INTRO TO AI/ML-2

1. **Unsupervised Learning -** No supervisor or teacher. For.eg - A son learning to ride a bicycle from his father is supervised learning, whereas he learning to ride it by looking at others without any assistance is unsupervised learning.

   A. Clustering -

- E.g - Consider an employer, looking and analyzing at the log records of all the employees during office hours. He records the Employee ID and the hours of work done by the employee.
- Here, the **machine will just plot** the Employee ID on the x-axis & hours of work on the y-axis.
- The **employer/owner will manually analyze the graph** and find out who is spending how much time where.
- Say, the employer notices that the employees can be divided into 4 clusters based on the graph. He/She analyzes the history of some points from each cluster and finds out that the following info -

Cluster	Analyzed Data
A	Spends only 3 hr on office work and spends rest time on FB.
B	Spends only 4 hr on office work and the rest of the time on Insta.
C	Spends 6 hr on work and the rest of the time searching on Linkedin or Naukri.com
D	Dedicates all 7 hr on work.

- Analyzing this data will help the employer take appropriate measures on appropriate people. Like he/she can organize motivational sessions for employees in cluster A & B to make them work harder. Also, he can notice that the employees in Cluster C are unhappy with their current job, so he/she can look into the matter and help them out.

B. Dimensionality reduction-

- It's all about identifying the useful features and discarding the others!
- E.g - You want to analyze whether a student, should be given a grant for a project. The data you have is - ~~Name, Phone no., Email id~~, grades, Scores. Here, the output is only dependent on grades and scores, whereas the other features are of no use to determine the output.
- This problem of having many features which are of no use to obtain the o/p --> **"Curse Of Dimensionality".**
- Several algorithms (such as PCA) are used to resolve this problem(in an automized way).

C. Association Rule-Mining -

- Apriori Algorithm - Examine the dataset and find patterns and rules.
- E.g- Amazon giving **generic recommendation**s such as - whenever you buy a mobile, it recommends you to buy a back cover or a screen guard because it has observed the pattern that the ones who buy a mobile phone also usually buy these things.
- !!! Remember that the **personal recommendations** given by amazon based on your personal buying interests **do not come** under this category.

2. Reinforcement Learning -

- Learning from own observations.
- It has an agent who works on the basis of reward(when beneficial) and penalty(when detrimental) policy.
- e.g - Google's Autonomous Car - Here, Agent---> Car . Say it is learning to drive on a speed breaker. It has a particular score at the beginning.

  First Trial - It has no idea, it drives at 70km/hr speed. So, it realizes its detrimental and penalizes itself with -ve 10 score.

  Second Trial - It now decreases its speed to 40,

although it has a better experience than the previous time but still it is not optimal. So, its penalized with -ve score.

- These trials continue until the car learns the best optimal speed to cross the speed breaker and reward itself with a +ve score.

3. Classification based on Incremental Data Samples-

A. Batch Learning

- The model is trained in with a particular data set in one go and then it stops learning.
- e.g - In a bread manufacturing factory, once the model of conveyors is trained on how to bake it, slice it, pack it and be ready, the training stops.
- So, if the machine is given another size of bread(other than the usual ones used for training), it will fail to work on it.

B. Online Learning

- Training data is fed in multiple batches.
- e.g - Stock market forecasting - where we continuously update the model with new training sets.
- !!!! This model may have a drawback- Suppose we take the previous e.g in which we had to analyze whether a student, should be given a grant for a project based on the data - Name, grades & Scores. Now suppose we have 4 Vaishnavi's in the class out of which 3 of them have got good grades and scores and are eligible for the grant. But, that does not mean that all the Vaishnavi's should be given the grant! The machine doesn't understand that the name feature isn't required for finding the output. So, training by **bad data samples** can lead to problems!

4. CRISP-DM Process Model -

1. Business Understanding -

- The data scientist goes to the client and understands how the business problem can be converted to a data science problem.

2. Data understanding -

- Collect the data, understand it, analyze its quality.
- Very imp as bad data can lead to adverse problems.

3. Data preparation -
    - Clean and curate the data before actual ML implementation.
    - Takes most of the time(70-80%)
4. Modeling-
    - Build and implement the ML model(Linear Regression or classification etc)
5. Evaluation -
    - Check if the algorithm is working fine. If not, it iterates the process back to build a new model.
6. Deployment -
    - When its working fine, its deployed to the clients.

5.ML pipelines -

1)Data asset - Data available in .csv or .db,.XML,son etc files.

2) Data Retrieval - extracting data from these files.

3) Data Preparation -

- Pre-processing data, f**eature extraction**(finding out the required features)
- **Feature Engineering** - e.g You have to find whether a student's performance is good or bad based on his performance in P, C, &M. So, you "engineer" a new feature called Total, which was initially not present in the data, for convenience of analyzing.
- **Feature scaling** - It involves normalizing the data by proportionally reducing it so that all the values lie between 0-1. Then, you can rank the data and find out which is good and bad.

4) Modeling - deciding ML Algorithm.

5) Evaluation - if it is not fine, reiterate to Data preparation

6) Deployment to client

6. Supervised Learning Pipelines -

Considering the same example - whether a student, should be given a grant for a project based on the data - Name, Email-id, grades & Scores.

TRAINING-

- **The** Training Data - Name, Email-id, grades, Scores,o/p(Yes or No) is fed through all data preparation processes.
- The required features -grades, Scores - are the training features and the o/p labels are the outcome labels. These form the Supervised ML model.

PREDICTION -

- The new data - Name, Email-id, grades & Scores.- has to go through all the data preparation processes. If we directly give it to the model it may give garbage value.
- Then these new features - grades & Scores.- produce the predicted output.

7.Unsupervised learning Pipelines -

- It is almost the same as the supervised one, but here we do not have the o/p(Yes or No) labels for the training data.
- Here, once the Unsupervised ML model will plot the graph for the training data. the, it will plot the new feature into that graph.
- Then, we need to analyze the graph with clustering, patterns and association rules.

8.

- **Underfitted Model** - Model in which the LOBF passes through fewer data points.
- **Overfitted mode**l - Model in which LBOF passes through all points (including noise), just like a mugging up data, which is of no use.

- **Good Fit/Robust** - Model which neither underfit nor over-fit and is descent.