SPLITTING THE DATASET

SPLITTING THE DATASET INTO DEOENDENT AND INDEPENDENT VARIABLES:

- In machine learning, the concept of dependent and independent variables is an important key concept.
- In the dataset, if you look closely, the first four columns (Item_Category, Gender, Age, Salary) determine the outcome of the fifth, or last, column (Purchased).
- Intuitively, it means that the decision to buy a product of a given category (Fitness item, Food product, kitchen goods) is determined by the Gender (Male, Female), Age, and the Salary of the individual.
- So, we can say that Purchased is the dependent variable, the value of which is determined by the other four variables.
- Using this we need to split our dataset into the matrix of independent variables and the vector or dependent variable.
- Mathematically, Vector is defined as a matrix that has just one column.

SPLITTING DATASET INTO INDEPENDENT FEATURE MATRIX:

```
1X = df.iloc[:, :-1].values
2print(X)
```

Pd

OUTPUT:

```
1[['Fitness' 'Male' 20 30000]
2['Fitness' 'Female' 50 70000]
```

```
3['Food' 'Male' 35 50000]

4['Kitchen' 'Male' 22 40000]

5['Kitchen' 'Female' 30 35000]]
```

Pd

EXTRACTING DATASET TO GET THE DEPENDENT VECTOR:

```
1Y = df.iloc[:, -1].values

2print(Y)
```

Pd

OUTPUT:

```
1['Yes', 'No', 'Yes', 'No', 'Yes']
```

Pd

CONCLUSION:

- There are many other sophisticated methods available in Python Pandas that can help the user to import data from different sources to its data frame.
- Once you have the data in the data frame, it can then be used for various kinds of analysis.
- We also saw how to segregate the data into dependent and independent variables.
- In the next guide, we will see how to carry on a few more pre-processing steps before data can be presented to the machine learning models.