassName, umlsSemanticTypeId, umlsSemanticTypeName, Approved symbol, HGNC ID, Approved name, NCBI gene ID, Ensembl gene ID, GENE, TARGET\_NAME, TARGET\_CLASS, SWISSPROT, ACT\_VALUE, ACT\_TYPE, ACT\_SOURCE**

1312, COMT, CATECHOL-O-METHYLTRANSFERASE, C0001723, Affective, dipressive disorder, Group, F03, mental disorders, DOID:150, disorders of mental health, T048, MENTAL OR BEHAVIORAL DYSFUNCTION, HGNC:2228, ENSG00000093010

Above all bold names are the dataset column names these all columns contains values are the dataset values and these data is cleaned for analysis and it is stored in cleanbook.csv file.



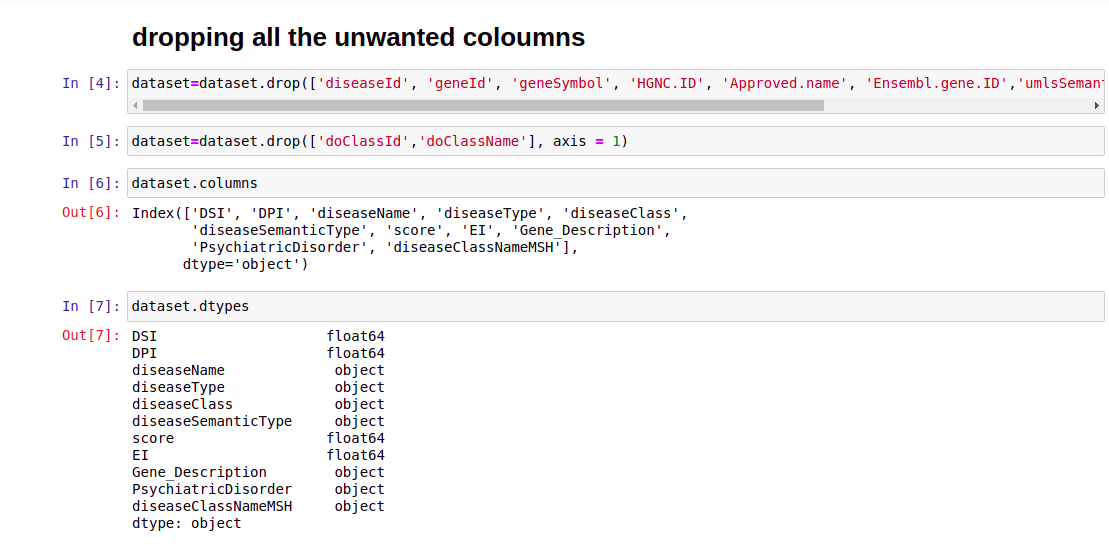
We are importing the required package for the analysis and identifying the Genetic Risk based on the datasets available from the site Kaggle and the dataset has been imported through pandas library and stored in the variable.

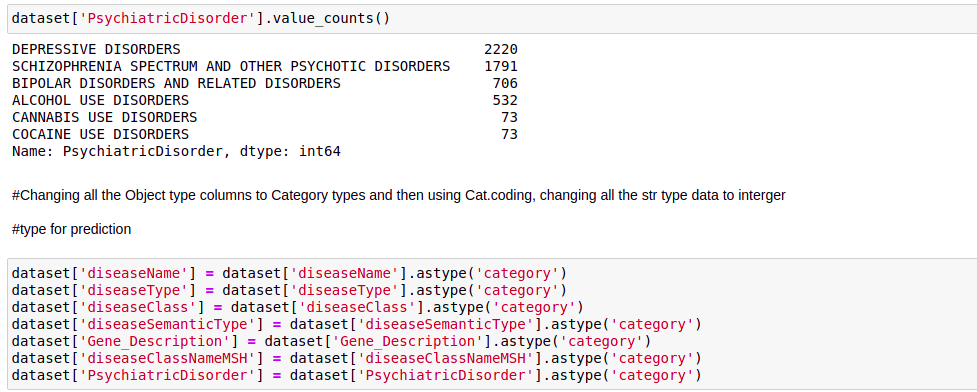
Then data has been extracted

here we divided the protocol from the entire URL. but need it to be divided it seperate column

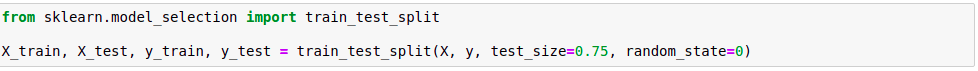
Data Exporting & Cleaning & Data Preprocessing

Dropping the unwanted columns and cleaning the data





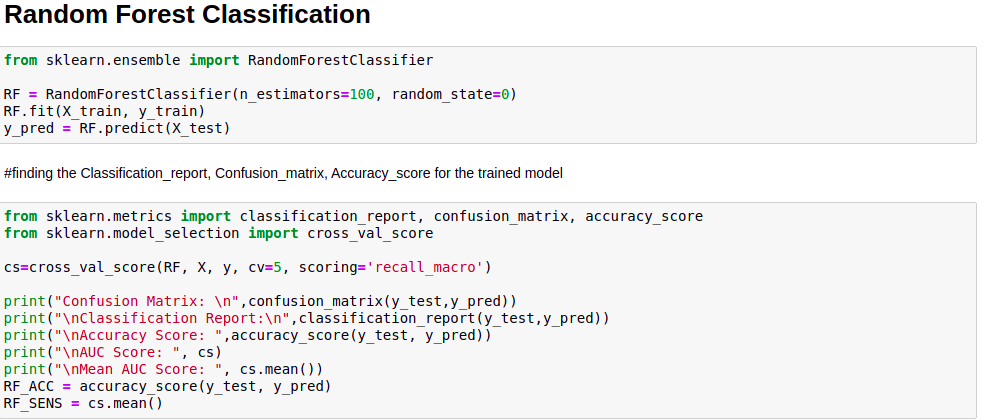
# finding all the types of Category present present in the Target variable



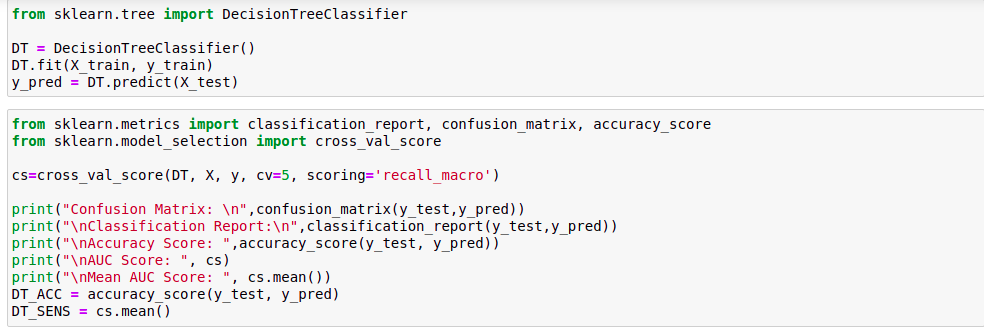
Training & splitting the data

Deploying ML algorithms for evaluating the accuracy value and we choose the better algorithm for prediction the brain status from the given dataset.

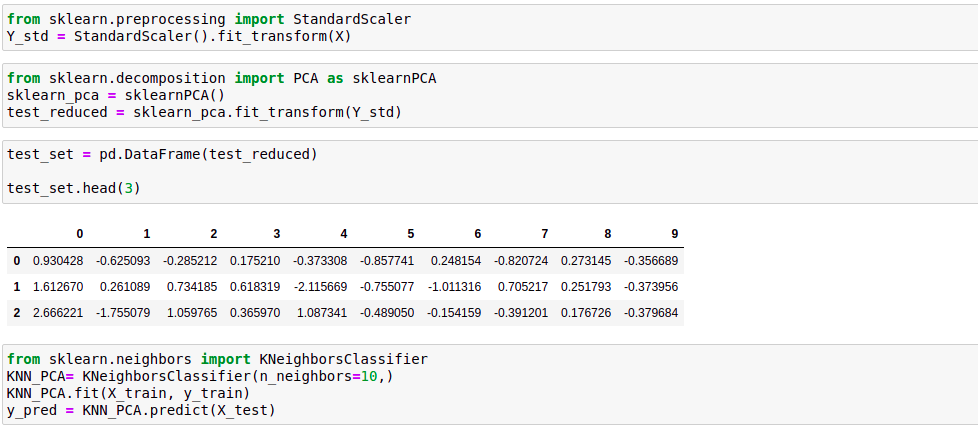
Random Forest



Decision Tree

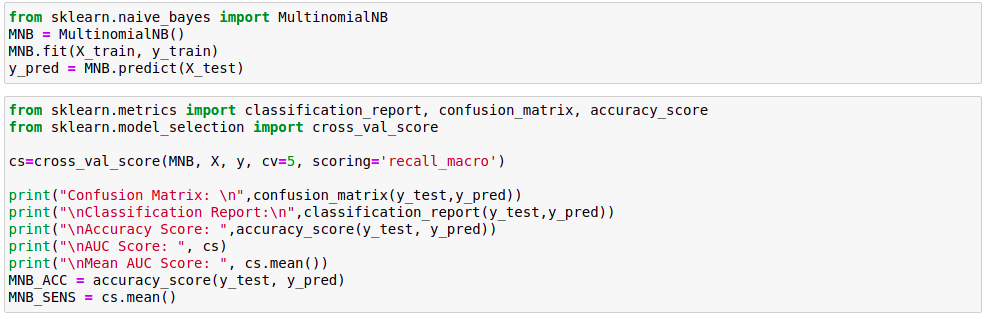


PCA with KNN



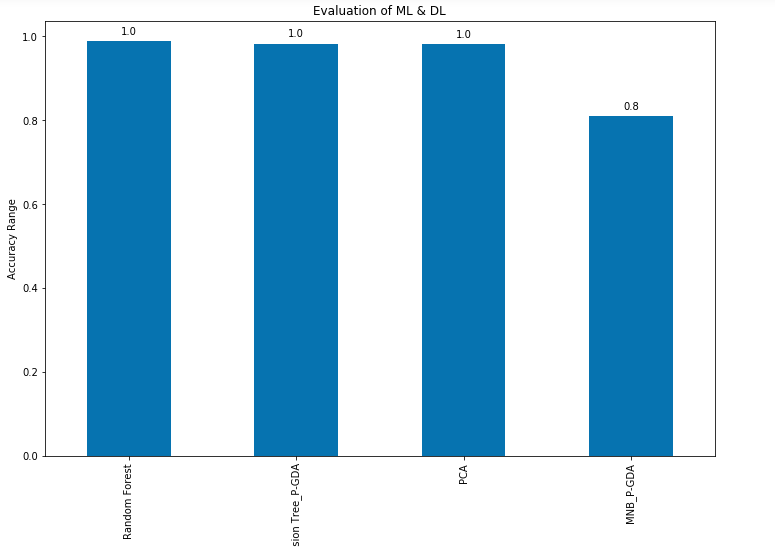
P-GDA : Proposed Genetic Disease Analysis PCA : Principal component analysis

MLP

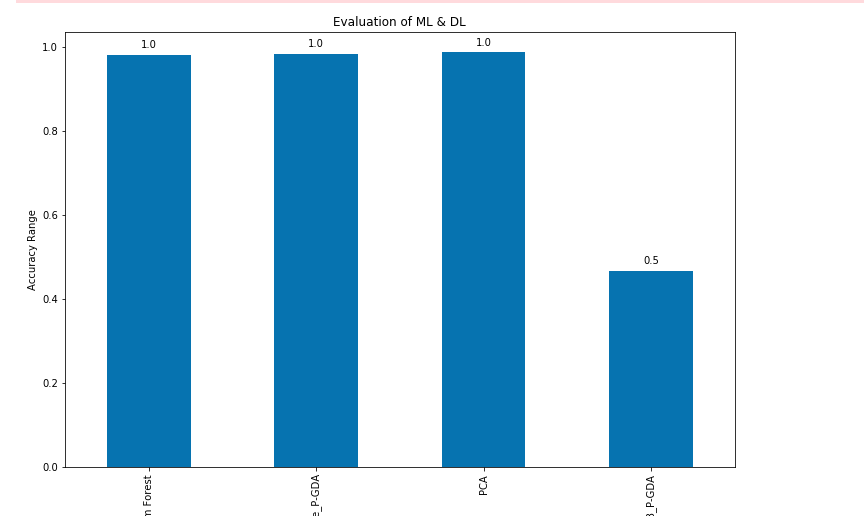


Model Comparsion

Accuracy Comparsion with P-GDA



Sensitivity



From the above comparison we can concluded that DT or RFC as proposed genetic disease analysis (P-GDA)