PES UNIVERSITY

B. TECH. (CSE)
V SEMESTER

UE19CS323-BIG DATA

Team ID -BD2_006_038_050

FINAL PROJECT REPORT
ON
SPARK STREAMING FOR MACHINE LEARNING

SUBMITTED BY

NAME	SRN
ABHIGYAN MANASVI	PES2UG19CS006
ANANYA BHATNAGAR	PES2UG19CS038
ANSHUMAN MANDAL	PES2UG19CS050

AUGUST - DECEMBER 2021
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ELECTRONIC CITY CAMPUS
BENGALURU - 560100 KARNATAKA INDIA

DATASET CHOSEN- Email Spam

STEPS FOR EXECUTION-

For training-

We streamed our given dataset through this command:

python3 stream.py

In a new terminal we run

spark-submit bdproj.py

Using this **bdproj.py** we create models which will be used to detect spam in our test data.

For testing-

For our test data we do the same ,we first stream using our stream command

python3 stream.py

In new terminal we run the command,

spark-submit testdata.py

Then we get the results

MODELS USED-

We have used three models:

Logistic Regression-

Logistic Regression is a Machine Learning algorithm which is used for the **classification problems**, it is a predictive analysis algorithm and based on the concept of probability. ... The hypothesis of logistic regression tends it to limit the cost function between 0 and 1 .

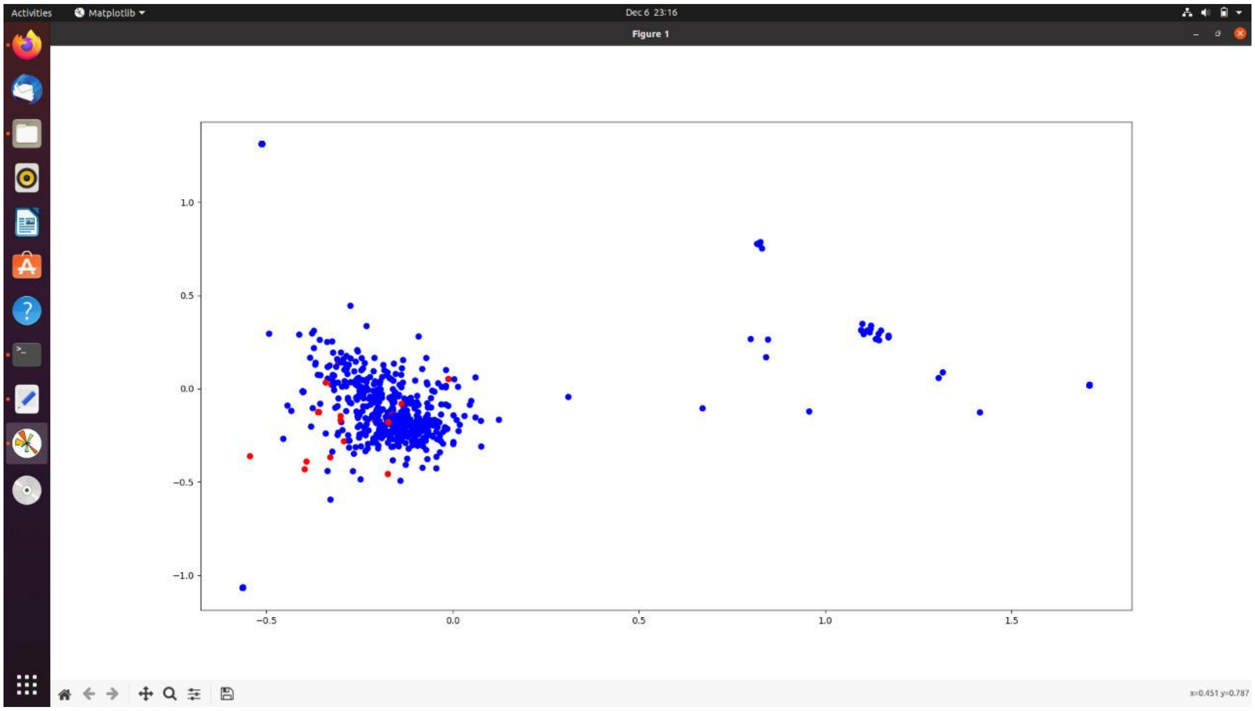
MLP-

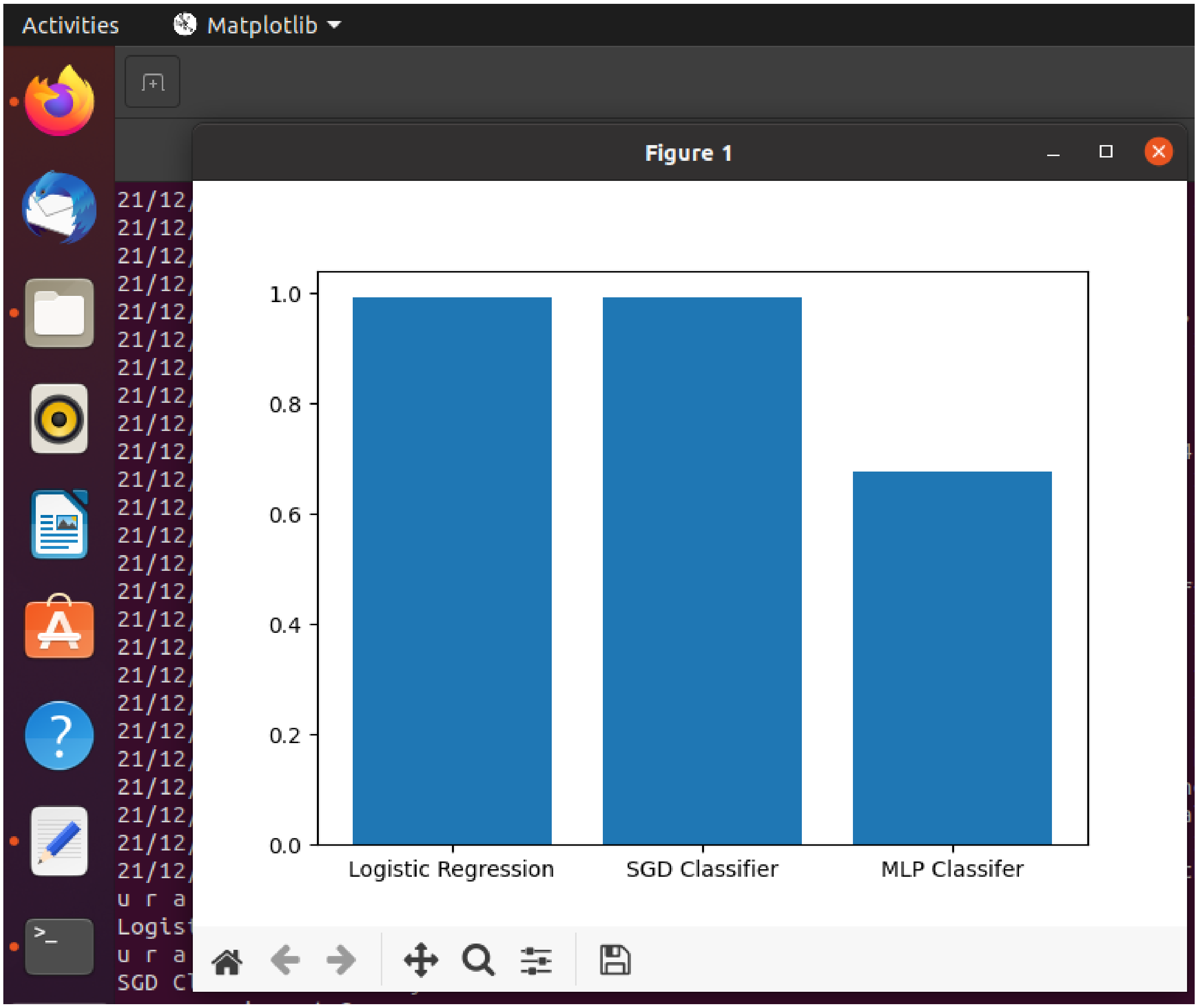
MLPs are suitable for **classification prediction problems** where inputs are assigned a class or label. They are also suitable for regression prediction problems where a real-valued quantity is predicted given a set of inputs.

Stochastic Gradient Descent (SGD):

It is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression.

OUTPUT SCREENSHOTS-





```
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
21/12,
u r a
Logis
u r a
SGD C
```

TAKEAWAYS:

This project made us familiar with the fundamentals of pyspark and sklearn. This has enriched our knowledge regarding the machine learning algorithms. We got hands-on experience on pyspark and sklearn.