CREATIVE INSTITUTE DATA SCIENCE

Business Problem

How many hours need to do the study to get 99% marks?

How many hours need to do the study to pass the exam?

If I will do study x(4) hours per day so how much marks I will get

Import the library and load dataset

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df  = pd.read_csv("student_info.csv")
df.head()
```

now find the shape of data set

```
df.shape
```

Discover and visualize the data to gain insights

```
df.info()
```

we seen that there is sum missing values in dataset

also check describe of the data

```
df.describe()
```

now visualize the data through scatter plot

```
plt.scatter(x =df.study_hours, y = df.student_marks)
plt.xlabel("Students Study Hours")
plt.ylabel("Students marks")
plt.title("Students Study Hours vs Students marks")
plt.show()
```

we have clearly seen that there is linear related data

now preprocess the data

data cleaning

find out there is null value present in the dataset.

```
df.isnull().sum()
```

there is null value in study hours variable

now find out mean of variable

```
df.mean()
```

now fill the missing value thorugh mean

```
df2 = df.fillna(df.mean())
df2
```

now check the new dataset that there is any missing value

```
df2.isnull().sum()
```

now split the dataset

```
X = df2.drop("student_marks", axis = "columns")
y = df2.drop("study_hours", axis = "columns")
print("shape of X = ", X.shape)
print("shape of y = ", y.shape)
```

now use train_test_split to train the model

```
from sklearn.model_selection import train_test_split
X_train, X_test,y_train,y_test = train_test_split(X,y, test_size = 0.2,
random_state=51)
print("shape of X_train = ", X_train.shape)
print("shape of y_train = ", y_train.shape)
print("shape of X_test = ", X_test.shape)
print("shape of y_test = ", y_test.shape)
```

Select a model and train the model

Now import linear regression library

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
```

now use fit method to implement data

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

now find co-efficient of the dataset

```
lr.coef_
```

now find intercept of the dataset

```
lr.intercept_
```

now we have value of slope, intercept and if student study 4 hours .what will be the result? Based on maths

```
m = 3.93
c = 50.44
y  = m * 4 + c
y
```

now predict the value of if student study 4 hours.

```
lr.predict([[4]])[0][0].round(2)
```

now predict the value of y

```
y_pred  = lr.predict(X_test)
y_pred
```

we have get value of 2D array . now convert into this dataframe

```
pd.DataFrame(np.c_[X_test, y_test, y_pred], columns = ["study_hours",
"student_marks_original","student_marks_predicted"])
```

now find accuracy of the model

```
lr.score(X_test,y_test)
```

now use scatter plot the visualise the data

```
plt.scatter(X_train,y_train)
```

```
plt.scatter(X_test, y_test)
plt.plot(X_train, lr.predict(X_train), color = "r")
```