**Steps for execution of the Project**:

**Project Demo**:

<https://www.youtube.com/watch?v=WOXEJLsxbzg>

🡺Please see the project demo before going further.

**DATABRICKS PUBLIC LINKS:**

**Task1 Preprocessing:**

[https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5008245478875404/900645327760563/440415051426816/latest.html](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5008245478875404/900645327760563/440415051426816/latest.html%20)

**Task1 Analysis with LR, RF and GBT using SparkMLlib:**

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/855389465014467/3665561683253945/4847935780219074/latest.html>

**Task2 Preprocessing:**

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5008245478875404/727104593191945/440415051426816/latest.html>

**Task2 Analysis with LR, RF and GBT using SparkMLlib:**

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5008245478875404/4245257127570951/440415051426816/latest.html>

1) We have given 4 databricks public links here above.

2) Import those pyspark notebooks in to your databricks cluster.

3) Upload the csv training and test files given in the Dataset folder to your databricks cluster in tables.

Once upload is finished you will path to the files, copy those all paths and paste it somewhere in a notepad.

**Task-1 Execution:**

4) Now, open the "**KDDTask1\_Preprocess**" file and replace the paths in the sc.wholeTextFiles("") appropriately.

**I think only changing the underlined bold portion of file path suffice since file name remains same.**

For example, if a file path is like this:

/FileStore/tables/**urq6xroz1501511577436**/trajectories\_table\_5\_\_training-2c00c.csv

You need to change only the undelined bold one above since the file name remains same.

If this doesn’t work out please change the complete file path.

5) Once you change all paths click run all and results will be displayed after each cell execution. At the end of this preprocessing, the preprocessed csv file be stored into dbfs file storage.

6) Now, open the "**KDDTask1LR\_RF\_GBT**" file, here there is no need to change any paths, just click run all and results, predictions and metrics will be displayed after each cell execution.

**Task-2 Execution**:

7) Now, open the "**KDDTask2\_Preprocess**" file and replace the paths in the sc.wholeTextFiles("") appropriately.

8) Once you change all paths click run all and results will be displayed after each cell execution. At the end of this preprocessing, the preprocessed csv file be stored into dbfs file storage.

9) Now, open the "**KDDTask2LR\_RF\_GBT**" file, here there is no need to change any paths, just click run all and results, predictions and metrics will be displayed after each cell execution.