# **Assignment Name - Analytics basics**

**Problem Statement -**

**Answer the following questions to the best of your knowledge including the concepts taught to you in the level. You can find the IMDB\_data in the folder containing this file.**

1. Write R code using data “IMDB\_data” to
   1. Load CSV in R by skipping second row.
2. fileEncoding="UCS-2LE"

#remove all the objects stored

rm(list=ls())

#set current working directory

setwd("Documents")

df=(read.csv("imdb\_data.csv",header=T,sep=",",fileEncoding="UCS-2LE")[-2,])

* 1. Extract the unique genres and its count and store in data frame with index key.
  2. Convert the required data types
  3. Sort the genre by its name
  4. Create new variable whose values should be square of difference between imdbrating and imdbvotes.

1. Write Python code using data “IMDB\_data” to
   1. Load CSV in R by skipping second row.
   2. Extract the unique genres and its count and store in data frame with index key.
   3. Convert the required data types
   4. Sort the genre by its name
   5. Create new variable whose values should be square of difference between imdbrating and imdbvotes.
2. Define problem category for below problem statement  
    “A chemist wants to find some interesting patterns in which patients are behaving upon administering the drug”

Ans : The chemist type problem statement falls in Unsupervised Learning problem category. Since the chemist will not focus on predetermined attributes, nor does it predict a target value. Rather, chemist wants to find hidden structure and relation among data. Unsupervised learning refers to techniques that find patterns in unlabeled data, or data that lacks a defined response measure. In Unsupervised learnings, we have less information about objects. So here we try to find some similarities between groups of objects and include them in an appropriate cluster which is called as segmentation clustering. Some object can differ hugely from all clusters, in this way we can say that these objects are anomalies.

1. How will you select suitable machine learning algorithm for a problem statement

Ans : understanding the problem statement and type of problem is the first go.Depending on dependant variable or (lack of it), we need to define the task. They are classified into below 3 categories.

* regression if the dependant variable is continuous
* classification if the dependant variable is categorical
* unsupervised if there’s no dependent variable

there are more possible tasks and even specialised tasks. Unsupervised learning is a world in itself for instance. There are also many types of regression and classification.

The take away point here though is that each algorithm is often (but not always) specific for a task or a specialised area of that task. Knowing the task creates a shorter list of algorithms.

**Understand the data**

Not all data is adequate for all algorithms. In order to choose the best algorithm we need to know their constraints and assumptions. we can transform data to sort some of the problems. This is often the case of feature engineering independent variables.

Other times the dependent variable is not adequate, for instance if we have a multi class classification problem we won’t use a binary classifier.

1. Define one problem statement for Education industry?

Ans :

Problem statement : The problem statement is to identify the quality of education system in the particualar school. To check whether the students are getting the better understanding and knowledege from our faculty. What is the average of the students who were successful in the completion of the based on their standards. What needs to be changed to have a better outcomes. What is the current percentile and count of the faculty and students.

🡪 this problem statement can be classifed based on the classes and then analysis can be done for the no of students and faculty and the average , max and min academic progress can be calculated based on which the organization can identify the drawbacks to improve their quality of education.

# **Evaluation Basis**

This project will be evaluated on following basis -

1. **Comprehensibility of the answer :** You have to answer the given question with full explanation, background information and examples(if required).
2. **Quality of code** : Write the best possible implementation from your side ensuring the coding practices and conventions taught to you during training.

## **Deliverables from Candidate**

1. A Doc or Pdf file containing your answers to theory questions
2. Code files - R and Python separate.
3. Notes to explain your code logic (if required)

**Warning** - Do not submit incomplete or wrong assignments. They will result in negative skill score. Also, you will not a get a second attempt at the assignment! So whatever score you get will be the final score