The Sparks Foundadtion

Data Science and Business Analytics Internship | GRIPJUL'21

Task-1 : Prediction using Supervised ML

Author - Tanushree gaur

Problem Statement : Prediction of a student based on the number of study hours.

Importing necessary libraries

In [11]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

Importing data

In [13]:
```
url="http://bit.ly/w-data"
df=pd.read_csv(url)
print("Data imported succesfully")
df.head(25)
```

```
Data imported succesfully
```

Out[13]:

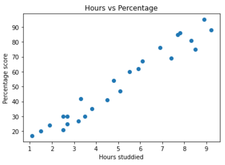
	Hours	Scores
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30
5	1.5	20
6	9.2	88
7	5.5	60
8	8.3	81
9	2.7	25
10	7.7	85
11	5.9	62
12	4.5	41
13	3.3	42
14	1.1	17
15	8.9	95
16	2.5	30
17	1.9	24
18	6.1	67
19	7.4	69
20	2.7	30
21	4.8	54
22	3.8	35
23	6.9	76
24	7.8	86

In [14]:
```
df.shape
```

Out[14]:
```
(25, 2)
```

Plotting the distribution of scores

In [18]:
```
plt.title('Hours vs Percentage')
plt.xlabel('Hours studdied')
plt.ylabel('Percentage score')
plt.scatter(df.Hours,df.Scores)
plt.show()
```



Creating training and test dataset

In [38]:
```
x=df.iloc[:, :-1].values
y=df.iloc[:, 1].values
```

Splitting the data into training and test tests

In [39]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y)
```

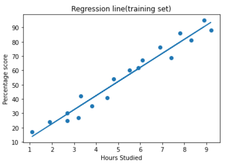
Creating regression model and training the model

In [41]:
```
from sklearn.linear_model import LinearRegression
regressor=LinearRegression()
regressor.fit(x_train, y_train)
print("score:", regressor.score(x_train,y_train))
print("trainig complete")
```

```
score: 0.960767489465233
trainig complete
```

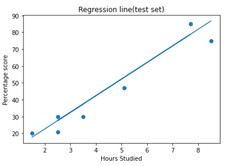
Plotting the regression line on training set

In [44]:
```
line=regressor.coef_*x_train +regressor.intercept_
plt.title('Regression line(training set)')
plt.xlabel('Hours Studied')
plt.ylabel('Percentage score')
plt.scatter(x_train,y_train)
plt.plot(x_train,line)
plt.show()
```



Plotting the regression line on test set

In [46]:
```
line=regressor.coef_*x_test +regressor.intercept_
plt.title('Regression line(test set)')
plt.xlabel('Hours Studied')
plt.ylabel('Percentage score')
plt.scatter(x_test,y_test)
plt.plot(x_test,line)
plt.show()
```



Predicting the scores

In [56]:
```
y_pred=regressor.predict(x_test)
y_pred
```

Out[56]:
```
array([86.66171395, 53.18549541, 27.58603418, 37.43198081, 27.58603418,
       78.78495665, 17.74008755])
```

Comparing actual vs Predicted scores

In [57]:
```
df1=pd.DataFrame({'Actaual': y_test, 'Predicted': y_pred })
df1
```

Out[57]:

	Actaual	Predicted
0	75	86.661714
1	47	53.185495
2	21	27.586034
3	30	37.431981
4	30	27.586034
5	85	78.784957
6	20	17.740088

evaluating the model

In [59]:
```
from sklearn.metrics import r2_score
print('Accuracy:',r2_score(y_test,y_pred)*100,'%')
```

```
Accuracy: 92.2269211719642 %
```

Our model is giving 92% accuracy

Predicting the score

In [61]:
```
pred=regressor.predict([[9]])
print('No. of Hours studied={}'.format(9))
print('Predicted Score={}'.format(pred[0]))
```

```
No. of Hours studied=9
Predicted Score=91.58468726323764
```

Conclussion:

From the above result we can conclude that if a student studies for 9 hours, then his score will be 91.58 marks

Completed TASK-1

Thankyou

Tanushree Gaur

In [ ]:
```
```