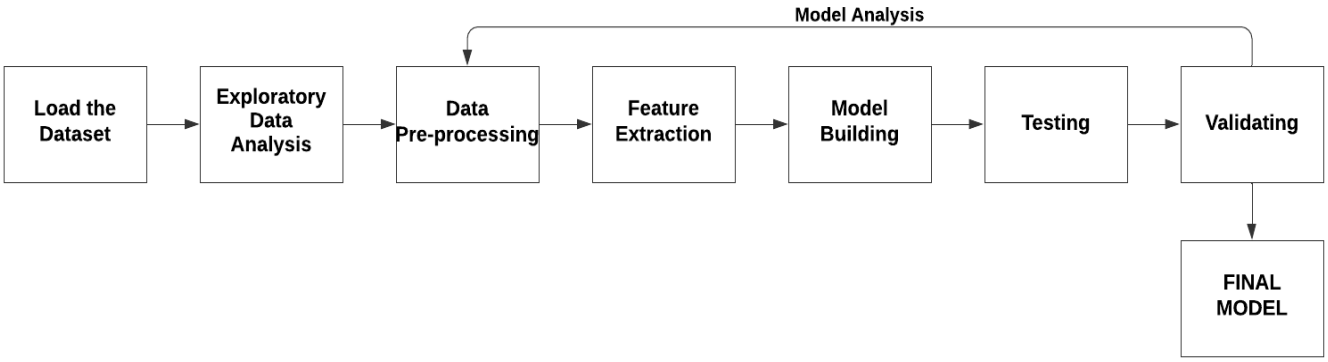
Commit Classifier in Version Control System

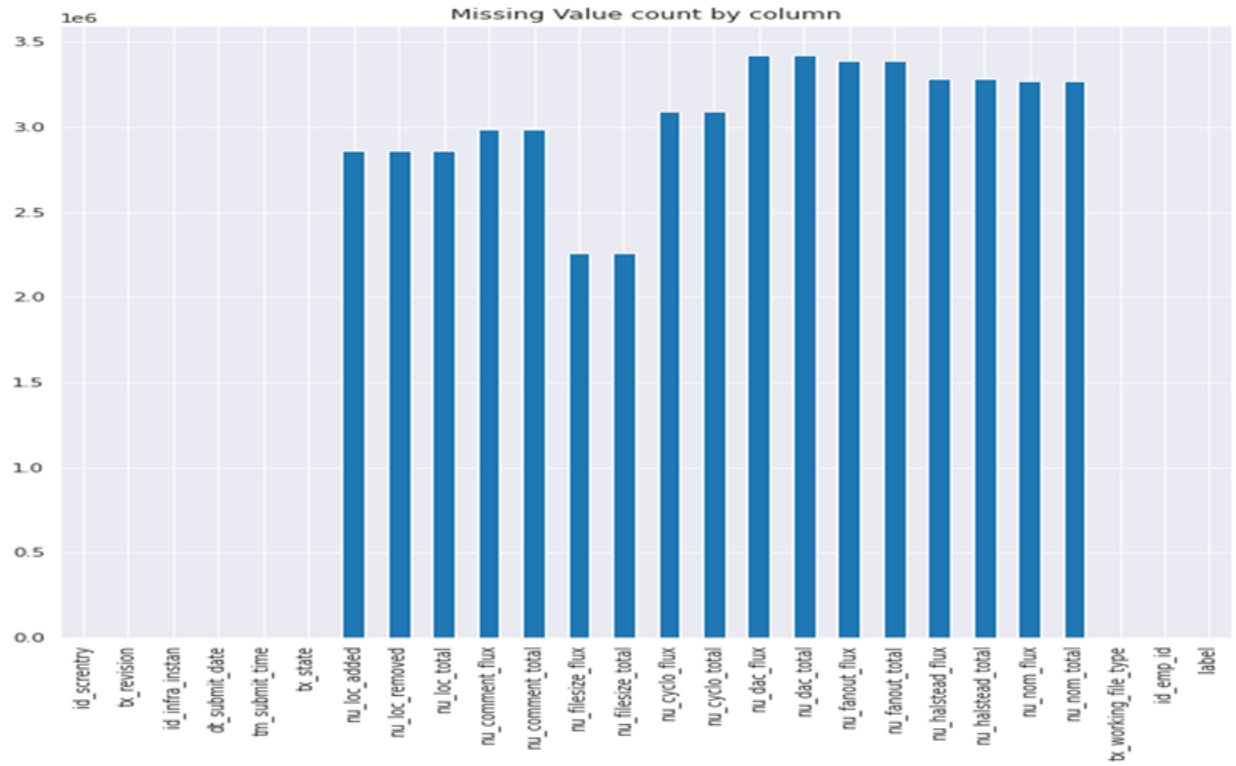
- **Data Pipeline:**



After importing the libraries and **Loading the Dataset** , we analyse the patterns in the data by **Exploratory Data Analysis.** To make data relevant we undergo **Data preprocessing** to make data free from missing values and outliers. This preprocessed data is followed by **Feature Extraction** where constant, quasi-constant and correlated features are removed. This processed data is passed to the **Model Building** phase where classifiers are trained and then **Testing** and **Validating** the model in order to handle overfitting and underfitting of the data and get more accuracy. The validation score is used to analyses the data preprocessing techniques.

- **Exploratory Data Analysis**

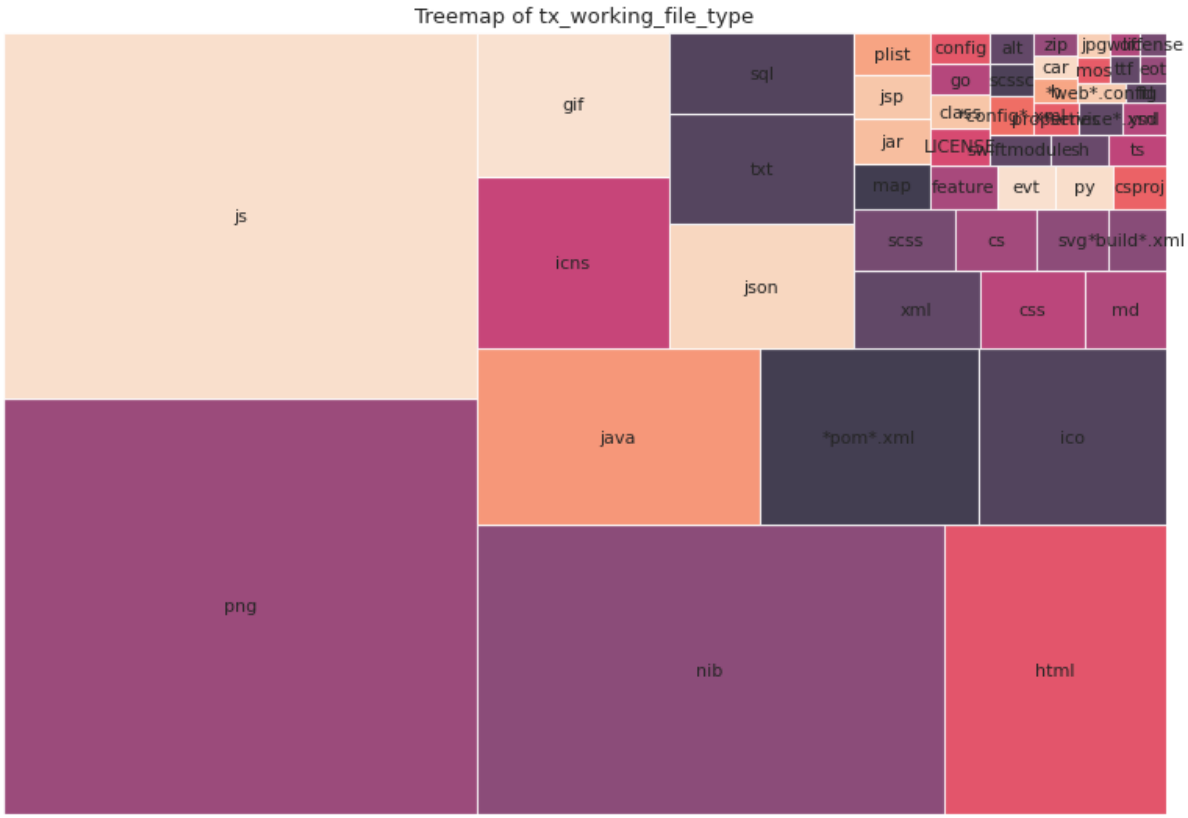
  Dealing with Missing value:



The total count of missing values in the dataset is 52,001,555 which constitutes 57.51% of the dataset. A bar of the same is shown below. The plot here represents the count of missing values per column. The missing values are handled by filling those values with the median of the column

- **Categorical to Non Categorical:**

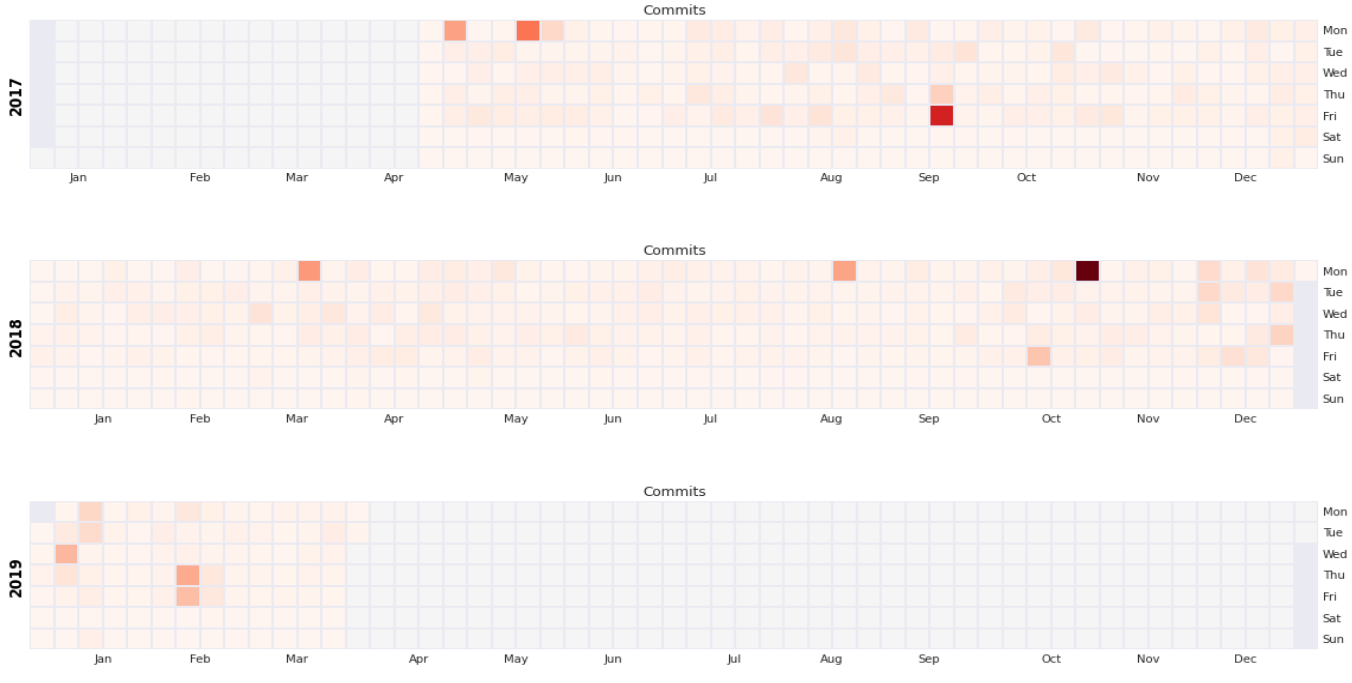
Working File Type:



There are 1603 types of working files in tx_working_file_type column out of which 5 were crossing threshold of 1.5lacs and rest types are coined as “OTHERS”

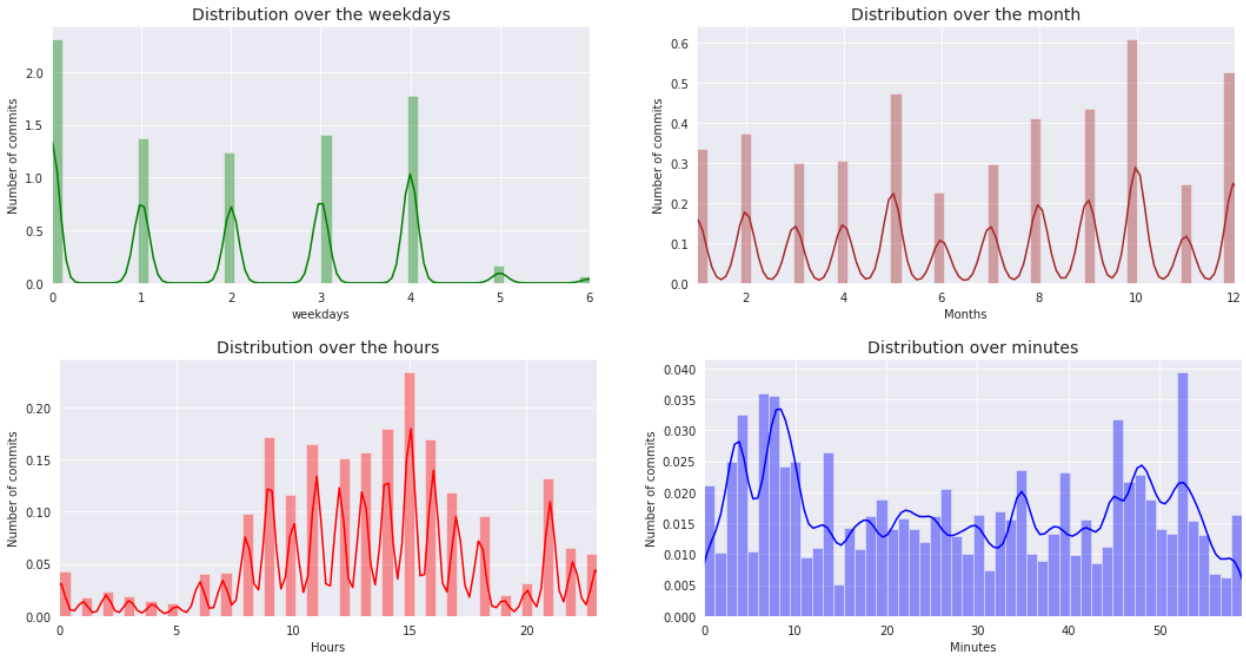
- **Dealing Dates and Time**

The calendar heatmap demonstrates the frequency of commits distributed throughout the year. The dataset contains commits from 2017-2019. The map was plotted using the information from dt_submit_date column.



- **Date time Distribution:**

The following distribution plots illustrates how the number of commits vary over the years, months and weekdays. They helped us analyse what time of the day, what month of the year and which day of the week bags the most number of commits.



- **Label Distribution:**

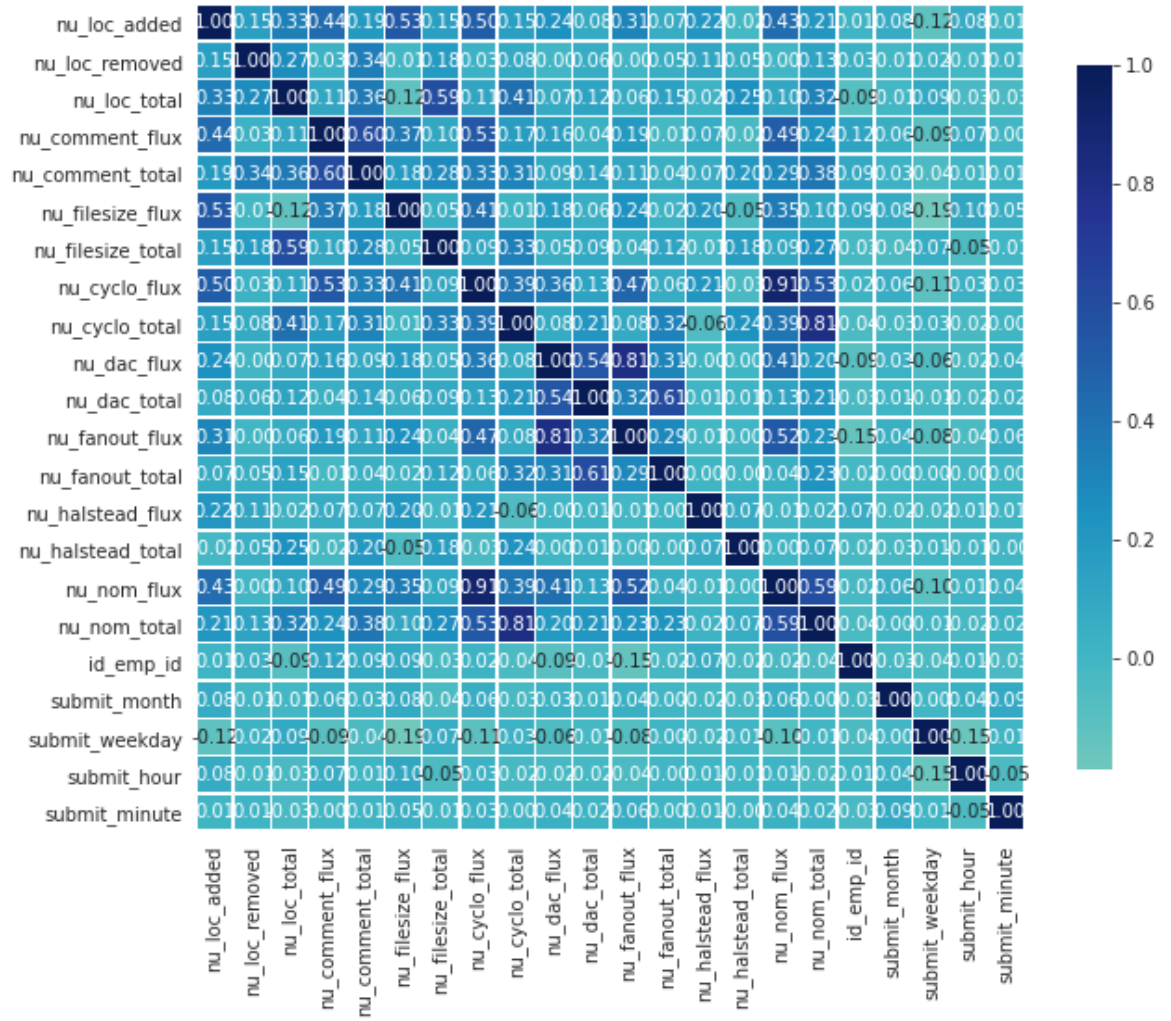
Dataset is composed of 60.7% of False i.e. Developer Activity and 39.3% of True i.e. Build Activity. As the ratio is almost 60:40 so it didn't fall under category of class imbalance and thus no need of sampling

- **Outliers:**

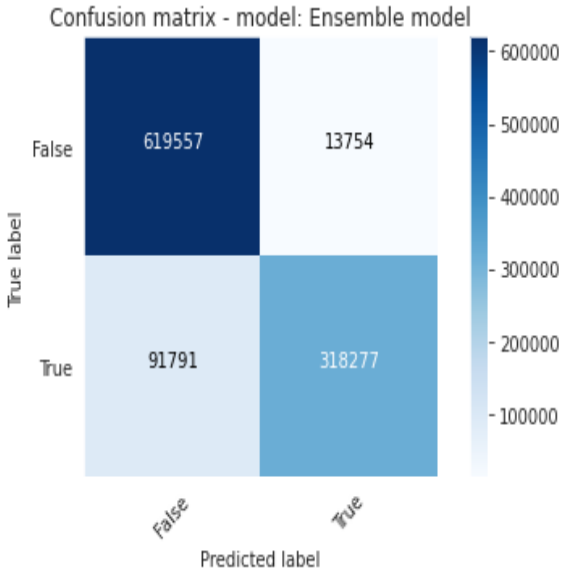
Presence of outliers in the dataset is realized by plotting these box plots. Box plots use the concept of Inter Quantile Range (IQR) which is the difference between 1st (0.25) and 3rd (0.75) quantiles of the dataset. Points are marked as extreme outliers when they fall out of the min and max values. A threshold value of 1.5 is used to calculate the min and max values. These outliers are handled by applying log transformation on the skewed variables which help correct its distribution.

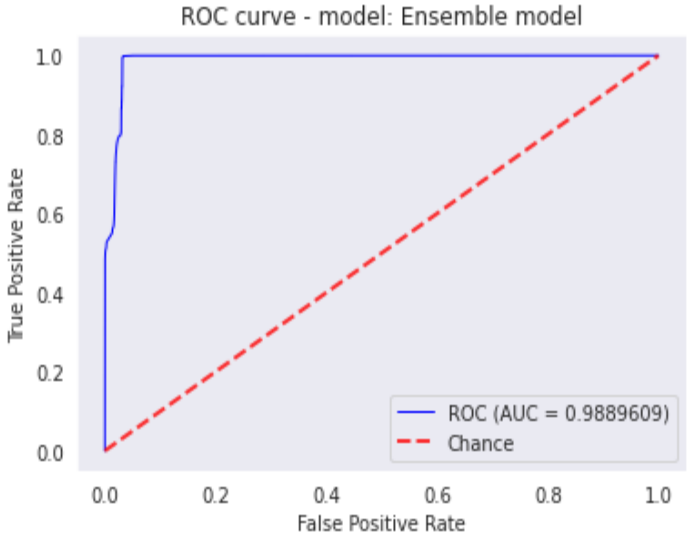
- **Features extraction :**

After constant and quasi constant filters correlation matrix is plotted to analyse the correlation among the columns. The columns having correlation value more than 0.8 are removed .The columns which are negatively correlated will result in True (Build activity) .The columns which are positively correlated will result in False (Developer activity).

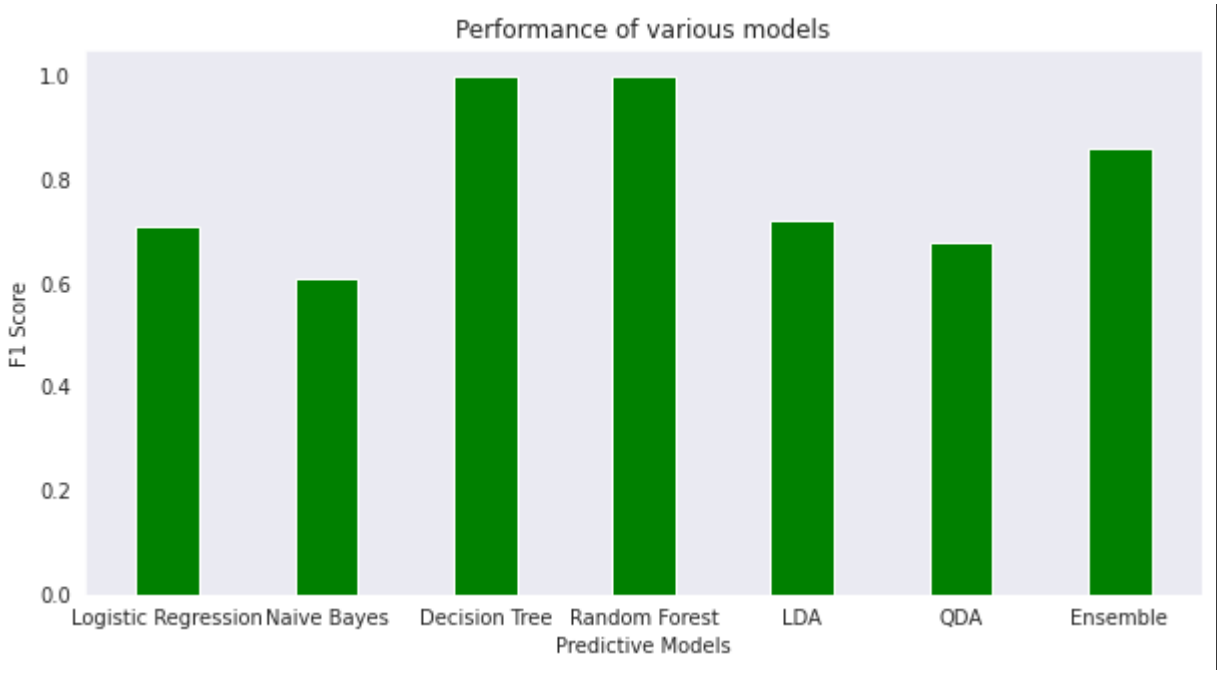


- **Model**





Ensemble Model is trained, tested and validated over the dataset and overall accuracy is 90%
The models deployed under ensemble learning are below with their accuracies



Summary

- Yes, the model is efficient to handle such large dataset as it is already tested on huge dataset and getting an accuracy 90%. Moreover we can also implement Dimensionality Reduction Techniques such as PCA etc. to deal with more larger datasets

- The model can be scaled up with minimum effort by tuning hyper parameter using Grid Search CV provided a good specs system is used

- Productionsation stand for the product in which we are representing our work to be implemented in real world by Data Scientists, Business Analysts, and Developers. It can be represented as open source ,application and library , module , packages and API's
- So the project can be deployed as libraries, applications, dashboards, or API endpoints so that other members of the data science team can leverage to further extend, disseminate, or collaborate on their results.
- We would extend our solution by productionising it in form of an API or deploying on platforms like Flask or Heroku and if more time is provided our focus would be optimizing the code and using Dimensionality Reduction techniques and grid search in order to tune hyper parameters so as to provide a more accurate and robust product to the outer world.