Noor Mustafa
Rajpar.

⇒ ML : is the science of getting computers to learn and act like human do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.

⇒ How ML is Different from Traditional programing?

⇒ Traditional programming :- it is basically the manual process in this one human or one programer bilt one program. and that we feed to the computer and then ~~he~~ it generate the output.

ex :-

Data : ex ? " input

⇒ Output.

Programe : pre-written logic

⇒ Machine learning : it is basically automated process. so in this we give input as data and output and through the help of Data and output our machine return the program. or logic or model.

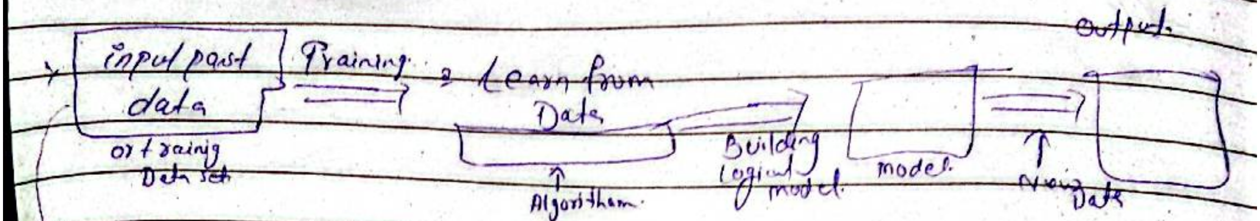
Data ~~input~~ input

⇒ program.

Output

so how machine learn by it self so the simple answer is Data. it is like a fuel

⇒ Flow it workrse.



This input past Data set is feed into the Algorithams. and what is Algorithams:? inside the Algorithams there are set of rules which tells How you this that have to be processed.

2nd phase:- in training phase the machine learning Algorithm. in which we feed the data into the Algorithams. that Algorithms learns from that Data's patterns and it does classifications. so in training Phase our Machine leaning Algorithm finds the patterns from that Data, and then it will classified it.

so it is known as training phase from, ~~for~~ machin learn from data

3rd Phase:- Testing Phase:- so in this we provid New Data sets to the Machines, Those data sets which they have never seen before, so what-ever Machine had lean in previous phase and find out patterns and classifications and other things, so it will use all learning in new Dataset. and then it will generate you output and provide you the predictions. so here the thing is if the more you have data the the more your model will be learn and then you will get more accurate answer.

⇒ So this is how Machine learning Model works.

⇒ classification of Machine Learning / Types.

1, Supervised Learning 2, Unsupervised Learning 3, Reinforcement Learnings

⇒ Data wrangling: is the process of cleaning, organizing and preparing data for use.

⇒ What is Dataset?

it is the collection of Records.

and what is Record?

A Record is like a single row in a table or dataset, Each row represents a single record, containing specific information or Data points

ex:-

ID	Name	Age
1	John	30
2	Jane	25

in this example, there are 2 records (rows)

Record1 : ID = 1, Name = John, Age = 30
Record2 : ID = 2, Name = Jane, Age = 25.

→ Note:- and that type of Data or Data sEts definitly will be stored in file. and that file can be different file format.

ex:- most commonly use formats.

1, CSV:- comma seprated value

2, JSON:- Javascript object Notations. in this type of format we store the ~~value~~ Data in the key value format. like tree like structure.

3, Excell format. and so many other format

⇒ Types of Datasets:

1. Numeric Data (Quantitative) :- exe Heigh weight price of house.
2. categorial data. (Qualitative) :- divide into the groups either yes or No
   T-F, M-F, Y-N
3. Ordinal data.
   ↳ it is the combination of both.

{ Note:- categorial Data mai agar ji Numerical Data bachi waandi ahy. ex T = 1 and F = 0

exe. if you have order something and then that App wants from so you rating then ex. 1 to 5
1 is like poor quality 5 is like excellent quality so in this case, both thing exist.

⇒ we can get or download data sets from:-
1. Kaggle Datasets
2. Amazon data sets
3. UCI Machine Learning Repository.
4. Google's Data search Engine.
5. Microsoft Data sets.

⇒ what is Data preprocessing in ML :-

:- Data preprocessing is a process of converting your raw Data into suitable form.

Means the data we get from the any source Most probably it is not clear format.
Inside that May be Noise, Missing value occur so by using Data preprocessing we set our data in suitable format.

⇒ Data preprocessing steps involves:-

1, Getting Dataset. 2, importing Libraries.

3, Importing Datasets 4, finding Missing values.

5, Encoding categorical Data.

6, spliting Dataset into Training and Test set.

7, feature scaling.

⇒ step1 create the new file and import your .csv file.

⇒ step2: Import Libraries.

→ Import numpy as np.

→ Import pandas as pd.

→ numpy Library is used for the scientific calculation

→ pandas Library is used for to manage the data set..

→ numpy helps you store and process numbers fast, especially when dealing with arrays or tables of Numbers, like in data science, machine Learning or eng.

⇒ Using panda we will display the files like.

step3 dataset = pd.read_csv('data.csv.')

↳ it is just the variable.

dataset # it will display that dataset.

step4: we have to find which is dependent and which is independent.

exe-

$x+y=z$

$x+y$ is dependent.

and z is dependent.

so: we have data set in which we have three independent variables.

country Age salary

and the dependent variable is purchased.

so what we have to do is the seprate out them.

→ so country, age and salary we are going to store into the X.

exe.

```
dataset = df.read_csv( 'data.csv')
→ X = dataset [['country','Age', 'salary'] ] # two dimeshonal array
```

X. → it will display: in tabular form, but if we need array then [[ ]].values

so this is for independent values

now similarly for the dependent variable.

```
Y = dataset [[ "Purchased"] ].values.
Y
```

→ so our next step of Data preprocessing is:- to find out the Missing data from the data.

like in Age NaN and salary NaN both have the Missing values.

so it is very Empotant to handle the Missing values.

There are two Methods:-

1 is to remove that particular rows or column from the data set.

but it is not much effiecient, it may arrize data loss issue.

Yeah we can use 1 method when in particular rows or column there Most be a 70 to 75% value Missing, then we can Delet or remove it. or Drop.

→ but in give scenario there is only one value is missing on particular rows and coulms.

→ so the second way is the to replace the particular value to find the medin or Mode or Mean then we will replace it.

so for this we have to import one Library.

→ from sklearn.impute import simpleImputer

Now call the simple Imputer() class.

// simple Imputer(): and the pass the values:

imputer = Simple Imputer (missing_values = np.nan, strategy = 'mean')

↳ it is just variable.

mean is the value. by defaut we can change it medin or mod.

Now

imputer.fit()

so

imputer = imputer.fit (x[:, 1:3]) # this will take all row and colums wherever the missing values is occur.

this means it will takes the all rows

Now, other method.

x[:, 1:3] = imputer.transform (:, 1:3)

↳ all rows 1 to 3.

↳ what it will do what ever the missing value is it will replace them with the means

→ X # it will now display and fill NaN will mean value.

=> Next step is categorial data:
it is known as categorial encoding data
categrial data hamari wo data hoti hai jis mai labels hoti hai
like country and purchased these are labels not
like France No the numeric values.
spanich Yes

-> so catogerial encoding data kiya hota hai yai jo hamary datasets mai country and purchased hai us ko hum numeric mai convert karygai
or hum is liye karty hai because hamary jo most of the algorithms hai woo categrial data par kam nh karty. bas
bas khuh hee algoritham hai like decisiontree jo is type ko easily handle kar sakta hai

-> so we have country and purchased variable. so for country
-> step1:- import module, named label encoder jo hamary preprocessing library ka part hai

so:

```
from sklearn.preprocessing import labelEncoder
   Label_encoder_x = LabelEncoder()
n[:,0]=label_encoder_n.fit_transform(n[:,0]) of n.
```

-> all rows
-> country exisist on 0

```
n // it will show in nmeric
```

⇒ Now Dummy encoding:

like

country	France	Germany	spain.
France	1	0	0
Spain	0	0	1
Germany	0	1	0
Spain	0	0	1
Germany	0	1	0

→ so domy encoding mai just 0 and 1 values hoti hai.

so for that:

```
from sklearn.preprocessing import OneHotEncoder.
 onehotencoder = OneHotEncoder()
x= onehotencoder.fit-transfor(dataset.Country, values.reshape(-1,1)).toarray()
   x // it will display
 // Now same for y.
labelencoder_y = LabelEncoder()
y= labelencoder_y.fit-transform(y.)
 y.
```

→ Now train and test.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y test = train_test_split(x,y, test_size=0.2,
                                                     random_state=0)
 x_train.
 x_test.
 y_train
 y_test.
```

⇒ Now last is feature scaling: ex in our dataset if there is a data who have huge magnitude and Lenghth then we have to made it to the same scale.