**Big-Data-Analytics**

**Analysis using MachineLearning and Pandas**

**Highlights:** Classification and Regression Machine learning libraries, Pandas, sklearn, matplotlib, seaborn.

**File:** *Q1-classifiers.ipynb*

**Problem statement:**

An organization would like to keep its good employees because they usually possess valuable experience. Therefore, it is useful for a company to predict the likelihood an employee will leave so mitigations can be done.

Applied various Classification techniques to build model and evaluated model:

* Decision Tree
* Random Forest
* Naïve Bayes
* Support Vector Machine(SVM)

Dataset splitting into training and testing dataset.

Evaluated model using various techniques like:

* Confusion matrix
* Accuracy score
* Classification Report
* ROC

**File:** *Q1-classifiers.ipynb* (For main problem) *and Q2-SUB-quest.ipynb* (For 3 sub questions)

**Problem statement:**

Which country has the happiest citizens in the world? The World Happiness Report may tell you something about that. In this problem, you are given world happiness survey result data sets for three years (2015, 2016, and 2017). The ranking in terms of happiness of the countries is provided. Your job is to use regression to create a model that can predict the ranking (note that ranking may be an indirect results) as close as possible to the actual results. To help you understand better of the meaning of the attributes, you can refer to **http://worldhappiness.report/faq/**. You could start your model building using one year’s data and apply it on other two years.

In addition, use the basic statistics and other data manipulation methods to answer the following questions (with code):

* From 2015 to 2017, which country’s happiness ranking increased the most?
* From 2015 to 2017, which country’s happiness ranking decreased the most?
* For each year, provide the ranking of the happiest continents.

For main problem: (*Q1-classifiers.ipynb*)

Model building using linear regression to predict the happiness rank.

* Model was based on 2015 dataset which was split into training and testing dataset.
* Applied same model on 2016 and 2017 dataset which has small variation in the structure.

Evaluated model using various techniques like:

* Mean Absolute Error
* Mean Squared Error
* Root Mean Squared Error

For main problem: (*Q2-SUB-quest.ipynb*)

1. Country’s happiness ranking increased the most

Max of difference between Ranking of 2017 and 2015 was calculated.

1. Country’s happiness ranking decreased the most

Max of difference between Ranking of 2017 and 2015 was calculated

### For each year, ranking of the continents based on Happiness Rank

Median statistics usage by grouping based on continents.

========================================================================

**Map Reduce Programs**

**Problem statement:**

**1) *To fetch 10 top most visitors to White House in year 2012 by (NAMELAST, NAMEFIRST, NAMEMID)***

**To run the program:** python prog1.py 10 // (10 is passed as parameter to get the top 10 most visitors)

The prog1.py script internally runs the ‘map\_reduce\_prog1.py’ program which runs in the emr. (with option -r emr)

**Description of program:** The main program runs the “map\_reduce\_prog1.py” program internally and writes output to the file ‘out\_file.txt’. The main program then reads the out\_file.txt’ and creates dictionary with key (visitor’s name) and value (frequency of count) and displays the output of top n most\_common visitor details.

**2) *To fetch 10 top most visited people in White House in year 2012 by (visitee\_namelast, visitee\_namefirst)***

**To run the program:** python prog2.py 10 // (10 is passed as parameter to get the top 10 most visitors)

The prog1.py script internally runs the ‘map\_reduce\_prog2.py’ program which runs in the emr. (with option -r emr)

**Description of program:** The main program runs the “map\_reduce\_prog2.py” program internally and writes output to the file ‘out\_file.txt’. The main program then reads the out\_file.txt’ and creates dictionary with key (visitee name) and value (freqeuncy of count) and displays the output of top n most\_common visitees details.

**3) *To fetch Meeting Location and Room which was used for highest number of time was programmed***

**To run the program:** map\_reduce\_prog3.py

**Description of program:** This program takes sample\_white\_house file as input , and does 2 stage map-reduce program. From First stage of map-reduce Meeting-location and Meeting room with frequency of visited is obtained.

In second stage, Location and room with maximum count is obtained.