**CASE STUDY**

***Perform Following Machine Learning Tasks on data given in related folders. Also show related charts.***

***Try to perform case study in Python.***

**Regression**

1. **Thunder Basin Antelope Study**  
     
   The data (X1, X2, X3, X4) are for each year.  
   X1 = spring fawn count/100  
   X2 = size of adult antelope population/100  
   X3 = annual precipitation (inches)  
   X4 = winter severity index (1=mild,  
   5=severe)
2. **Systolic Blood Pressure Data**  
     
   The data (X1, X2, X3) are for each patient.  
   X1 = systolic blood pressure  
   X2 = age in years  
   X3 = weight in pounds
3. **Test Scores for General Psychology**  
     
   The data (X1, X2, X3, X4) are for each student.  
   X1 = score on exam #1  
   X2 = score on exam #2  
   X3 = score on exam #3  
   X4 = score on final exam
4. **Hollywood Movies**  
     
   The data (X1, X2, X3, X4) are for each movie  
   X1 = first year box office receipts/millions  
   X2 = total production costs/millions  
   X3 = total promotional costs/millions  
   X4 = total book sales/millions
5. **Health**  
     
   The data (X1, X2, X3, X4, X5) are by city.  
   X1 = death rate per 1000 residents  
   X2 = doctor availability per 100,000 residents  
   X3 = hospital availability per 100,000 residents  
   X4 = annual per capita income in thousands of dollars  
   X5 = population density people per square mile

**Classification**

1. Apply Classification algorithm on given data of San Francisco International Airport Report on Monthly Passenger Traffic Statistics by Airline.

**DESCRIPTION**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

**SUMMARY**

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).  
   The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Input variables:

**bank client data:**

1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')

**related with the last contact of the current campaign:**

8 - contact: contact communication type (categorical: 'cellular','telephone')  
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

**other attributes:**

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

**social and economic context attributes**

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)  
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)  
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

**Clustering**

# *Clustering the Countries by using Unsupervised Learning for HELP International*

### **Objective:**

To categorise the countries using socio-economic and health factors that determine the overall development of the country.

### **About organization:**

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

### **Problem Statement:**

HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, CEO has to make decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.