THE SPARKS FOUNDATION

DATA SCIENCE AND BUSINESS ANALYTICS INTERNSHIP

AUTHOR – UMANG AGGARWAL

Task Name - Prediction using Supervised ML

- Predict the percentage of a student based on the no. of study hours.
- This is a simple linear regression task as it involves just 2 variables.
- You can use R, Python, SAS Enterprise Miner or any other tool
- Data can be found at http://bit.ly/w-data
- What will be predicted score if a student studies for 9.25 hrs/ day?

Solution

Step 1: Import Necessary Libraries

import pandas as pd → For reading dataset

from sklearn.linear_model import LinearRegression → For implementing linear model

from sklearn.model_selection import train_test_split → For splitting data into train and test set

from sklearn.metrics import mean_squared_error → For Evaluating Model

import matplotlib.pyplot as plt → To plot best fit line

Step 2: Read Dataset

data = pd.read_csv("http://bit.ly/w-data")

data.head(10) → To see first 10 rows of loaded data

Step 3: Data Preprocessing - Checking for missing value

data.isnull().sum() → For checking missing value

data.info() → For getting more info

Step 4: Setting Dependent and Independent Variable

```
x = data['Hours']

y = data['Scores']

x = x.values.reshape(len(x),1)

y = y.values.reshape(len(y),1)
```

Step 5: Building Model

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

lr = LinearRegression()
```

→ Building model

```
lr.fit(x_train, y_train)
```

→ Fitting the model

Step 6: Evaluating Mean Squared Error

```
y_pred = lr.predict(x_test)

mse = mean_squared_error(y_test,y_pred)

print("Mean Square Error = ", mse)
```

Step 7: Plotting Best Fit Line

```
line = lr.intercept_ + lr.coef_ * X

plt.scatter(X, y)

plt.plot(X, line, color='green', linewidth=3);

plt.show()
```

Step 8: Making Prediction

```
Y = lr.intercept_ + lr.coef_ * 9.25

print("Predicted Score if a student studies for 9.25 hrs/day =  ", Y)
```

**********THANKYOU**********