**BIG DATA PROJECT (UE19CS322)**

**Topic: Machine Learning with Spark MLlib**

**Dataset: Spam**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Name** | **SRN** | **Class/Section** |
| 1 | Sammed Tapire | PES1UG19CS426 | 5G |
| 2 | Samarth Kumar | PES1UG19CS424 | 5G |
| 3 | Peruru Shyam Prasad | PES1UG19CS327 | 5E |
| 4 | Naveen Kumar | PES1UG19CS292 | 5E |

**Introduction:**

Ours final project is simulating a real world scenario where we have handle an enormous amount of data for predictive modelling. Basically, we streamed the dataset as dstream and converted into dataframe in batches.at each batch arrival we trained the different models in an incremental fashion

**Design details**:

1)In this implementation, we are streaming dataset through socket connection and storing the dstream.

2)Then we created dataframe from each rdd in the dstream by applying transformations inside a helper function.

3) After obtaining the dataframe from each batch dstream, we are preprocessing.

4) we are training multiple models namely SVM, Random Forest Classifier, Naïve bayes and linear SVC.

5) Finally we tested the model using the test dataset to obtain accuracy of each model.

**Surface level implementation details**:

1)Preprocessing –

1. Using “nltk” module to remove stopwords in message and subject column values.
2. Regular expression to remove unwanted words
3. Using pyspark feature like RegexTokenizer,CountVectorizer and StopwordRemover.

2)models -SVM, Random Forest classifier, Naïve Bayes

3)storing the trained model in a pickle file and loading at each batch.

4)testing the data with the model stored in the pickle file after all iterations.

**Reason behind design decisions:**

1)doing the Preprocessing first to obtain data which is in a suitable manner for training the classifiers.

2)after training the model with preprocessed data, storing in a “.pkl” file to use in next iteration to train model incrementally.

3)after final model is created, testing to get accuracy of each classifier.

**Takeaway from the project:**

A accurate model to predict the new data object as “spam” or “ham” which can be used in many applications.