ML Assignment

Dataset Preparation and Preprocessing

We exported 3 datasets of our group members of at least 1000 mails each Using Gmail's labels and Microsoft Outlook's File handling.The columns which we felt were irrelevant like BCC,CC were removed from the csv files. Our dataset had 5 unique labels i.e -'Updates', 'Social', 'Promotion', 'Forum', 'Spam'. After converting the data into a Pandas Dataframe. labels were encoded using LabelEncoder of sklearn.preprocessing following which Subject and body of email were concatenated, following which variable X and Y were defined where X was the "subject+body" while Y was "labels". Then, we randomly split our datasets into training and testing dataset of 80:20 ratio by using sklearn.model_selection module, following which a deeper cleaning of dataset was done using re module and hence html tags, slangs, extra spaces, links etc were removed and dataset was updated, therefore reducing significant noise from the datasets .

Feature extraction

After significant cleaning, we vectorised our data using TfidfVectorizer (using sklearn.feature_extraction.text module). We extracted 1-gram,2-gram,3-gram and 4-gram features using the vectoriser. Following which the sparse matrix we had after vectorisation were converted to array type structure for convenient use ahead.

Model training and testing

We selected 5 different models which would probably be best suited for the Multi label classification problem, namely- Support Vector Machine(SVM), Multinomial Naive Bayes (MNB), K nearest neighbours (KNN), Decision trees and Random Forest models.
We trained all the models using the training dataset, immediately after which the test dataset was used on the trained models to find the accuracy. This process was repeated thrice owing to the 3 different test datasets we had. We also plotted various graphs using matplotlib and seaborn and made comparisons.