BIG DATA
SPARK STREAMING WITH MACHINE LEARNING

NIDHI BHARITYA - PES2UG19CS254
SHATAKSHI BHARDWAJ - PES2UG19CS378
MANOJ KUMAR K - PES2UG19CS222
DATA SET - Email Spam Analysis

We streamed the dataset from the spark streaming. The dataset is then processed which provides the cleaned dataset.

Models used are
Multinomial Naïve Bayes
Bernoulli Naïve Bayes

Implementation:
We initially stream the data in batches of a particular batch_size and check whether the RDD is empty or not. Later we clean the data by ENGLISH_STOPPING_WORDS so that we can get better accuracy and performance of the models will be increased. We chose these models because they had better accuracy and it was basically a hit and trial method.

Thank You