**DataTrained (Padte Raho Badte Raho)**

**Name:- Vikas Rajput**

**Batch No.:- 1831**

**From :- Jammu and Kashmir.**

**Hello to all of you I never make a blog even I do not know how to make blog but I will try to make it, I am not very good in English but I will try.**

**Continue from next page.**

**Data Science**

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analysing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

As the given choice I can choose a **Flight Price Prediction Project.**

1. **Problem Statement**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

Size of training set: **10683** records

Size of test set: **2671** records

In this Dataset the features are(Airline, Date of journey,Source,Destination,Route,Dep. Time,Arrival Time,Duration,Total Stops,Aditional Info,Price.

Problem Statement by which we can understand that what to do with the dataset, to understand the dataset like a AIM that we have to achieve or a question that we have to answer,each and every dataset has a problem statement that defines the value of the dataset and a challenge to complete.

1. **Data Analysis**

Data Analysis is to analyse the data, to understand the data first, to see what type of data you have, categorical or continuous,numerical or alphabateical,float or integer,everything that u have to see ,to understand,to visualize and many much more things that u have to self understand and to work on it.

1. EDA(Exploratory Data Analysis)

EDA is the first step or you can also say that data analysis starts with this process,in eda that u have to start operations with your dataset,lets see the processes we can do in EDA.

* To check the Dataset classified or continuous
* To check the shape of data
* To check the information of the dataset
* To check their null values and replace them.
* To check the statistical view of dataset
* Their outliers and skewness,visualisation in dataset.
* To convert the string data into numeric.
* Making object columns in int. or float to be understandable for python
* Correlations with the features to each other.
* Divide the date columns into day month and year.
* Encoding processes
* Relationship of features
* Target variables problem sorting

There are many king of processes that we have to do in EDP this is the most important step by this we can get the best output if our all the can be sorted properly.

1. **Pre Processing Pipeline**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task. Generally a data contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

* **Dataset**

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the **dataset**.

Dataset may be of different formats for different purposes, such as, if we want to create a machine learning model for business purpose, then dataset will be different with the dataset required for a liver patient. So each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV **file**. However, sometimes, we may also need to use an HTML or xlsx file.

* **Importing Libraries**

In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:

**Numpy:** Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:

**( Import numpy as np)**

**Matplotlib:** The second library is **matplotlib**, which is a Python 2D plotting library, and with this library, we need to import a sub-library **pyplot**. This library is used to plot any type of charts in Python for the code. It will be imported as below:

**(import matplotlib.**pyplot as plt)

**Pandas:** The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library. It will be imported as below:

**(import Pandas as pd)**

**Building Machine Learning Models**

**Before I can explain about Building machine learning models,lets discuss about machine learning models.**

**There are three types of machine learning algorithms.**

**1.-Supervised Learning:-** This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, [Decision Tree](https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/), [Random Forest](https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/), KNN, Logistic Regression etc.

2.- Unsupervised Learning:- In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

3.-**Reinforcement Learning:-** Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

There is list of some common machine learning algorithms are:-

* Linear Regression
* Logistic Regression
* Decision Tree
* SVM
* KNN
* Random Forest
* Gradient Boost Algorithms

1.- GBM

2:-\_XGBoost

Linear regression:- It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a \*X + b.

Logistic Regression :- Don’t get confused by its name! It is a classification not a regression algorithm. It is used to estimate discrete values ( Binary values like 0/1, yes/no, true/false ) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a [logit function](https://en.wikipedia.org/wiki/Logistic_function" \t "_blank). Hence, it is also known as **logit regression**. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

Decision Tree :- This is one of my favorite algorithm and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

Building Machine learning model is a process in which we can divide our data into features and target variables,after performing this we can scale our data in x and y,after this process we can use train test tecniues now we can use the model for building a machine learning model,by using algorithms we can get the Scores .