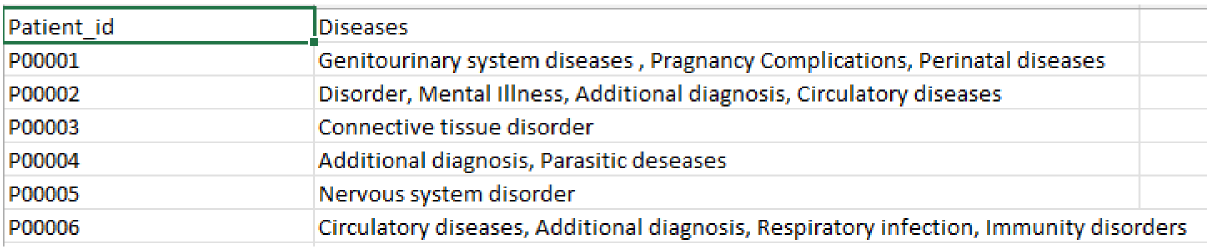
Report

I acknowledge and hereby confirm that the submitted work is entirely my own and I have not (i) used the services of any agency or person(s) in the preparation of the work I submit for this assignment; (ii) given assistance other candidates submitting assignment.

Your Name : Ashish Rana Roll No. : 18171

Problem 1) : - The output that I have gotten is like this:-



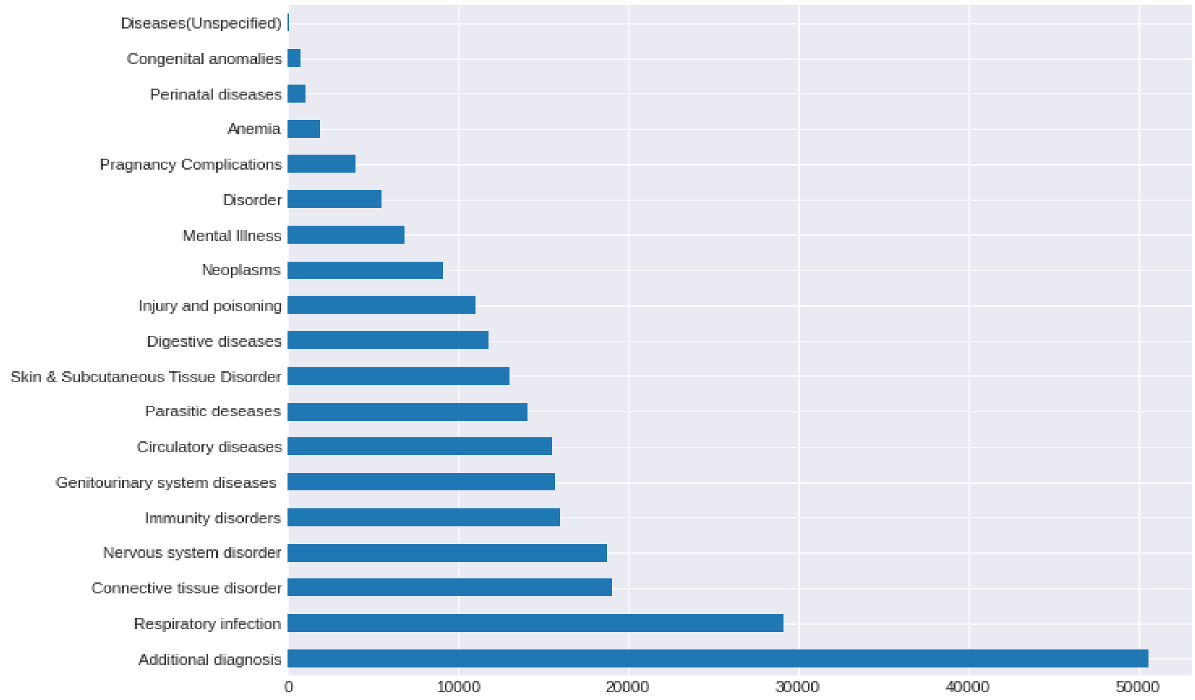
Patient_id	Diseases
P00001	Genitourinary system diseases , Pragnancy Complications, Perinatal diseases
P00002	Disorder, Mental Illness, Additional diagnosis, Circulatory diseases
P00003	Connective tissue disorder
P00004	Additional diagnosis, Parasitic deseases
P00005	Nervous system disorder
P00006	Circulatory diseases, Additional diagnosis, Respiratory infection, Immunity disorders

First as the ICD10 code in ccs is not similar to the ICD10 code in Diagnosis. Therefore I first make them similar by removing the "." From the ICD10 code of Diagnosis dataset. And then I replaced those ICD10 code with ccs_1_desc. But as the ccs_1_desc contains large statement which makes the output hodgepodge, so I replaced them with the concise description. Down below I mentioned special case that comes into picture during that operation, how I delt with them

For example Endocrine; nutritional; and metabolic diseases and immunity disorders -- > Immunity disorders

Special Case:-

1) Residual codes; unclassified; all E codes [259. and 260.] -- > Disorder
*** As there is not enough clinical evidence for diagnosis, therefore, I used Disorder rather than Diseases because disease is distinct and measurable.
2) Codes (not in ccs map) :- As I don't have any information about those codes in ccs dataset, but these code still represent a diseases, so I will replace these code with diseases(unspecified)



3) Ill defined condition :- . Additional diagnosis

Conclusion:-

Patients mostly suffering from resipiratory infection other than Additional diagnosis. This predominance of respiratory issues might allude to the derogatory effect of the increasing air pollution.

Problem 2) Prescription dataset has 5 columns i.e. Patient_id, Prescription_date, drug_category, drug_group and drug_class.

Objective :- Generate an output column.

1) Approach 1 :- Use prescription year to classify whether patient is curable or not curable.

For example :- Case 1) Patients whose prescription year does not include 2018. As only two case happened in such case either the person is curable or dead. But I consider those patient curable ("1").

Case 2) Patients whose prescription year including 2018 and other year as well. As those patients suffering long term health problem, therefore they are not curable ("0").

Case3) Patients whose first visit in 2018. For such kind of patients, their curability can be decided by their corresponding drug category. (check?)

Limitation :- For example One patient i.e. P00806 has Antiparkinson . The person only visits in the year 2016. Therefore above approach conclude that person to be cured. But in reality that person is not curable.

2) Approach 2 :- In which I go through the drugs (drug_category) which appears maximum number of times and check the curability (on Internet) and if the drug is simple like Antifungal, cough/cold , I give it "1". The drugs which is used for non-curable disease like Antiparkinson, I give "-1" And those drugs which I have not checked as the number of unique value in drug_ Category is 89, I give it "0".

The output column by approach 2 is generated :- Sum of the medicines (numbers mentioned approach 2)

- If positive number comes out, the patient is cured. (represent 1)
- If negative number comes out, the patient is cured. (represent 2)
- If zero comes out, No idea. But Suggestion :- 1) Additional Diagnosis 2) Check drug category. (represent 0)

The value_counts of output column is :-

Cured (1)	52043
Not Cured(2)	18855
Neutral(0)	11648

This indicates the dataset is biased. Therefore, in order to deal this imbalance, I will use SMOTE.

Evaluation Metrics :-

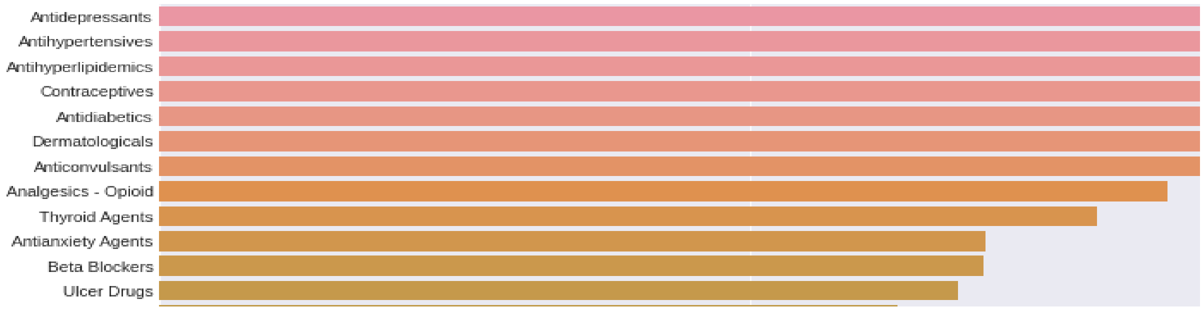
I have decided to take only confusion matrix as my evaluation paramtere. What specifically I'm looking in the confusion matrix is that non-cured patient (Actual) ---> cured patient (model prediction) should be less.

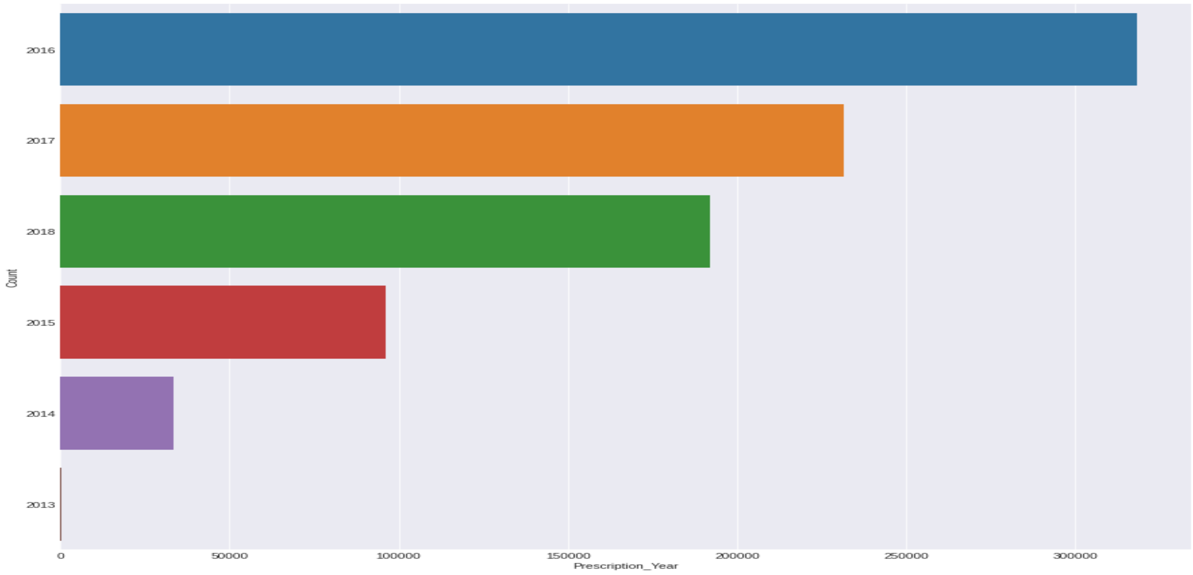
Say patient A is non-cured, and if our model predicts non-cured then great. What if our model predict it cured, then that's problematic. Lets say our model gives "0" output for that patient, then, still it is fine because in that scenario either we will send patient again for diagnosis or check its drug_category.

Say this time patient B is cured but our output predicts it is non-cured, then, patient B will go to doctor and doctor will declare patient B cured. And if for B our model predicts 0, then stillit is fine.

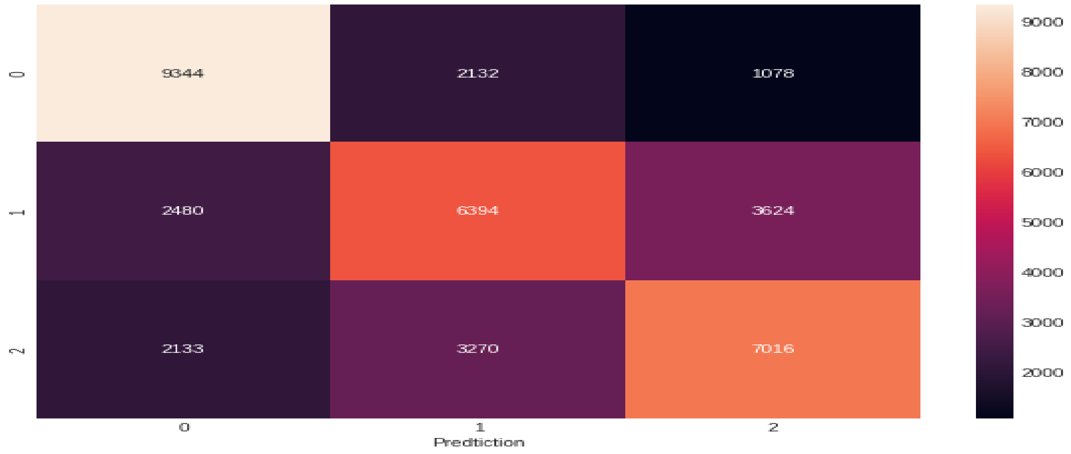
The above two case clearly indicates why I’m specifically watching the variation of specific elements of confusion matrix rather than other parameter. And why I’m calling “0” as a neutral as it works well in every case.

Below Drug Category vs Count(appearance in dataset) , I have pasted a small part of that portion which gives Drug name that appear maximum number of times and patients visits to asylm(below to bleow). Below graph indicates that people are suffering from depression. It is matter of concern. Because depression leads to sucide that’s why India’s sucide rate is between 10-15 (“wikipedia source”).





Model :-   This is the best I come up with (use XGBoost) ..


Prediction