Page :
Date: / /

⇒ Machine Learning

It is an application of AI that enables system to learn and improve from experience without being explicitly programmed. It focuses on developing computer programs that can access data and use it to learn for themselves.

Types :-

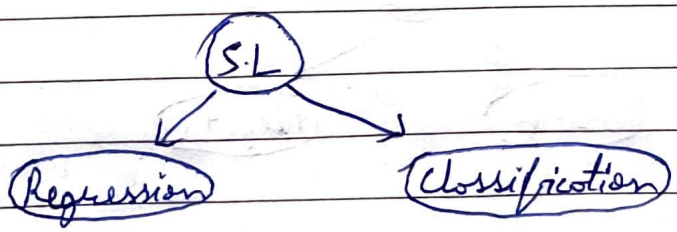
1) Supervised :- Machines are trained using well "labelled" training data and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

Advantages :-

- The model can predict the output on the basis of prior experience.
- We can have an exact idea about the classes of objects.
- It help us to solve real-world problems such as fraud detection, spam filtering etc.

Disadvantages :-

- They are not suitable for handling the complex tasks.
- It cannot predict the correct output if the test data is different from training dataset.
- Training required lots of computation times.
- We need enough knowledge about the classes of object.



Page :
Date: / /

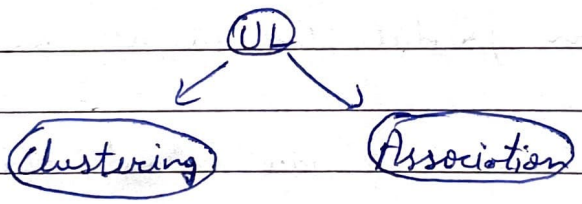
2) Un-Supervised :- Models are trained using unlabeled dataset and are allowed to act on that data without any supervision. Eg :- Human Brain.

Advantages :-

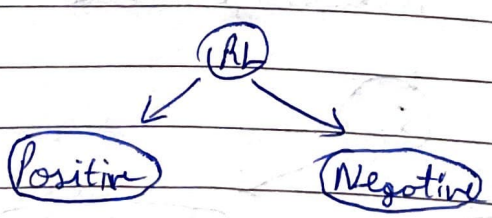
- Used for complex tasks as compared to supervised
- It is easy to get unlabeled data in comparison to labeled data.

Disadvantage :-

- More difficult than supervised because it does not have corresponding output.
- The result might be less accurate as input data is not labeled and algorithms donot know the exact output in advance.



3) Reinforcement :- It is a feedback-based ML technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback and for each bad action, the agent gets penalty.



Page :
Date: / /

⇒ Life Cycle :-

1) Gathering Data :- We need to identify the different data sources, as data can be collected from different sources such as files, database, internet or mobile devices. The quantity and quality will determine the efficiency of data.

Steps :- • Identify various data sources
• Collect data
• Integrate the data obtained from sources

By this step(1) we got coherent set of data called as dataset.

2) Data Preparation :- After collecting the data we put our data into a suitable place and prepare it to use in our ML.

• Data Exploration (Explore quality of data)
• Data Pre-processing

3) Data Wrangling :- It is the process of cleaning and converting raw data into a usable format. We use various filtering techniques to clean the data. It is mandatory to detect and remove the issues bcz it can negatively affect the data outcome.

Issues - Missing Values
Duplicate Data
Invalid Data
Noise.

4) Data Analysis :- The aim is to build a ML model

to analyze the data using various analytical techniques and review the outcome. We select the ML techniques such as Classification, Regression, Cluster analysis, Association etc. then built the model using prepared data and evaluate the model.

- Selection of analytical techniques.
- Building models.
- Review the result.

5) Train Model :- We train our model to improve its performance for better outcome of the problem. We use datasets to train the model. Training is required so that it can understand the various patterns, rules and features.

6) Test Model :- We check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage of accuracy.

7) Deployment :- If the prepared model is producing accurate result as per our requirements with acceptable speed then we deploy the model in the real system.

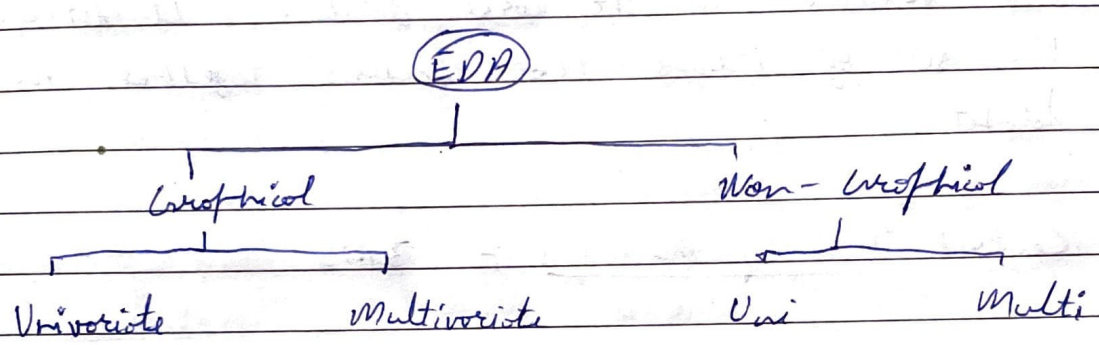
Data Discovery is same as step 1, 2, 3.

⇒ EDA (Exploratory Data Analysis) :- It is a technique to analyze datasets. It involves mostly graphical techniques.

Page :
Date: / /

Data Science

Life cycle :- ① Business Understanding
2) Data Acquisition
3) Data Preparation
4) ~~Model Planning~~ EDA
5) Model Building
6) Model Evaluation
7) Operationalize



⇒ Regression :- They are used if there is a relationship b/w the input variable and output variable. It is used for the prediction of continuous variables such as Weather forecasting etc..

Algorithms :- Linear Regression
Regression Trees
Non-Linear Regression
Bayesian Linear Regression
Polynomial Reg.

Classification :- They are used when the output is categorical which means there are two classes such as Yes-No, Male-Female etc.

Spam Filtering
Random Forest
Decision Trees
Logistic Reg
Support Vector Machines

Page :
Date: / /

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.

Association :- It is used for finding the relationships b/w variables in the large database. It determines the set of items that occurs together in the dataset.

⇒ Central Tendency Measures :- 3 M's

1) Mean :- It is the average of values. Eg :-

{ 10, 20, 30 }

$$\text{Mean} = \frac{10+20+30}{3} = 20$$

Not good when outliers is present.

2) Median :- It is the centrally located value of the dataset sorted in ascending order.

If odd values is present Eg :- There are 3 value, $\frac{3}{2} = 1.5$, So Median will be 2.

If even values is present like $\frac{4+4}{2} = 4$

So Median is 4. (Two mid values are added)

3) Mode :- It is the most frequent value in the dataset. Most repeating element in the dataset. Eg :- 1, 2, 3, 3, 3

Mode = 3

Page :
Date: / /

⇒ Correlation :- It is a measure of association. It is used for bivariate analysis. It is a measure of how well two variables are related.

Covariance :- It is a measure of association. Relationship measure b/w two variables or measure of strength and direction.

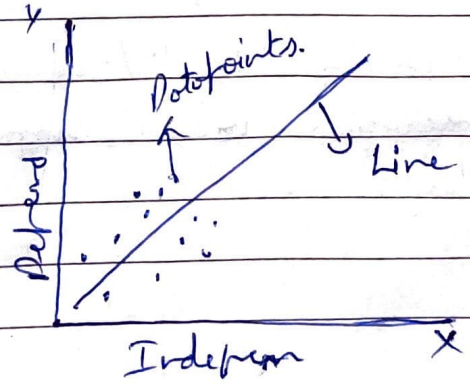
⇒ Quartile :- The values that divide a list of numerical data into three quaters. The middle part is central point of distribution and shows the data which are near to the central point.

(25%) $Q_1 = \frac{N+1}{4}$

(50%) $Q_2 = \frac{2N+1}{4}$

(75%) $Q_3 = \frac{3N+1}{4}$

⇒ Linear Regression :- It shows a linear relationship b/w a dependent and independent variable. It finds how the value of dependent variable is changing according to the value of independent variable.



Page :
Date: / /

Simple — Multiple

Simple LR :- A single independent variable is used to predict the value of a numerical dependent variable.

$$y = mx + c$$

m = slope
c = Reg. Coefficient
x = independent
y = Dependent

Multiple :- If more than one independent variable is used to predict the value of a dependent variable.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \ldots\ldots \alpha_m x_m$$

α = Reg. Coefficient (alpha)
x = Independent
y = Depend

⇒

ML Errors
- Reducible
  - Bias — Low, High
  - Variance — Low, High
- Irreducible

Irreducible :- These error will always be present in the model.

Bias :- While making predictions a difference occurs b/w prediction values made by the model and actual value.

Page :
Date: / /

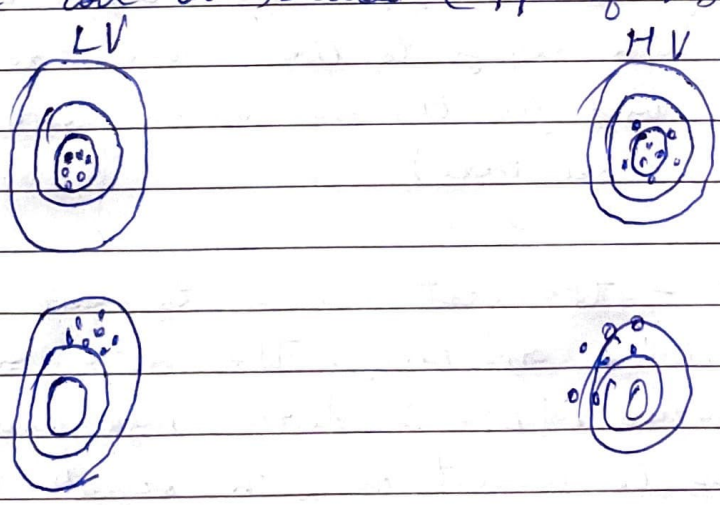
Low Bios :- Predicted value is very close to actual value. Eg :- Decision Tree, SVM.
High Bios :- Predicted value is far from the actual value. Eg :- Linear Reg, Logistic Reg.

Ways :- High bios mainly occurs in a simple model. Using more complex model or increasing the input will fix this.

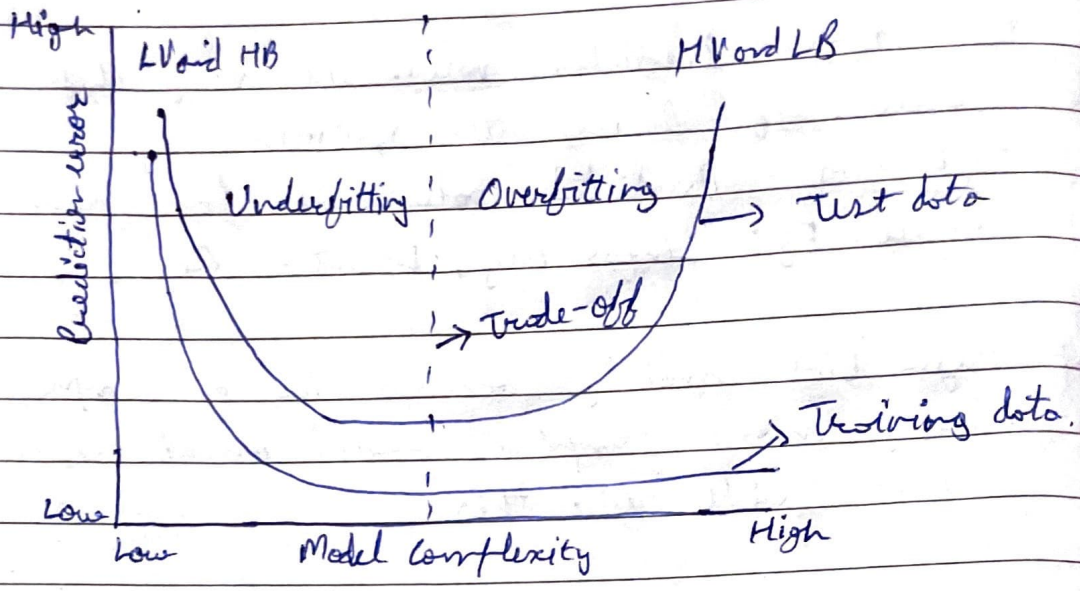
Variance :- It tells how much a random variable is different from its expected value.

Low Var :- Small variation in the prediction. (Data are close to each other not scattered.).
High Var :- Large variation in the prediction (Data is scattered.). Eg :- Vice-versa of bios for both.

Ways :- High Var can be reduced (opp of high bios)



Bios-Variance Trade off :- It is required to make a balance b/w bios and variance errors, this balance is called ~~~.



⇒ Model Validation :- It is the procedure of evaluating the wellness of models performance against the real data.

Approaches :-

- Resubstitution :- Using all the data for training the model, the validity of the model evaluated by comparing the output value with an actual value which belongs to the same training dataset. (All data used for training nothing left for test so errors are there).

- Hold-out :- The best way is to divide the dataset into training and test. The ratio can be 80/20 70/30 60/40. (Data is divided into two but some data can be present in both or all the data goes in one (specific data like a class) resulting in error.

Page :
Date: / /

- K-Fold Cross :- Dataset is divided into k number of subsets where K-1 subsets for training and one for testing. (Overlapping of data doesn't happen error count is less (K=3))

- LOOCV :- Only one record used for testing and the rest of other records are used for training.

- Scatter Plot :- This a graphical representation of the value predicted concerning the actual values. It calculates the accuracy of the model.

- Random Subsampling :- The number of subsets selected and then combined to form a super subset used for testing and the rest of the data used for training.

⇒ Mean Squared Error (MSE) :- We calculate the error by squaring the difference b/w the predicted values and actual values and averaging it (Mean) across the dataset. It is also known as Quadratic loss. It can never be negative bcz we square it.

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i → Actual value
$\hat{y}_i$ → Predicted value
N → No. of values
Σ → Sum of all values

Page :
Date: / /

Root Mean Squared Error (RMSE) :- It is computed by taking the root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MSE is bad ~~for~~ when outlier are there as it squares them but in RMSE outliers have to be first removed for RMSE to function properly.

Unit - 2