2.

Python code for given problem Statement**:(Pyhton code Attached with assigment as jupyter notebook)**

import os

k = os.getcwd()

print(os.listdir(k))

#importing CSV file without second row

import pandas as pd

df = pd.read\_csv("IMDB\_data.csv", encoding= 'ISO-8859-1' , skiprows=range(2,3))

print(type(df))

df.head(5)

# Extracting the unique genres and its count

vc = df.Genre.value\_counts() #gives unique genre and thier respective count

lis1 = list(vc.index)

lis2 = list(vc.values)

# Alternatives to get unique values

#df.Genre.unique()

#pd.unique(df.Genre)

#df['Genre'].unique()

pd.DataFrame({"Unique\_Genre":lis1,"Count": lis2}) # storing the Series in Dataframe with index

#since all the columns are of dtype object(String) we only need imdbRating and imdbVotes as integers for further transformations

#converting imdbvotes and imdbRating to numeric

df['imdbVotes']= pd.to\_numeric(df['imdbVotes'],errors='coerce')

df['imdbRating']= pd.to\_numeric(df['imdbRating'],errors='coerce')

print("After conversion\n")

df.info()

# Check if any feature has missing values

df.notnull().all()

# fill missing values with previous record

df['imdbRating']= df['imdbRating'].fillna(method = 'pad')

df['imdbVotes']= df['imdbVotes'].fillna(method = 'pad')

#calculate square of difference of imdbRating and imdbVotes

df['Sqr\_diff'] = (df['imdbRating']-df['imdbVotes'])\*\*2

print(df['Sqr\_diff'])

3. According to given problem statement

'A chemist wants to find some interesting patterns in which patients are behaving upon administering the drug'

We shall break the above problem in following way:

* Input: The Drug which has been prescribed to patients.
* Output: Behaviour of different patients upon application of the given drugs.

Here we must understand that drugs given to set of patients are known i.e input to the model is known, but we do not have predefined groups of known behavior of patients upon application of known drugs.

i.e Labels are missing which is fundamental requirement of any classification problem.

So to implement this problem we have to study the behavior patterns among patients and draw similarity among available sets of pattern which is nothing but

Clustering .

So this problem can be termed as Unsupervised learning problem.

4.

Following factor must be considered while choosing the machine learning algorithm :

* Understanding the Data:
* The size and type of data must be considered while choosing ML algo

One must check if data is numerical(continuous) or categorical. We use different algorithm for different type of data e.g Regression needs numerical data while classification needs categorical data.

* Spread of data , central tendency and correlation among data must be analysed to select appropriate algorithm
* Principal component Analysis must be performed for dimensionality reduction for some algorithms which are sensitive towards range of features.
* Appropriate measures must be undertaken to tackle Missing values in the data sample since some algorithm are sensitive to missing values .
* If not taken care of Outliers can affect the accuracy of models especially regression algorithms, Outliers must be normalized for such scenarios.
* Feature Engineering:
* Number of features to be made out of existing data and the type of feature is very crucial in ML lifecycle.
* Model can perform to it’s fullest if appropriate feature is selected.
* Categorize the problem
* Categorize by the input: If it is a labeled data, it’s a supervised learning problem. If it’s unlabeled data with the purpose of finding structure, it’s an unsupervised learning problem. If the solution implies to optimize an objective function by interacting with an environment, it’s a reinforcement learning problem.
* Categorize by output: If the output of the model is a number, it’s a regression problem. If the output of the model is a class, it’s a classification problem. If the output of the model is a set of input groups, it’s a clustering problem.
* Requirement of the Application:
* There is always a tradeoff between accuracy , Run time and Utilization of computing resources.
* One can get 100% accuracy and minimal run time provided infinite computing resources which is hypothetical scenario.
* So algorithm must be selected according to application reqirement.
* E.g For problems were real time prediction is required like self driving cars, A algorithm must be selected which can bulid the model in real time with newly arrived data.
* Once the model is decided it is advisable to test it with different sets of train data to analysed it’s behavior .
* Scalability and complexity of the model must be considered .

5. Problem statement for education indusrtry is as follows:

Problem:

Inspite of Government's streneous efforts to to provide free and compulsory primary education to every kid under 14 years of age under ,The Right of Children to Free and Compulsory Education Act or Right to Education Act. The number of students getting further education is considerably low .

In my opinion following factors are reponsible for above phenomena.

1. Scarcity of sufficient teaching staff.

2.Socio-Economic background of student.

3. Lack of interest in further studies due to delayed primary education resulting in increased drop out rates.

Solution:

Levaraging the UADAI- AADHAR data available with centre and census data of year 2011 as well as by conducting live polls data can be collected in a single consolidated data bases and can be modeled into multidimentional data warehouse having many Dimension tables to address above problem.

Once data is collected and data pre-processing is carried out with appropriate EDA analysis. Feature Engineering phase can be initiated to get important features from data with help of SME's.

Following are some of important features which can be considered:

For Students

1. parents alive

2. family's annual income.

3. No. of siblings

4. Rank of the student among the siblings

5. Mother Tongue

6. Parents Education.

7. Parents Occupation.

8. Caste/Category/Religion of student

9. Gender

10. Domicile of State

11.Age

12. Age at which he/she is enrolled in school.

13. Grades till date

For Teachers/Applicants applying as a teacher

1. Age

2. Education

3. Domicile of State

4. Mother Tongue

5. Reason of applcation

6. Marks in entrance exam

Once important features are extracted data must be divide in train and test.

The goal of this model would be to decide if student will continue education or not

i.e Binary Classifier

Since data will be collected over the years It can be labled and Supervised classification can be performed.

Following ML algorithm can be used to achieve this:

1. Logistic Regression.

2. Random Forest

3. XGboost

ROC curve and AUC score can be used to compare this models on given dataset.