CS 5433: Bigdata Management
Programming Assignment 2
ReadMe for Task3

GROUP 4

Task2 – Measure accuracy for the models (a) Linear Regression (b) Random Forest

Steps to run code:

For Part A (Linear Regression) and Part B (Random Forest)

1. After logging into Hadoop cluster, create a python file.

   For Part A

   File Name: Assign3_Group4_Task3_PartA.py

   Refer to "Group_4_Task_3_Part_A_code.pdf"

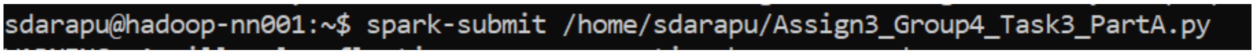
   For Part B

   File Name: Assign3_Group4_Task3_PartB.py

   Refer to "Group_4_Task_3_Part_B_code.pdf"

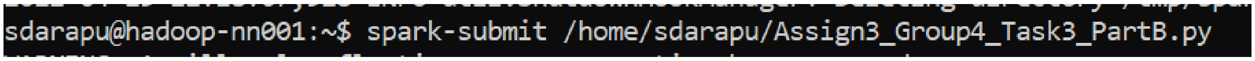
2. Execute the below command to run the file. After this, it takes some time to get executed.

   For Part A



```
sdarapu@hadoop-nn001:~$ spark-submit /home/sdarapu/Assign3_Group4_Task3_PartA.py
```

   For Part B



```
sdarapu@hadoop-nn001:~$ spark-submit /home/sdarapu/Assign3_Group4_Task3_PartB.py
```

3. Once, the file gets executed. The output gets displayed on the console.

<u>Comparison of Results from PartA and PartB</u>

While we compared the results of RMSE value for both Linear Regression and Random Forest, we got RMSE value higher in Linear Regression when compared with Random Forest. But R2 value is higher in Random Forest than Linear Regression which is quite opposite compared to above scenario.

Note: We have given permissions to username “aanmol” for all our input and output file under username “sdarapu”. So that, you can access all my files. If there is any issue with permissions, then please let us know.