Assignment No. 01

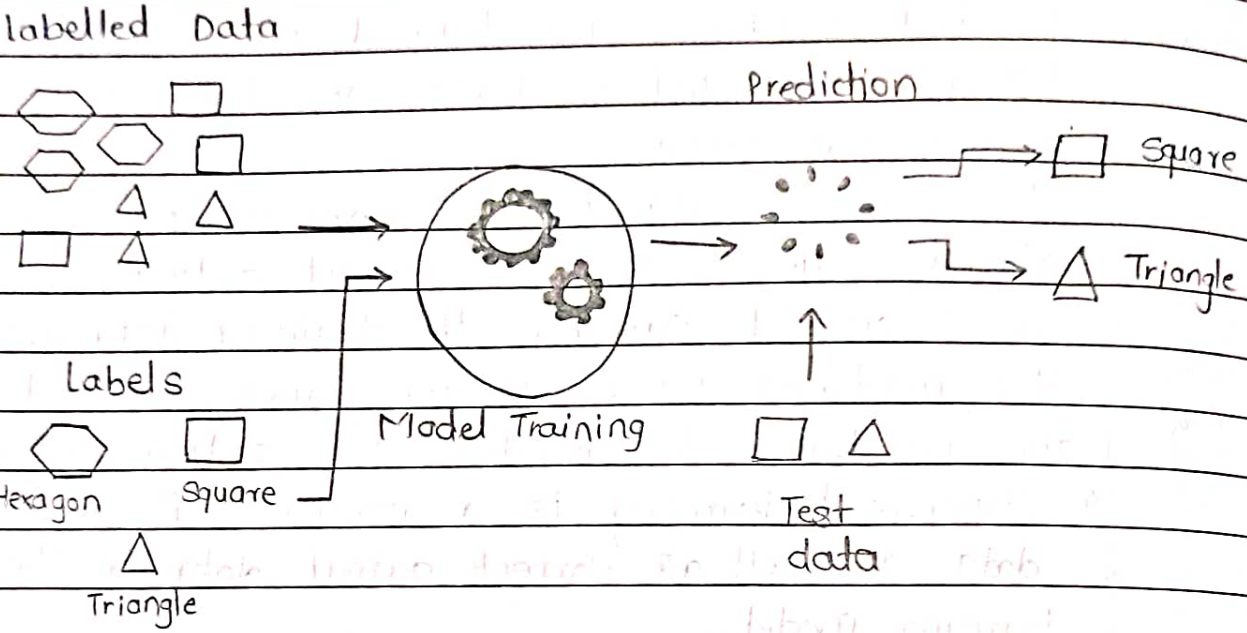
1. Explain Supervised learning.

1 Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.
2. The labelled data means some input data is already tagged with the correct output
3 In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly.
4 Supervised learning is a process of providing input data as well as correct output data to the machine learning model.
5. The aim of a supervised learning algorithm is to find a mapping function to map the input variable (x) with the output variable (y)
6 In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud detection, spam filtering etc

Working of Supervised learning :

1 In this, models are trained using labelled dataset, where the model learns about each type of data.
2 Once the traing process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output

3. The working of supervised learning can be easily understood by the below example and diagrams:



Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

i] If the given shape has four sides, and all the sides ~~and~~ all are equal, then it will be labelled as a Square

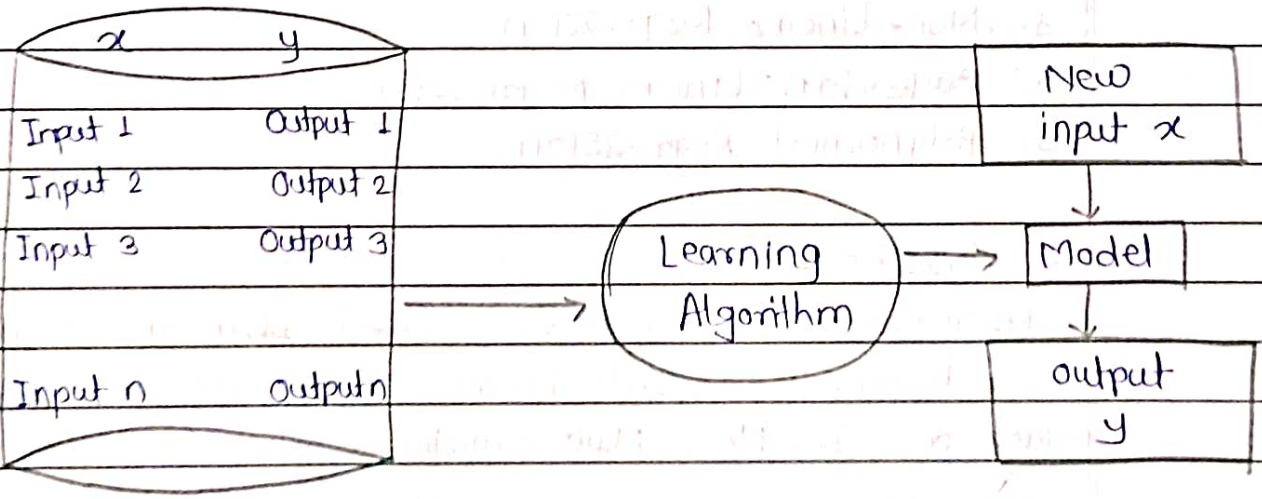
ii] If the given shape has three sides, then it will be labelled as a triangle

iii] If the given shape has six equal sides then it will be labelled as hexagon.

Now, after training, we test our model using the test set, & the task of the model is to identify the shape. The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides and predicts the output

Steps involved in Supervised learning :

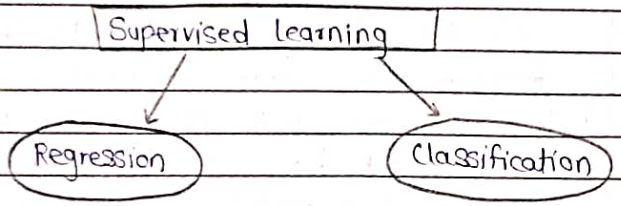
1. First Determine the type of training dataset
2. Collect / Gather the labelled training data.
3. Split the training dataset into training dataset, test dataset and validation dataset.
4. Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
5. Determine the suitable algorithm for the model, such as vector machine, decision tree, etc.
6. Execute the algorithm on the traing dataset Sometimes we need validation sets as the control parameters, which are the subset of traning datasets.
7. Evaluteate the accuracy of the model by providing the test set If the model predicts the correct output, which means our model is accurate.



Trained dataset

Types of Supervised Machine learning Algorithms :

Supervised learning can be further divided into two types of problems :



1 Regression :
Regression algorithms are used if there is a relationship between the input variable and the output variable
It is used for the prediction of continous variables, such as Weather forecasting, Market Trends etc.
Below are some popular Regression algorithms which come under supervised learning :
1. Linear Regression
2. Regression Trees
3. Non-Linear Regression
4. Bayesian Linear Regression
5. Polynomial Regression

2. Classification
Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-female, True-false etc

Spam filtering
1. Random forest
2. Decision Trees
3. Logistic Regression
4. Support vector Machines

Advantages of Supervised learning :
1. With the help of supervised learning the model can predict the output on the basis of prior experiences
2. In supervised learning, we can have an exact idea about the classes of objects
3. Supervised learning model helps us to solve various real-world problems such as fraud detection, spam filtering etc.

Disadvantages of supervised learning :
1. Supervised learning models are not suitable for handling the complex tasks
2. Supervised learning cannot predict the correct output if the test data is different from the training dataset
3. Training required lots of computation times
4. In supervised learning, we need enough knowledge about the classes of object

2. Explain unsupervised learning.

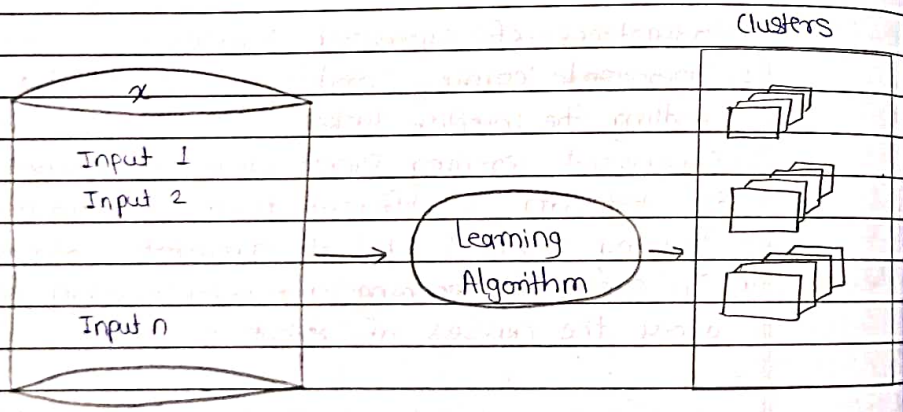
Unsupervised learning is a machine learning technique in which models are no supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data
It can be compared to learning which takes place in the

brain while learning new things. It can be defined as

3] Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

4] It cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data.

5] The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities and represent that dataset in a compressed format.



Example :

Suppose the unsupervised learning algorithm is given as input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset.

The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities bet^n images.

The need of unsupervised algorithm -

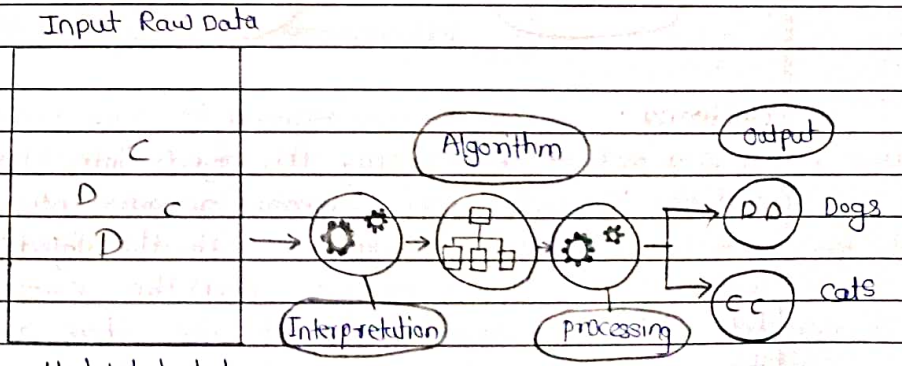
1] It is helpful for finding useful insights from the data.

2] It is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.

3] Unsupervised learning works on unlabeled & uncategorized data which make unsupervised learning more imp.

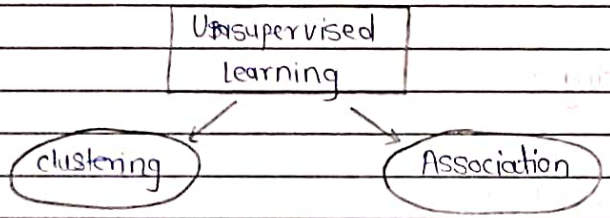
4] In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Working :



1] Here we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given
2] Now, this unlabeled input data is fed to the machine learning model in order to train it
3] Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc
4] Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference bet^2 the objects.

Types of unsupervised learning algorithm:



Clustering:
It is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities bet^2 the data objects and categorize them as per the presence and absense of those commonalities.

Association:
An association rule is an unsupervised learning method which is used for finding the relationships bet^n variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy x item (bread) are also tend to purchase Y (Butter / Jam) item. A typical example of association rule is Market Basket Analysis.

Unsupervised Learning Algorithms:
1] K-means clustering
2] KNN (K-nearest neighbors)
3] Hierarchal clustering
4] Anomaly detection
5] Neural Networks
6] Principle Component Analysis
7] Independent Component Analysis
8] Apriori algorithm
9] Singular value decomposition

Advantages of unsupervised learning -
1. Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.

2. Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of unsupervised learning -

1. It is intrinsically more difficult than supervised learning as it does not have corresponding output

2. The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, & algorithms do not know the exact output in advance

9. Explain Reinforcement Learning :

i] It works on a feedback based process in which an AI agent (software component) automatically explore it's surrounding by hitting & trail, taking action, learning from experiences & improving it's performance.

ii] Agent gets rewarded for each good action & get punished for each bad action hence goal of reinforcement learning agent to maximize the rewards

iii] In that there is no labelled data like supervised learning and agents learn from their experience only

iv] It is similar to human being

v] e.g. - child learns various things by experiences in his day-to-day life.
Play a game, where the game is the env. move of an agent at each step define states & goal of the agent is to get high score. Agent receives feedback in terms of punishment & rewards

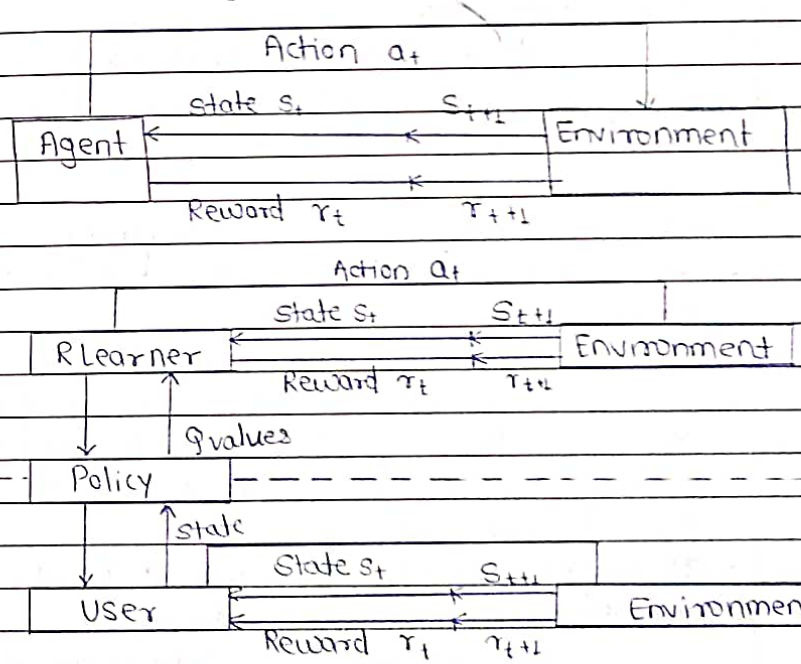
vi] A reinforcement learning problem can be formalized using Markov Decision Process (MDP) In MDP agent constantly interacts with the end & performs at each action, the env. respons & generates a new state

vii] It is categorized mainly into two types of algo.

1] Positive Reinforcement Learning -
It specifies increasing the tendency that the requirement behaviour would occurs again by adding something

2] Negative RL
It works exactly opposite to the positive RL
It increase the tendency that the specific behaviour would occur again by avoiding negative conditions

viii] Uses - video games, Resource management, Robotics, Test mining.



Advantages of Reinforcement Learning -

1. Reinforcement learning can be used to solve very complex problems that cannot be solved by conventional techniques
2. The model can correct the errors that occured during the traning process
3. In RL, training data is obtained via the direct interaction of the agent with the environment
4. It can handle enviroments that are non-deterministic, meaning that the outcomes of actions are not always predictable. This is useful in real-world applications where the enviroment may change over time or is uncertain
5. It can be used to solve a wide range of problems including those that involve decision making, control and optimization
6. It is a flexible approach that can be combined with other ML techniques, such as deep learning to improve performance

Disadvantages of Reinforcement learning -

1. It is not preferable to use for solving simple problems
2. It needs a lot of data and a lot of computation
3. It is highly dependent on the quality of the reward function. If the reward function is poorly designed, the agent may not learn the desired behaviour.
4. It can be difficult to debug and interupt. It is not always clear why the agent is behaving in a certain way, which can make it difficult to diagnose and fix problems.

4. What is machine learning ? Explain types of ML.

A branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been fed into it.

ML is a scientific discipline that is concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data, such as from sensor data or databases.

Types of ML -

1. Supervised learning
2. Un-supervised learning
3. Semi-Supervised learning
4. Reinforcement learning

Reinforc- Semi-supervised learning -

1. It is a type of machine learning that falls in between supervised and unsupervised learning.
2. It is a method that uses a small amount of labelled data to train a model.
3. The goal of Semi-Supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning.
4. However, unlike supervised learning the algorithm is trained on a dataset that contains both labeled & unlabeled data.

Semi-supervised learning is particularly useful when there is a large amount of unlabeled data available, but it's too expensive or difficult to labe all of it.

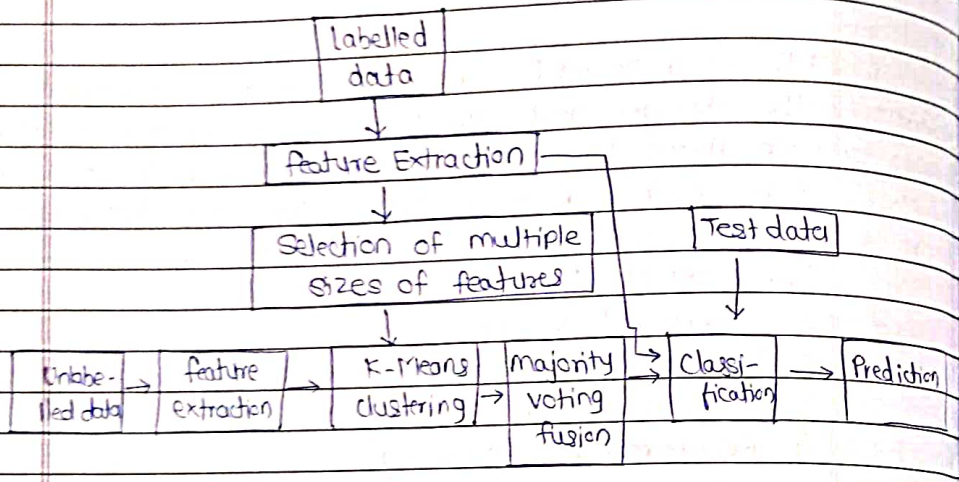
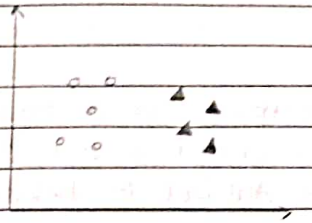
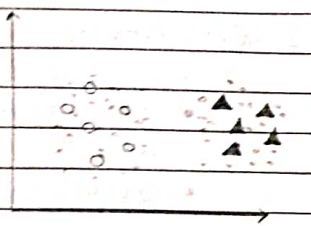


fig. Semi-supervised learning flow chart

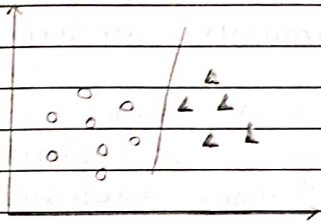
Examples -
1. Text classification
2. Image classification
3. Anomaly detection.



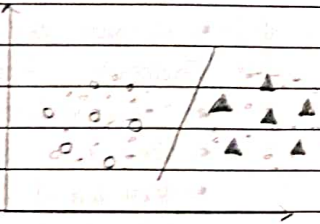
a] Labelled data



b] Labelled + unlabelled data



c] Supervised data



d] Semi supervised data

Real-world applications of Semi-supervised learning-
1] Speech Analysis
2] Web content classification
3] Protein sequence classification
4] Text document classifier

5. Explain ML problem categories -

ML algorithms are also classified under these learning problems -
1. Classification
2. Clustering
3. Regression
4. Optimization
5. Simulation

1. Classification -
   1. Classification definition -
      - Identification of groupings in a dataset based on the value of a target or output attribute
      - Qualifies the entire dataset to belong to specific classes
   2. Purpose of classification :
      - Aids in recognizing patterns in data behaviour
      - Serves as a discrimination mechanism.
   3. Example Scenario :
      - Illustration involving a sales manger
      - Objective : Identify prospective customers
      - Key input for the manger : Total lifetime value (TLV) of the customer.
   4. Application in Business :
      - Helps the sales manager decide if its worth investing time & effort in a potential customer
      - TLV is a common metric used in this context

2. Clustering -
   1. Data analyst Task :
      - Often given data with the expectation to discover interesting patterns for deriving intelligence.
   2. Difference from classification :
      - Classification : Business user specifies what to look for (e.g., good or bad customer)
      - Analyst task : Unearth patterns without a predefined target
   3. Example Expansion - customer classification :
      - clustering, as a form of classification analysis, identifies patterns without specific predefined targets
   4. In clustering, the example of classifying customers doesn't have a predetermined target, the results may vary based on factors like initial centroid selection.
   5. K-means clustering is an example modeling method used for unsupervised analysis in clustering.
   6. Clustering does not start with a specific target in mind, making it distinct from classification based on criteria like "good / bad" or "will buy / will not buy".

6. Explain supervised learning problem categories.

3. Forecasting, prediction or regression
   1. Similar to classification, forecasting or prediction involves anticipating future events based on past experience or knowledge.
   2. Forecasting may involve regression when there is insufficient data to define the future.
   3. Results of forecasting & prediction are presented with a degree of uncertainty or probability.
   4. The problem type of forecasting is sometimes referred to as rule extraction.
   5. An example illustrate an agricultural scientist predicting crop yield based on attitude & relate data points

6. Regression is a technique used in forecasting where a relationship between variables is determined by plotting a graph & finding an eq^n that fits the data points
7. In regression, data points that do not fit the curve may be discarded.

4. Simulation
1. Uncertainty in data context
- Situations may arise where the data itself carries inherent uncertainty
- Example : An outsourcing manager estimating task completion time based on team skills and experience

2. Variable Input Parameters :
- The context involves variables with varied possibilities
- Example : Cost of input material ranging from $100 to $120 and the no. of employees bet^n 6 & 9.

3. Experience-Based Estimation
- Professionals, like an analyst, rely on experience to estimate project duration
- Example : An outsourcing manager estimating a task to be done by a team with specific skills in 2-4 hours

4. Simulation Requirement :
- Addressing such uncertainties often necessitates simulating numerous alternatives
- Ex : Simulating various scenarios to account for the range of possible input parameters

5. Challenges in Machine Learning :
- In tasks like forecasting, classification and unsupervised learning, the data interconnections may be unknown
- Ex : lack of eq^n describing one variable as a function of others in these types of learning scenarios

Essentially, data scientists combine one or more of the preceding techniques to solve challenging problems, which are
- Web search & information extraction
- Drug design
- Predicting capital market behaviour
- Understanding customer behaviour
- Designing robots

5. Optimization :
1. Definition of optimization
- Optimization is a mechanism to enhance something or establish a context for a sol^n, making it the best

2. Production Scenario Example :
- Two machines in a production scenario
- Machine 1 requires more energy for high-speed production & fewer raw materials
- Machine 2 requires higher raw materials & less energy for the same output in the same time

3. Understanding Output Patterns :
- It's crucial to comprehend output patterns based on input variations
- Identifying a combination yielding the highest profits is essential

4. Analytical Task :

- As an analyst, the goal is to identify the best way to distribute production bet^2 the machines
- The aim is to maximize profits for the production manager.

6. Explain supervised learning problem categories.

There are two categories

1] Regression
2] Classification

1] Regression

i] Regression algorithms are used if there is a relationship between the input variable and the output variable.

ii] It is used for the prediction of continuous variables, such as weather forecasting, Market Trends etc

iii] Below are some popular Regression algorithms which come under supervised learning :

1] Linear Regression
2] Regression Trees
3] Non-linear Regression
4] Bayesian Linear Regression
5] Polynomial Regression.

7. Explain unsupervised learning problem categories.

There are two categories

i] Clustering
ii] Association

i] Association

1] An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database

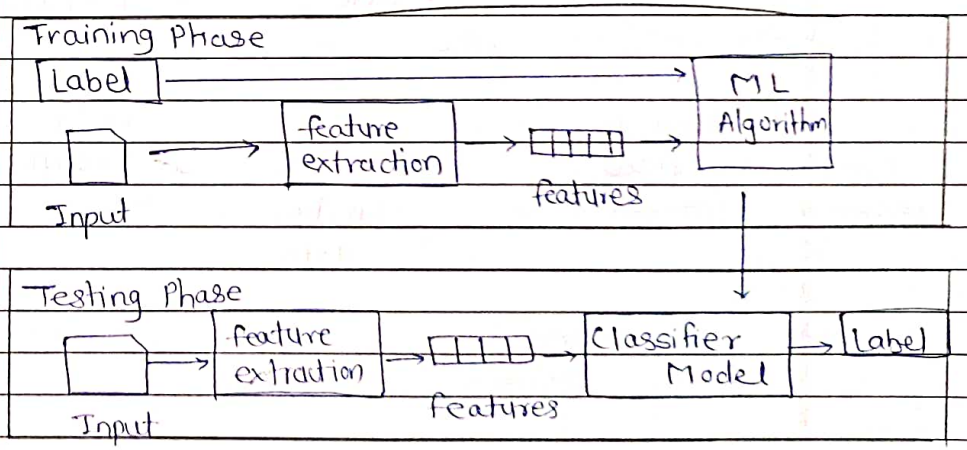
2] It determines the set of items that occurs together in the dataset

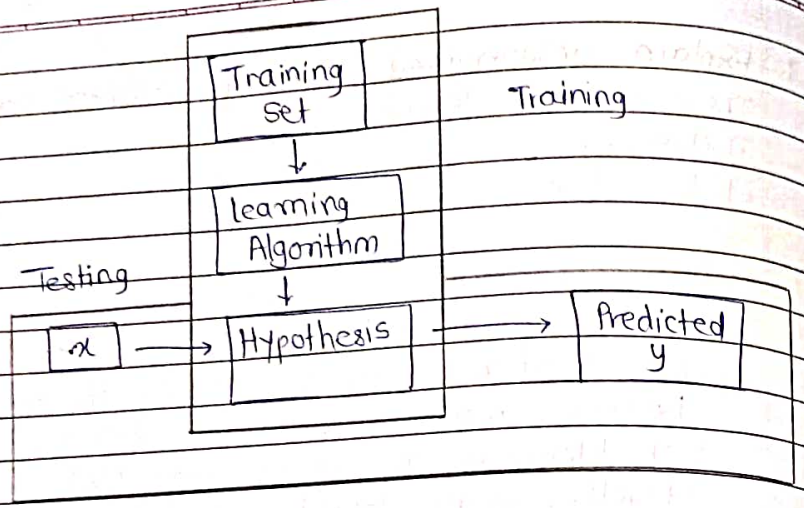
3] Association rule makes marketing strategy more effective.

4] Such as people who buy x item (suppose a bread) are also tend to purchase Y (Butter / Jam) item.

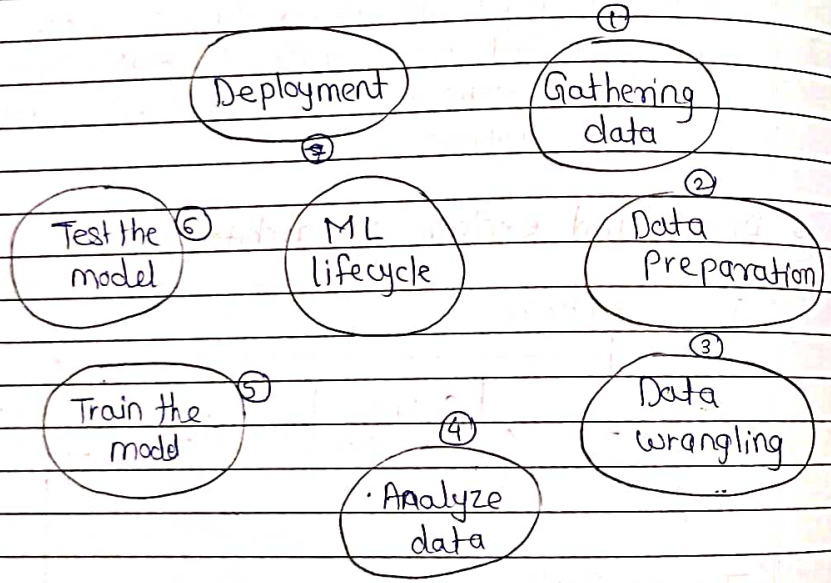
5] A typical example of association rule is market basket Analysis

8. Draw and explain ML architecture





9. Draw & explain ML lifecycle.



Basically ML life cycle contains two phases
1. Training Phase
2. Testing Phase

In the training phase it contains 5 subtypes
1. Gathering data
2. Data preparation
3. Data wrangling
4. Aanlyze data
5. Train the model

In the testing phase it contains 2 subtypes
1. Test the model
2. Deployment

1. Gathering data
   i] In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet or mobile devices.
   ii] It is one of the most important steps of the life cycle.
   iii] The quantity and quality of the collected data will determine the efficiency of the output
   iv] The more will be the data, the more accurate will be the prediction

This step includes the below tasks:
1. Identify various data sources
2. Collect data
3. Integrate the data obtained from different sources

2. Data preparation -

In this step, first, we put all data together & then randomize the ordering of data
This step can be further divided into two processes:

i] Data exploration :
1] It is used to understand the nature of data that we have to work with. we need to understand the characteristics, format & quality of data
2] A better understanding of data leads to an effective outcome. In this, we find correlations, general trends, & outliers

ii] Data pre-processing
1] Now the next step is preprocessing of data for its analysis

3. Data Wrangling :
i] Data wrangling is the process of cleaning & converting raw data into a useable format
ii] It is the process of cleaning the data, selecting the variable to use & transforming the data in a proper format to make it more suitable for analysis in the next step
iii] It is one of the most imp steps of the complete process. Cleaning of data is required to address the quality issues
iv] It is not necessary that data we have collected is always of our use as some of the data may not be useful

In real-world applications, collected data may have various issues, including :
- Missing values
- Duplicate data
- Invalid data

4. Analyse Data :
i] Now the cleaned & prepared data is passed on to the analysis step. This step involves :
a] Selection of analytical techniques
b] Building models
c] Review the result
ii] The aim of this step is to build a machine learning model to analyse the data using various analytical techniques & review the outcomes
iii] It starts with the determination of the type of the problems, where we select the ML techniques such as classification, regression, cluster analysis, Association etc., then build the model using prepared data & evaluate the model

5. Train the model :
i] Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
ii] We use datasets to train the model using various ML algorithms

6. Test the model:
Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it

7. Deployment:
The last step of ML life cycle is deployment where we deploy the model in the real-world system

10. Explain performance measure for ML.

On the basis of actual & prediction values we are finding different errors i.e. type 1 & type 2 error & using type 1 & type 2 error we are finding

1. Accuracy
2. Sensitivity
3. Specificity
4. Negative prediction value
5. Positive prediction value
6. False negative value / rate
7. false positive rate
8. false discovery rate
9. false aumission rate
10. F1 score

Actual values

Predicted values	positive	negative
positive	TP	FP Type-1 error
negative	FN Type-2 error	TN

dig. Confusion Matrix

TP = True positive
FP = false positive
FN = false negative
TN = True negative

True positive (TP) - It is the case when both actual class & predicted class of data point is 1

True negative (TN) - It is the case when both actual class & predicted class of data point is 0

false positive (FP) - It is the case when actual class of data point is 0 & predicted class of data point is 1

false negative (FN): It is the case when acutal class of data point is 1 & predicted class of data point is 0.

Ex

TP	FP
(30)	(30)
FN	TN
(10)	(930)
40	960

Accuracy : $\frac{TP+TN}{P+N} = \frac{30+930}{1000} = 0.96$

Precision : $\frac{TP}{TP+FP} = \frac{30}{30+30} = 0.5$

Recall : $\frac{TP}{P} = \frac{30}{40} = 0.75$

fallout : $\frac{FP}{N} = \frac{30}{960} = 0.031$

Sensitivity : $\frac{TP}{TP+FN} = \frac{30}{30+10} = 0.75$

Specificity : $\frac{TN}{FP+TN} = \frac{930}{30+930} = 0.968$