

MLP OPEN ENDED ASSIGNMENT:

PROBLEM STATEMENT:

- **Build and evaluate the machine learning model for determining the possibility of purchasing a product based on customers' economic capabilities.**

INTRODUCTION:

Predicting customer behavior in the context of e-commerce is becoming more important nowadays. By utilizing clickstream and additional customer data, predictions can be carried out, ranging from customer classification, purchase prediction, and recommender systems to the detection of customer churn. A variety of machine learning models and data are available to conduct these kinds of predictions.

DATA COLLECTION AND PREPROCESSING:

Collecting data for training the ML model is the basic step in the machine learning pipeline.

Main source of data for building this model is **KAGGLE**.

Data processing techniques are a very crucial part of building model. Real-world raw data and images are often incomplete, inconsistent and lacking in certain behaviors or trends. They are also likely to contain many errors. One of the data cleaning techniques include removing empty cells that we have used while building the model.

The function reads as:

```
data.drop(data[data['Income'] >= 300000].index, inplace = True)
```

```
// will return a new dataframe and changes the data frame.
```

DATA VISUALISATION:Data visualization provides a good, organized pictorial representation of the data which makes it easier to understand, observe, analyze. We have plotted the data using a scatter plot.

FEATURE SELECTION:

It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. The most relevant feature to build a model if a person can buy a laptop will be his source of income.

MODEL SELECTION:

Input variables:

1. ID (numeric)
2. Year_Birth(numeric)
3. Education(Categorical: "Graduation", "PhD", "Master")
4. Marital : marital status (Categorical: "divorced", "married", "single")
5. Income(numeric)
6. Sex(Categorical: Male,Female)
7. Can_buy_Laptop(categorical: "Yes" , "No")

MODEL DESCRIPTION:

We have used Logistic Regression Algorithm. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

```
reg = LogisticRegression()
reg.fit(x_train, y_train)
Y_pred = reg.predict(x_test)
print(Y_pred)
```

- **Libraries used:**

1. PANDAS:

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. One of the most used method for getting a quick overview of the DataFrame.

2. SKLEARN:

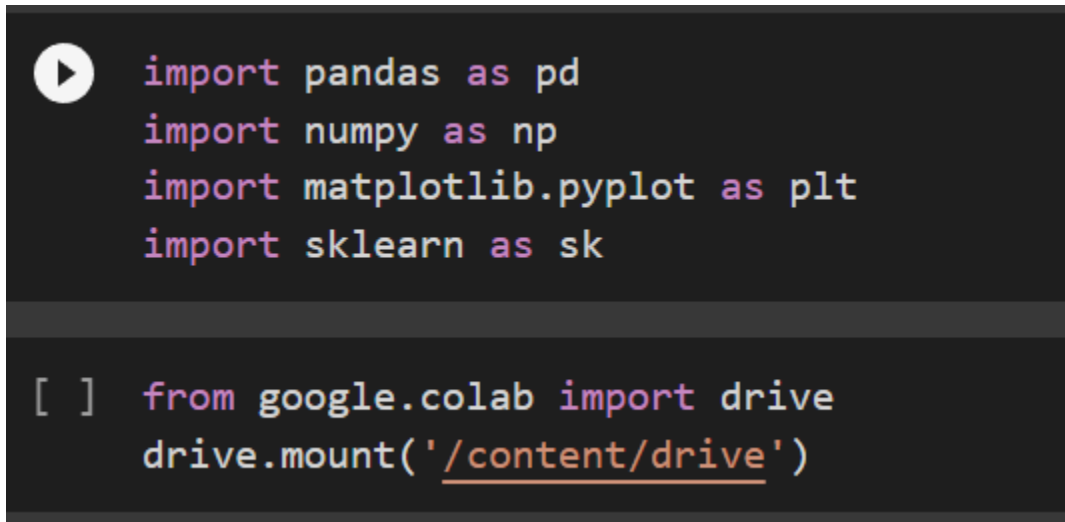
Scikit-learn is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation, and visualization algorithms using a unified interface.

3. MATPLOTLIB:

Matplotlib comes with a wide variety of plots. Plots help to understand trends, patterns, and to make correlations.

4. NUMPY:

NumPy, which stands for numerical python, is a library used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform, and matrices.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn as sk

[ ] from google.colab import drive
    drive.mount('/content/drive')
```

Testing

In machine and deep learning systems, data and desired behavior are the inputs and the models learn the logic as the outcome of the training and optimization processes. In this case, testing involves validating the consistency of the model's logic and our desired behavior.

They are two different classes of tests for Deep Learning systems:

- Pre-train tests
- Post-train tests

Pre-train tests: The intention is to write such tests which can be run without trained parameters so that we can catch implementation errors early on. This helps in avoiding the extra time and effort spent in a wasted training job.

We can test the following in the pre-train test:

- the model predicted output shape is proper or not.
- test dataset leakage i.e. checking whether the data in training and testing datasets have no duplication.

- temporal data leakage which involves checking whether the dependencies between training and test data do not lead to unrealistic situations in the time domain like training on a future data point and testing on a past data point.
- check for the output ranges. In the cases where we are predicting outputs in a certain range (for example when predicting probabilities), we need to ensure the final prediction is not outside the expected range of values.
- ensuring a gradient step training on a batch of data leads to a decrease in the loss.

```
[ ] from sklearn.model_selection import train_test_split

[ ] x = df[['Income']]
    y = df['Can_buy_Laptop']
    x_train,x_test,y_train,y_test=train_test_split(x,y,
    train_size=0.7,
    test_size=0.3,
    random_state=0)

[ ] print(y_test)
```

Post-train tests: Post-train tests are aimed at testing the model's behavior. It tests the learned logic and it could be tested on the following points and more:

- invariance tests which involve testing the model by tweaking only one feature in a data point and checking for consistency in model predictions. For example, in the case of titanic survivor probability prediction data, change in the passenger's name should not affect their chances of survival.
- Directional expectations wherein we test for a direct relation between feature values and predictions. For example, in the case of a loan prediction problem, having a higher credit score should definitely increase a person's eligibility for a loan.

CONCLUSION : Using logistic regression we can define persons ability to buy laptop using income as basic source

APPLICATION : This model can be used by companies to get targeted income consumer for specific product which can increase sell of specific product

GROUP MEMBERS:

- VAISHNAVI MORE 2205
- MANUSHREE MUNDRA 2208
- NIKITA PATIL 2211