BIG DATA PROJECT REPORT

SPARK STREAMING FOR MACHINE LEARNING

DATASET CHOSEN : Email Spam

TEAM DETAILS

NAME	SRN
MEELA DEEPTI	PES2UG19CS227
PRIYA MOHATA	PES2UG19CS301
LOHITH SRINIVAS T	PES2UG19CS203

DESIGN DETAILS :

1> We got our dataset for email spam analysis.
2> We then streamed the dataset.
3> After streaming, we applied pre-processing techniques on our dataset.
4> After pre-processing, we obtained the clean data frame with required features.
5> We then applied the first model which is **MULTINOMIAL NAÏVE BAYES** using sklearn.
6> We then applied our second model which is **SGD LINEAR REGRESSOR** using sklearn.
7> Our third model that we applied is **MINI BATCH K MEANS CLUSTERING** using sklearn.

IMPLEMENTATION DETAILS :

_For each RDD, we first check if it is not empty and then do the streaming. We then convert it into a dictionary and create data frame. The data we get from here is clean data. In the user defined function defined by us we have defined how to read the data stream and print the respective data read.
After Streaming of the dataset, we performed pre-processing on our dataset. In pre-processing we used :

1> TOKENIZER- This converts the input string to lowercase and then splits it by white spaces.
2> REMOVING STOP WORDS- A feature transformer that filters out stop words from the input.
3> COUNT VECTORIZER- Convert a collection of text documents to a matrix of token counts.
4> IDF- Convert a collection of raw documents to a matrix of TF-IDF features.
5> String Indexer- String Indexer encodes a string column of labels to a column of label indices

In multinomial naïve Bayes model, we have used the conditional independence between every pair of features given the value of the class variable.

In SGD Linear regressor ,the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule.

Mini batch K means algorithm uses mini-batches to reduce the computation time, while still attempting to optimise the same objective function.

We have used accuracy, confusion matrix, precision and recall to evaluate our model.

REASON BEHIND THE DESIGN :

We have used sklearn for the implementation of our models since it is easy to use.
The scikit-learn website provides elaborate API documentation for users who want to integrate the algorithms with their platforms. This feature has helped us to build and get a deep understanding of our models. The scikit-learn library is very versatile and free to use, with minimum legal and licensing restrictions.

TAKEAWAYS :

This project has exposed us to various fundamentals and intrinsic details about pyspark and sklearn. This has expanded our knowledge about machine learning algorithms .This project has given us a hands on experience of handling pyspark and sklearn.