**Machine Learning Assignment**

The aim of the assignment is to train ML models in order to classify emails into our defined labels and compare their accuracy. The attributes used for classification are the mail’s subject, sender and body. Following are the steps undertaken:

1. Create a labelled dataset (CSV format) with emails from 6 categories namely:
   1. Course
   2. Clubs
   3. Important
   4. Outside (Not within the BITS Organisation)
   5. Personal
   6. Promotion
2. Cleaning the dataset:
   1. Remove spaces and punctuation
   2. Remove stopwords
   3. Lemmatise the text
3. N-gram Feature Extraction
   1. Use unigram, bigram and trigram
   2. Set a threshold of 0.1 i.e. 10% of all mails
4. Pick out models for classification. The following models were used:
   1. Linear SVM
   2. Naive Bayes
      1. Multinomial
      2. Bernoulli
   3. Decision Tree
   4. Logistic Regression
5. Train the models with 80% of the labelled data and use the rest 20% for testing. Train and test 4 times with:
   1. Amulya’s mails
   2. Josh’s mails
   3. Tanmay’s mails
   4. All of our mails
6. Compile results from all the models in form of graphs
7. Make Observations
   1. The accuracy given by the models in decreasing order are:
      1. Logistic Regression: 91.9%
      2. Decision Tree: 90.5%
      3. SVM: 90.1%
      4. Naive Bayes (Multinomial) :77%
      5. Naive Bayes (Bernoulli): 70%