CS 5433: Bigdata Management
Programming Assignment 2
ReadMe for Task2

GROUP 4

Task2 – Implement prediction algorithms (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_Task2_PartA.py

   Refer to "Group_4_Task_2_Part_A_code.pdf"

   For Part B

   File Name: Assign3_Group4_Task2_PartB.py

   Refer to "Group_4_Task_2_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_Task2_PartA.py
   ```

   For Part B

   ```
   sdarapu@hadoop-nn001:~$ spark-submit /home/sdarapu/Assign3_Group4_Task2_PartB.py
   ```

3. Once, the file gets executed. First, the output gets displayed on the console and also the output gets saved in the path specified in the code.

   To view the data stored in the specified folder under hdfs, execute the below command.

   For Part A

   ```
   sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/Assign3_Group4_Task2_PartA_Output/part*
   ```

For Part B

```
sdarapu@hadoop-nn001:~$ hdfs dfs -ls /user/sdarapu/Assign3_Group4_Task2_PartB_Output
```

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.