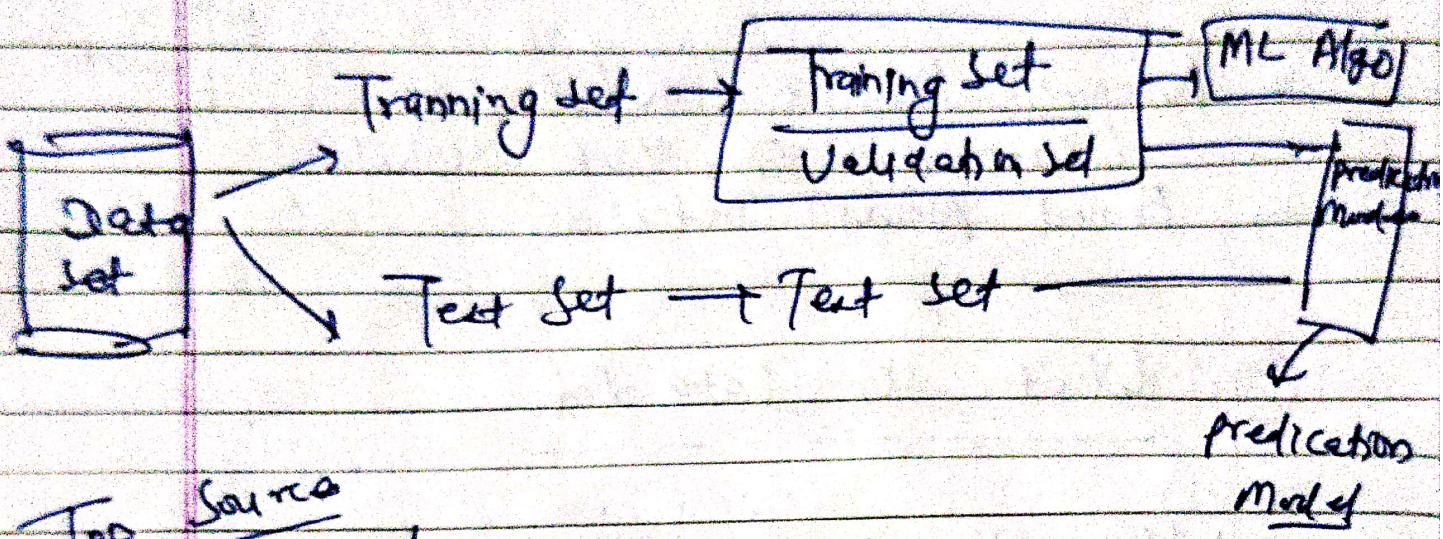
(11)

Need of Data set :



Predication Model

Top Source

→ Kaggle Data set

→ Amazon Data set

→ UCI ML Repository → Classified Already Clean data

→ Google data search Engine

Most data set Clean

→ Microsoft Data set

→ Awesome public Dataset Collection (Git repo)

→ Scikit-learn Dataset

→ Toy (packages)

→ Real world data set

→ Computer Vision Dataset (Upload with)

→ Kaggle dataset

data.gov.in → API

data.gov

12

7 Data Preprocessing in Machine learning

It is a process Raw data को suitable form में बनाता है।

→ Getting dataset → Import libraries → import dataset → finding Missing Values → Encoding Categorical data → Training & Test Data set → Feature Scaling

→ Dependent or In Dependent Variable

x + y = z

Google Colab file → Data Preprocessing .ipynb

All Step we are doing step by step

Features and Lable in ML

Column → Features

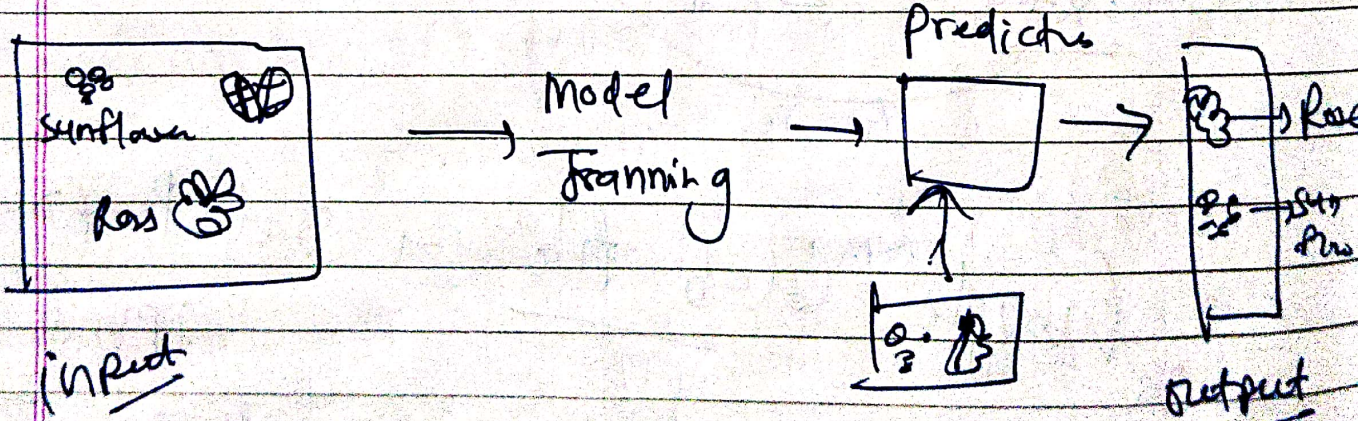
Tranning fase Comple होने के बाद

label = output

Super wise ML

Advantage & Disadvantage

→ Input Variable को match करता है के साथ में



2 types

Regression

जो output होगा Numerical होगा

Ex 1, 0, 22, 23

Unique Values

→ इसमा यह हम पहा करते है जहा depended और independed मे कोई relation होता है

→ Continous Values

Ex Weather forcast
market Trends

Most popular Algorithm.

① → Linear Regression
② → Polynomial Regression
③ → Regression Tree

Classification

जो output होगा Categoiel होगा

True false Yes, No
Male femal
Day night,

Most popular Algo.

Random Forest
Decision Trees
Logistic Regression
Support Vector Machine

Advantage of Supervised Learning

① full control over ML
② easily test and debug model
③ we can determine the number of classes

(15)

Dis Advantage of SL :

→ Have limited Scope
→ label data dat is expensive and time-Consuming
→ Wrong prediction

→ इसमे एक supervisor होता है उसके निर्देश मे होता है सब कुछ

→ supervise ML मे जो जिस data से Train करते है वो label data होता है

→ 2 types
- Regression (Numerical)
- Classfical (Categorical)

Ex:- Spam filter
↳ label को target करता है जैसे win lotery win, Jackpod mail मे मिला तो Spam मे डाल देगा

→ यदि हम Tranning data से अलग कुछ देते है तो वो गलत result देगा।

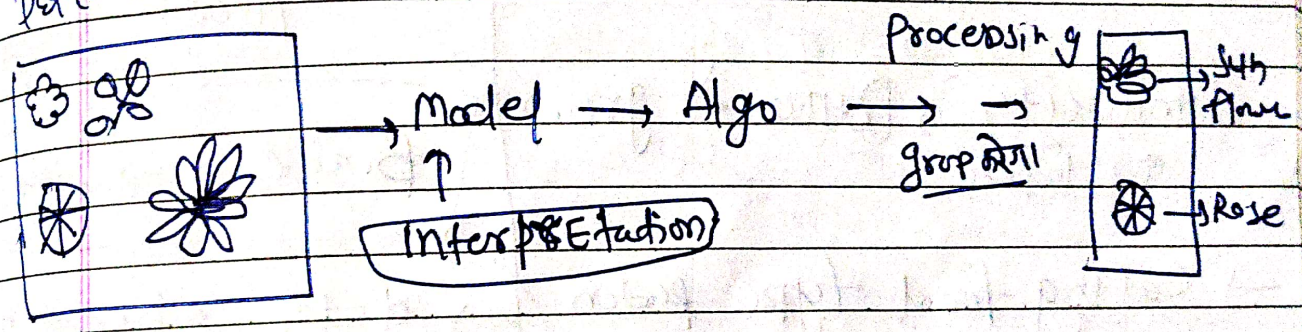
Unsupervised Machine learning

Unlabeled data

model खुद find करेगा hidden Patterns and insights से जो हम data दे रहे है

Data को sort करेगा similarities देखेगा।

कोई previous knowledge नहीं होता है



Unlabeed data

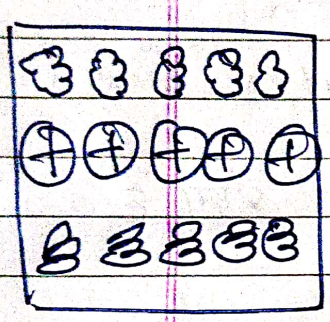
→ तब हमारा model Pattern search करेगा और category के Base पर बताएगा की flower कौनसा है

The goal of Unsupervised Learning is to group Unlabelled data acoading to the similarities, patterns and difference without any prior training of data.

Types of Unsupervised ML

Clustering

Raw Data → Algo → output



Similarity के हिसाब से group बना दिया

→ eating food type person

→ Onion, Potato सब mix है उनको अलग² कर देता है

Association

Bread + Butter + egg =
egg + Br + By =
Chocolate + Br + By =
Chips + Br + By

Most frequent

Bread + Butter

→ data object में relationship देखेगा

Ex → Grocery example

Most people

Home जरिदते है तो साथ खरीदते ही खरीदते है इससे marketing effective हो जाती है।

Bread के पास है कोई तो Butter भी होगा वो

Unsupervised Leaning Algorithms

- K-Means Clustering
- KNN (K-Nearest Neighbour)
- Hierarchical Clustering
- Neural Networks / Deep learning
- Single Value Decomposition
- Distribution Models
- Principal Component Analysis
- Apriori Algorithms

Advantages ⟹

(I) Most Complex task के लिए use होता है।

(II) It is help in finding pattern in data

(III) Saves lot of manual work and expense

यह easily मिल जाता है।

Dis Advantage → Less Accuracy
→ Time Consuming
→ More the features More the Complexity

→ Unlabeled data को Train करते है

→ Clustering & Association 2 types

→ result is less Accuracy

→ Clustering क्या होता

इसमें कोई supervisor नहीं होता है तो सारा काम Machine को ही करना होता है

relation ship देखना
Bread + Butter
Shoes के साथ दो Shoes भी तो लेगा ही लेगा

(11) * Training & Testing Data in ML

Training Data	Testing Data
80%	20%

training and testing.ipynb → git repo

(12) Linear Regression Single Variable

It is part of Supervise ML

→ बाकी famous and easy

→ इसमे हम जो Value predict करते है वो Continous होनी चाहिए or numerical Value होनी चाहिए

* Cost age, Salary temp. price

2 types

(I) Linear Regression with Single Variable

(II) Linear Regression with multipal Variable