VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

**School of Computer Science and Engineering**

**Practice Problem set**

Course Title: Machine Learning Lab Faculty: Prof.G. Manikandan, Prof.G.N. Balaji

Code: MCSE602P

1. Create a dataset using an API with Python (Use Web Scrapping/web crawling to create your own dataset) from anyone (discussed in class) of the following application domains.

    a. IMDB

    b. Flipkart

    c. Amazon

    d. Twitter

2. Apply pre-processing techniques such as

- Stopwords Removal
- URL Removal
- Stemming
- Lemmatization
- Convert Numbers to Words
- Tokenization
- Unigram/Bigram Approach

etc.,

Intermediate Result: Show Pre-processed data in each

3. Apply feature selection algorithms to extract the predominant features.

*[Note: FS algorithms which is discussed in class]*

4. Use Classification algorithms for classification such as

    a. Naive Bayes

    b. Multinomial Naive Bayes

    c. SVM

    d. Random Forest

5. Interpret the result

    a. Print confusion matrix

b. Use 10-fold cross validation

c. Give the summary of results such as accuracy, precision, recall, f-measure and Matthew Correlation Coefficient (MCC)

d. Compare the results with 4-classifier and suggest the best classifier. (Represent the comparisons with tabular format)