

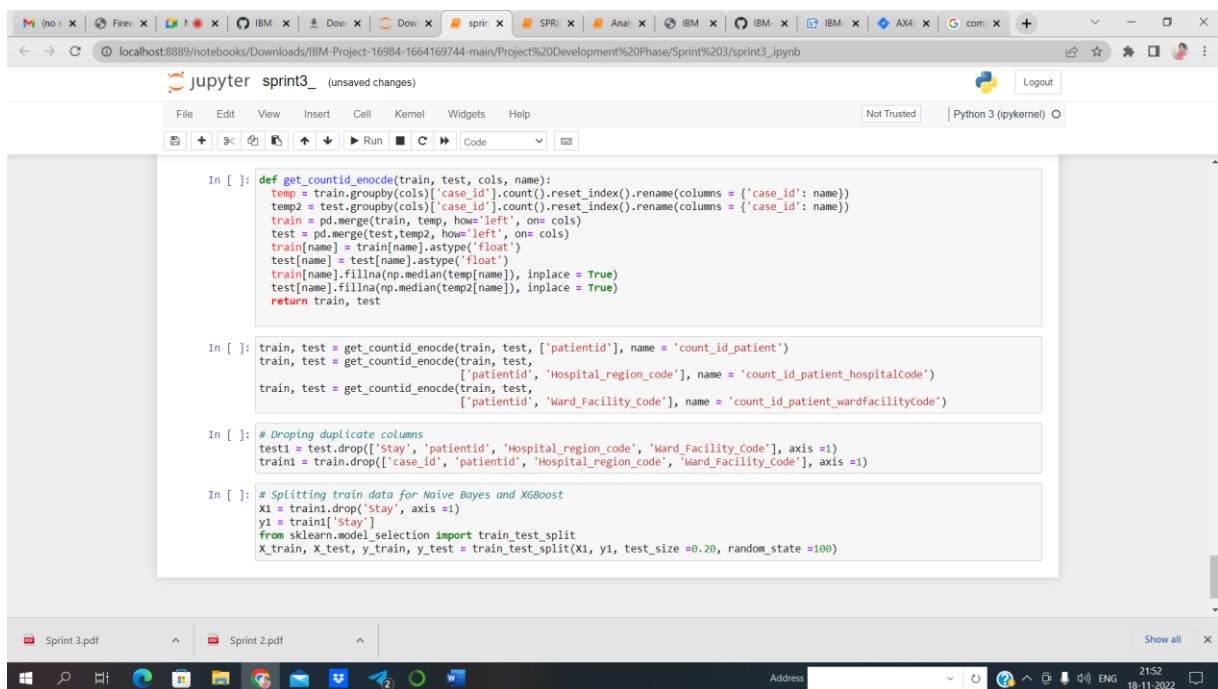
## SPRINT 3

Date	13-11-2022
Team ID	PNT2022TMID04305
Project Title	Analytics for Hospitals' Health-Care Data
Team Members	Abishankari S , Aravindh T, Dhivya A , Dhivya S

### Feature Engineering

Once the data is cleaned and prepared, we grouped patientid and case\_id to extract the new column “count\_id\_patient”. This variable contains the count of multiple admits of a patient under different case\_id. Further two more columns “Hospital\_region\_code” and “ward\_facility\_code” were grouped to patientid and case\_id. These two new variables “count\_id\_patient\_hospitalCode” and “count\_id\_patient\_wardfacilityCode” contain the count of multiple admissions in a hospital region and the count of multiple wards allocated to a patient.

Before getting into analysis, the train data must be split into two parts, the first part with all the feature variables and the second part with a target variable (“Stay”). Then preprocessed into train and validation sets. So, here we are portioning the train set with 80% and validation set with 20% of the data for Naïve Bayes and XGBoost models.



```
In [ ]: def get_countid_enocde(train, test, cols, name):
temp = train.groupby(cols)['case_id'].count().reset_index().rename(columns = {'case_id': name})
temp2 = test.groupby(cols)['case_id'].count().reset_index().rename(columns = {'case_id': name})
train = pd.merge(train, temp, how='left', on= cols)
test = pd.merge(test, temp2, how='left', on= cols)
train[name] = train[name].astype('float')
test[name] = test[name].astype('float')
train[name].fillna(np.median(temp[name]), inplace = True)
test[name].fillna(np.median(temp2[name]), inplace = True)
return train, test

In [ ]: train, test = get_countid_enocde(train, test, ['patientid'], name = 'count_id_patient')
train, test = get_countid_enocde(train, test,
['patientid', 'Hospital_region_code'], name = 'count_id_patient_hospitalCode')
train, test = get_countid_enocde(train, test,
['patientid', 'Ward_Facility_code'], name = 'count_id_patient_wardfacilityCode')

In [ ]: # Dropping duplicate columns
test1 = test.drop(['stay', 'patientid', 'Hospital_region_code', 'Ward_Facility_code'], axis =1)
train1 = train.drop(['case_id', 'patientid', 'Hospital_region_code', 'Ward_Facility_code'], axis =1)

In [ ]: # Splitting train data for Naive Bayes and XGBoost
X1 = train1.drop('stay', axis =1)
y1 = train1['stay']
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X1, y1, test_size =0.20, random_state =100)
```

## Choosing ML models for analysis:

The goal is to predict Length of Stay i.e., “Stay” column (Target Variable) and it is classified into 11 levels. We must find the probability of each patient’s length of stay using feature variables, which contain the patient’s condition and hospital-level information.

After analysis of different ML models for classification, we have decided to choose three models:

- 1) Naives Bayes Model
- 2) XGboost
- 3) Neural Networks

We have chosen Naïve Bayes model, as the feature variables are ordinal in nature and also Naïve Bayes Model is a perfect multilevel classifier.

So, the model 2 we have decided to use XGBoosting. Boosting is a sequential technique that works on the principle of an ensemble model. It combines the set of weak learners and improves prediction accuracy. The final prediction score of the model is calculated by summing up each and individual score.

The 3<sup>rd</sup> model we have chosen to be neural network because in NN, increasing the number of neurons from each layer to the other layer, will increase the hypothetical space of the model and try to learn more patterns from the data. So, this can better predict the Length of stay of patients.

Sprint 3 completed successfully