**Data Pre-Processing**

**In this milestone, we will be pre-processing the dataset that is collected.**

**Pre-processing includes**

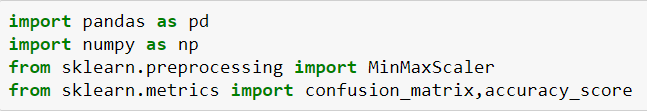
1. Handling the null values.
2. Handling the categorical values if any.
3. Normalize the data if required.
4. Identifying the dependent and independent variables.
5. Split the dataset into train and test sets.

**1)Import Required Libraries**

**Step 1**- Launch Jupyter notebook through anaconda navigator or anaconda prompt.

**Step 2**- Create a new notebook by clicking on "new" button on the top right corner of the page.

The libraries can be imported using the import keyword. Insert commands as shown below.



**2)Read The Dataset**

The dataset is read as a **dataframe** by using pandas library. Insert the commands as shown below

(Here ds is referred as **dataframe & pd** is the alias name given to pandas library).

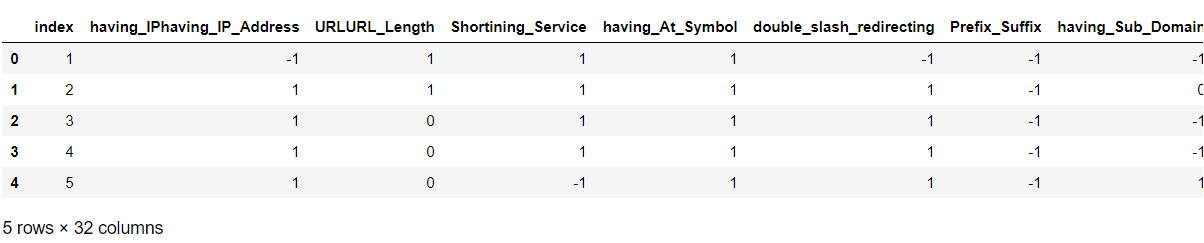
To read the data set we have use the following code:

#import Dataset

Ds=pd.read\_csv(“dataset\_website.csv”)

Ds.head()

Sample out put



**3)Handling Null Values**

Checking for Null values in a dataset and handling if any:

In this activity, we will check if there are any null values in a dataset and fill/handle them.

To know if there are any null values present in a dataset **isnull()** method can be used.

Input the commands as shown below to check for **null**values.

the command**ds.isnull().any()** returns true if null values are present.

Here, the dataset which we have used doesn’t have any null values.

**4)Splitting data into independent and dependent variables**

**Identifying Independent & dependent variables:**

In this activity, the dependent and independent variables are to be identified. The last column (Result) in the dataset is the dependent variable which is dependent on the 30 different factors. The independent columns are considered as  x and the dependent column as y.

#splitting data as independent and dependent

#removing the index coloumn in independent and dataset

X=ds.iloc[:,1:31].values

Y=ds.iloc[:,-1].values

Print(x,y)

**Splitting the data:**

After identifying the dependent and independent variables, the dataset now has to be split into two sets, one set is used for training the model and the second set is used for testing how good the model is built. The split ratio we consider is 80% for training and 20% for testing.

#splitting data into train and test

From sklearn.model\_selection import train\_test\_split

X\_train,x\_test,y\_test=train\_test\_split(x,y,test\_size+0.2,random\_state=0)