Project Team Members:

1) Dhavalkumar Patel (001665804)
2) Mohit Gupta (001688141)

Please follow the below steps to build and execute both the standalone and AWS Elastic MapReduce (EMR) versions of our program:

Step 1 - On unzipping our solution you will get following deliverables in one folder:

1_Report - pdf report as required with Excel document for selected features with index.

2_SourceCode – source code (Java Maven Project) & Makefile (see step 2 for running the code)

3_AWSSyslogFiles_Training – plain text syserr and sysout files for a successful run of model training program on the full labeled data set.

4_AWSSyslogFiles_Prediction - plain text syserr and sysout files for a successful run of prediction program on the full unlabeled data set.

5_AWSOutputFiles - Final prediction output file in CSV format for the full unlabeled data set.

Step 2: Steps to build and execute the program on local and AWS (loc.: 2_SourceCode/FinalProject)

2.1 Copy the labeled data set into FinalProject/input/labeled folder.

2.2 Copy the unlabeled data set into FinalProject/input/unlabeled folder.

2.2 Open terminal and cd to this FinalProject directory

2.3 There are two different programs, one for training the model ("`ModelTraining`") and other for prediction ("`ModelPrediction`"). To switch to any of this program, update below variable in makefile.

job.name=mr.scala.project.ModelTraining - for training
job.name=mr.scala.project.ModelPrediction - for prediction

You need to first run the model training program which will train the random fores classification model on entire labeled data set and writes it to outputModel folder.
After successful run of training program, modify the job name in makefile to prediction program, and run the second program. this will generate the prediction output using the previous training model.

2.4 To execute each program in Local

- Run the below command to execute the program in local
  "make alone"
- Output files will be generated in the same directory.

2.5 To execute each program on AWS

- Run the below commands to execute the program on AWS

"make upload-input-aws" : command to upload the input to aws (it will copy the input files into dspatel28 bucket)

"make cloud" : to execute the program/job on AWS

- After execution, you can find the output results and logs in dspatel28 bucket
- At last, you can delete all the data from dspatel28 bucket by running below command:

"make delete-s3-aws"