**Programming Homework2**

**Sai Rukesh Kantheti(SK429)**

**TASK 1 : Parallel training on 4 ec2 Instances**

This project is implemented with Spark data frames api and MLib libraries, with this Native spark implementation application is automatically parallelized and distributed natively.

Wine prediction application is developed using Spark DataFrames and MLlib. Running it on AWS EMR cluster automatically parallelize and distribute job execution. Hadoop Distributed Files system is used for locating dataset files and for storing trained models.

1.Create EMR Cluster

2.Upload files to EMR Cluster Master node

3.Copy files to HDFS

4.Launch ModelTrainer application

**TASK 2 : Predict wine quality on single ec2 instance**

At this stage we are interested in executing prediction code on a single ec2 instance. For that we need TestDataset.csv, wine-quality-predict.jar, model.tar.gz (from task1)

1.Ec2 instance Create:

2.Ec2 instance pre configuration:

2a.Install SCALA

2b.Install SPARK:

2b1.Upload trained model and jar files

2b2.Extract model.tar.gz

2b3.Disable unnecessary log4j

3.Run wine-predict application

**TASK 3 : Predict wine quality using docker**

For Predicting a wine quality on TestDataset.csv using docker. We need to have a full local file path of TestDataset.csv and provide it as an input argument while running a docker. So that TestDataset.csv can be copied to the local docker environment before running.

Test filename has to be **TestDataset.csv**and file has to be placed under data/ directory of container.

Docker Link: https://hub.docker.com/repository/docker/rukesh3663/ruke

GitHub Link: https://github.com/Ruke3663/Cloud-HW2/tree/master