Assignment No 3

- **Aim**

Split Sample data into training and test sets. (use suitable data set).

- **course objectives**

1. Student will apply data science concepts and methods to solve problems.

- **course outcomes**

CO2 : Demonstrate the classification, clustering and etc-in large data sets

CO6 : Apply data science concepts and methods to solve problems.

- **software and hardwares requirements**

sr. no.	requirements (softwares, hardwares)	specifications.
1.	Python Jupyter	version V.7.0.6
2.	Anaconda Navigator	version V.7.2.6
3.	computer /PC	I5 version, 64 bits, 8GB RAM.
4.	Excel, Google crome.	Excel software.

- Theory

Split a Dataset into Train and Test sets using python

1. Here we will discuss how to split a dataset into a Train and test sets in python.

2. The train-test split is used to estimate the performance of a machine learning algorithms that are applicable for a prediction-based algorithms.

3. This method is a fast and easy procedure to a perform such that we can compare our own machine learning model result to machine result.

4. By Default, the test set is split into 30% of an actual data and training set is split into 70% of the actual data.

5. we need a split dataset into train and tests to an evaluate how well our machine learning model, and the statistics of the train set are known.

6. The second test set is called the test dataset this set is solely used for predictions.

7. The simplest way to split the modelling dataset into training and testing sets is to assign 2/3 data points to the former and the remaining one-third to the latter.

Dataset Splitting

1. Scikit-learn alias sklearn is most useful and a robust library for a machine learning in python.

2. The scikit-learn library provides us with the model_selection module in which we have the spitter fuction train_test_split().

Syntax

```
train_test_split(*arrays, test_size = None, train_size = None
random_state = None, shuffle = True, stratify = None)
```

1. *arrays : Inputs such as lists, arrays, data frames, or matrices

2. test_size: This is a float value whose value ranges between 0.0 and 0.1 . it represents the propo. rtion of our test size. its default value is None.

3. Train_size: This is a float value whose range between 0.0 and 0.1 .it represents the proportions of our train size. its default value is none.

4. random_state : This parameter is used to control a shu. ffled applied to data before applying the split.

5. shuffle : This parameter is used to shuffle the data before splitting. its default value is true.

6. stratify : This parameter is used to split the data in a stratified data fashion.

Example of splitting data.

First download the csv file used in the following for splitting

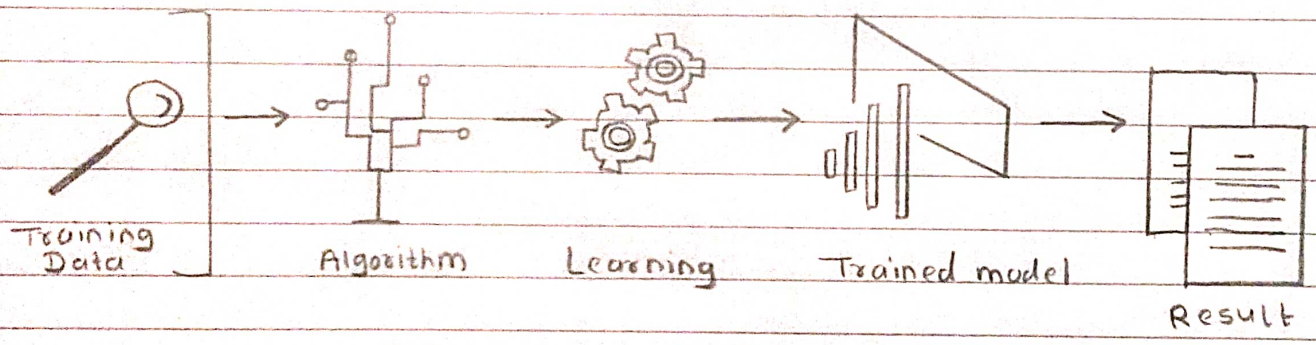
```
#import modules libraries
 import pandas as pd
 from sklearn.leaear_model import LinearRegression
 from sklearn. model-selection import train_test_split

# Read the dataset
  df = pd. read_csv ('Real estate.csv')

# get the locations
  X = df. iloc [:, :-1]
  Y= df. iloc [:, -1]

# split the dataset
  X_train , X_test , Y_train , y_test = train_test_split ( X, y,
  test_size = 0.05 , random_state = 0)
```

In above example , we import the required libraries , after that read the csv (dataset) file, The variable df now contain data frame of 'house price' then its column X, Y predict and then random state helps us get the same random split each other

What does train Dataset

1. The training data is the biggest (in size) subset of the origional dataset, which is used to train or fit the machine learning model.

2. firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.



(fig a: process of Training data)

Syntax

```
#Train a Data set

train_X = X [:80]
train_y = y [:80]
```

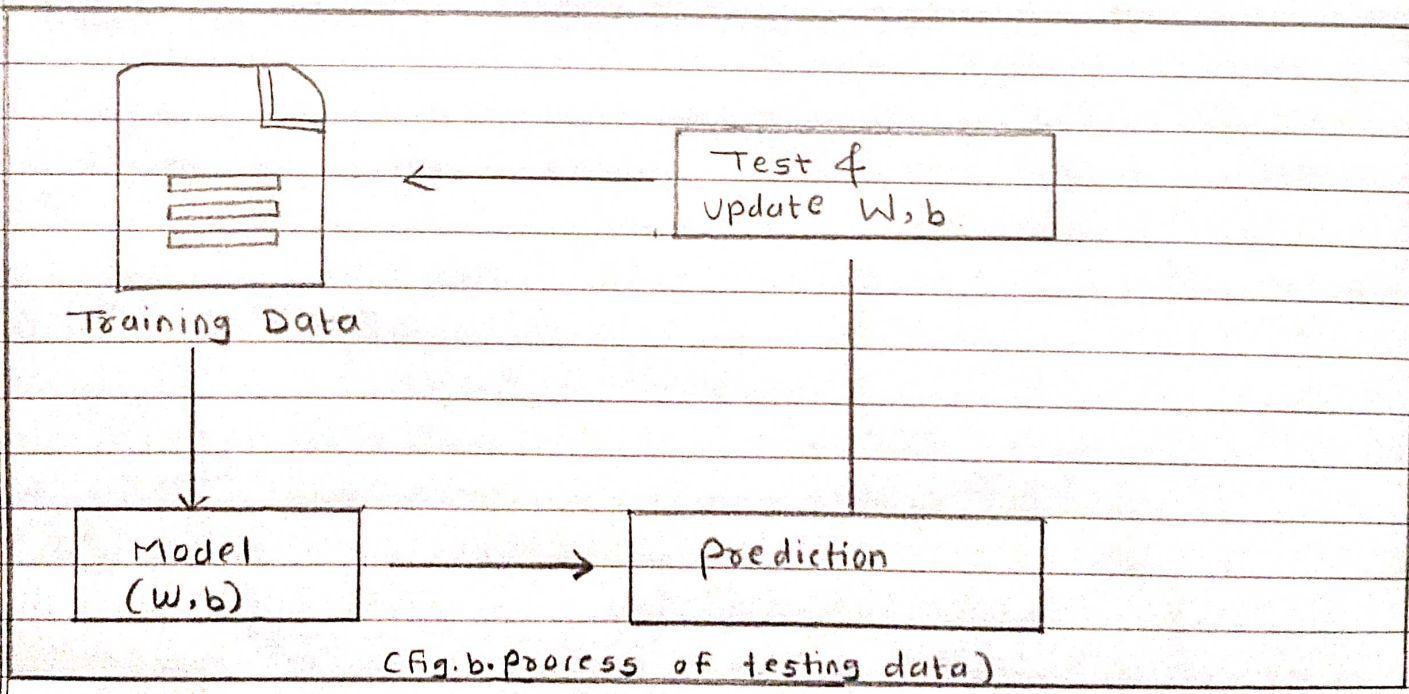
What does Test Dataset

1. In machine learning, we use testing data to ensure the model works for the given testing data.

2. The testing data should meet two criteria : It is represent the actual dataset that the model will be used on.

3. This means that the testing data should have the same distribution of features as the actual dataset.

4. Test datasets are small contrived datasets that lets you test a machine learning algorithm or testing.



(Fig. b. Process of testing data)

Sytax

```
test_x = X[80:]
test_y = y[80:]
```

- **conclusion**

In this practical, I have splitting data into training and testing sets is a crucial in machine learning to evaluate model performance. By training on a subset, we can data and testing on another unseen subset, wel can assess how well the model generalizes to new data. this helps prevent overfitting and provides a reliable estimate of the model's performance on unseen data.