Data Science and Business Analytics

Task1: Prediction using Supervised ML

Problem Statement: Predict the percentage of an student based on the no. of study hours.

submited by: Laxman Velip

In [ ]:

```
#lets change our current wiorking directory, where our dataset lies
import os
os.chdir('C:/Users/laxman/Documents/My Bluetooth/data_csv')
```

In [ ]:

```
#Importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:

```
#now we will import our dataset
dataset=pd.read_csv('students_score.csv')
print(dataset.head(5)) #first 5 rows of dataset
print(dataset.isna().sum()) #checking if there are any missing values
print(dataset.corr())
```

In [ ]:

```
#we need to divide our data into input and output variable
hours=dataset.iloc[:,[0]] #input data(study hours)
score=dataset.iloc[:, [1]] #output data(student score)
```

In [ ]:

```
#training the model
from sklearn.linear_model import LinearRegression
model=LinearRegression()
model.fit(hours,score)
```

In [ ]:

```
#finding intercept and slope
print('Intercept C: ', model.intercept_)
print('Coefficient m: ', model.coef_)
```

In [ ]:

```
#predicting the percentage of students
predicted_score=model.predict(hours)
```

In [ ]:

```
#checking performance of model
from sklearn.metrics import r2_score, mean_squared_error, accuracy_score
print('R Squared error: ', r2_score(predicted_score, score))
print('Root mean squared error: ', np.sqrt(mean_squared_error(predicted_score, score)))
```

In [ ]:

```
#visualizing our linear regression model
plt.scatter(hours,score, c='blue')
plt.plot(hours, predicted_score, c='black', linewidth=3)
plt.xlabel('Hours of study')
plt.ylabel('Student score')
plt.show()
```

In [ ]:

```
#now we will predict the score, if student studies 9.25 hours/day
test_hour=[[9.25]]
test_score=model.predict(test_hour)
print(test_score)
```

We can clearly see that if student studies for 9.25 hours/day, score will be 92.90985477