TASK 1: PREDICTION USING SUPERVISED MACHINE LEARNING

AUTHOR : Abarna

In [7]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [8]:

```
#reading the data
data=pd.read_csv("https://raw.githubusercontent.com/AdiPersonalWorks/Random/master/student_scores%2
%20student_scores.csv")
```

In [9]:

```
#return the first 5 rows by default
data.head()
```

Out[9]:

	Hours	Scores
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30

In [10]:

```
data.tail()
```

Out[10]:

	Hours	Scores
20	2.7	30
21	4.8	54
22	3.8	35
23	6.9	76
24	7.8	86

In [11]:

```
data.columns
```

Out[11]:

```
Index(['Hours', 'Scores'], dtype='object')
```

In [12]:

```
#returns all the element
data
```

Out[12]:

	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 [13]:

```
data.describe()
```

Out[13]:

	Hours	Scores
count	25.000000	25.000000
mean	5.012000	51.480000
std	2.525094	25.286887
min	1.100000	17.000000
25%	2.700000	30.000000
50%	4.800000	47.000000
75%	7.400000	75.000000
max	9.200000	95.000000

In [14]:

```
data.corr()
```

Out[14]:

	Hours	Scores
Hours	1.000000	0.976191
Scores	0.976191	1.000000

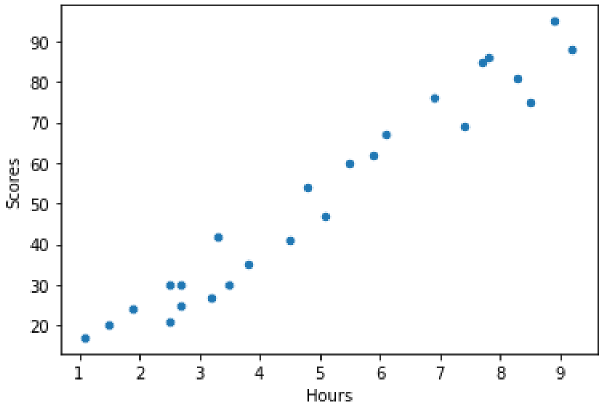
In [15]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25 entries, 0 to 24
Data columns (total 2 columns):
Hours     25 non-null float64
Scores    25 non-null int64
dtypes: float64(1), int64(1)
memory usage: 528.0 bytes
```

In [16]:

```
data.plot(kind='scatter',x='Hours',y='Scores');
plt.show()
```



In [17]:

```
data.shape
```

Out[17]:

```
(25, 2)
```

visualizing the data

In [18]:

```
x=data.iloc[0:,:-1].values
x
```

Out[18]:

```
array([[2.5],
       [5.1],
       [3.2],
       [8.5],
       [3.5],
       [1.5],
       [9.2],
       [5.5],
       [8.3],
       [2.7],
       [7.7],
       [5.9],
       [4.5],
       [3.3],
       [1.1],
       [8.9],
       [2.5],
       [1.9],
       [6.1],
        [7.4],
        [2.7],
        [4.8],
        [3.8],
        [6.9],
        [7.8]])
```

In [19]:

```
y=data.iloc[:,1].values
y
```

Out[19]:

```
array([21, 47, 27, 75, 30, 20, 88, 60, 81, 25, 85, 62, 41, 42, 17, 95, 30,
       24, 67, 69, 30, 54, 35, 76, 86], dtype=int64)
```

In [20]:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```

In [21]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=50)
```

In [22]:

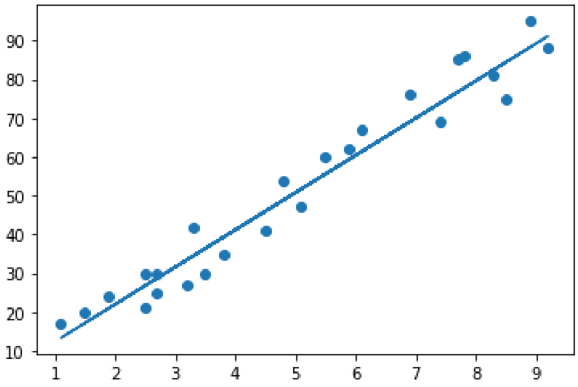
```
reg=LinearRegression()
reg.fit(x_train,y_train)
```

Out[22]:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

In [23]:

```
a=reg.coef_
b=reg.intercept_
line=a*x+b
plt.scatter(x,y)
plt.plot(x,line)
plt.show()
```



In [24]:

```
pred_y=reg.predict(x_test)
```

In [25]:

```
print(np.concatenate((pred_y.reshape(len(pred_y),1),y_test.reshape(len(y_test),1)),1))
```

```
[[88.21139357 95.        ]
 [28.71845267 30.        ]
 [69.02012231 76.        ]
 [39.27365186 35.        ]
 [13.36543566 17.        ]]
```

In [26]:

```
hr=[9.25]
result=reg.predict([hr])
print("The predicted score of a student who studies for 9.25hr/day = {}".format(result[0],2))
```

```
The predicted score of a student who studies for 9.25hr/day = 91.56986604454477
```

Evaluation

In [27]:

```
from sklearn import metrics
from sklearn.metrics import r2_score
```

In [28]:

```
print("Mean Absolure error: ",metrics.mean_absolute_error(y_test,pred_y))
print("Mean Squared error: ",metrics.mean_squared_error(y_test,pred_y))
print("R2 Score: ",r2_score(y_test,pred_y))
```

```
Mean Absolure error:  4.5916495300630285
Mean Squared error:  25.58407829653998
R2 Score:  0.971014141329942
```

In [ ]: