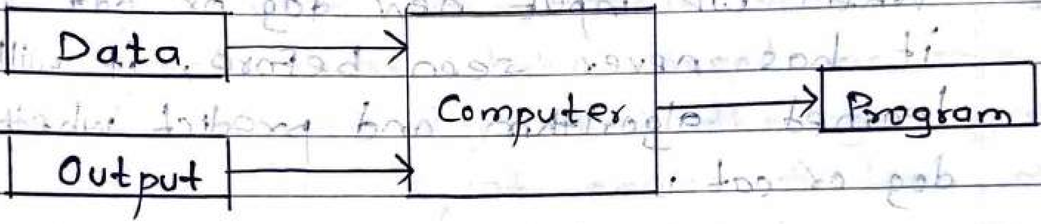
Q.B P.T - I

Q.1 What is learning? Explain different types of learning with example.

⇒

- Machine learning is a category of AI. In machine learning computers has the ability to learn themselves, explicit programming is not required.
- Machine learning focuses on the study and development of algorithms that can learn from data and also make predictions on data.

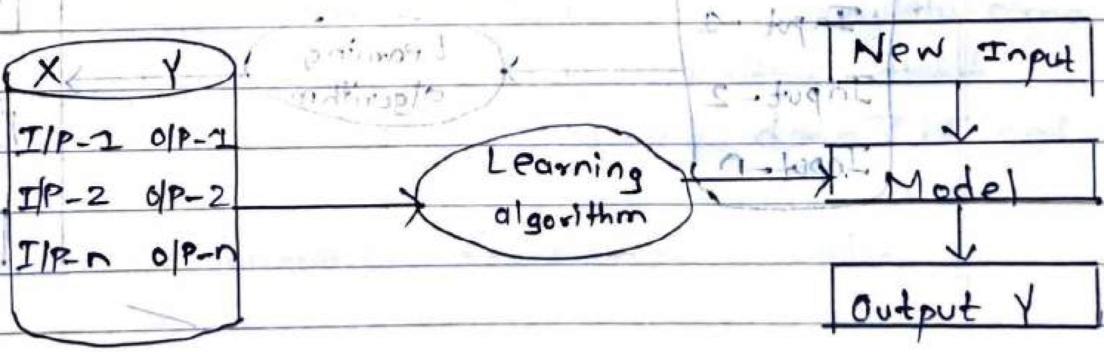


- Different types of learning :-

1) <u>Supervised Learning</u>

⇒ In this type of learning we use data which is comprises of input and corresponding output.

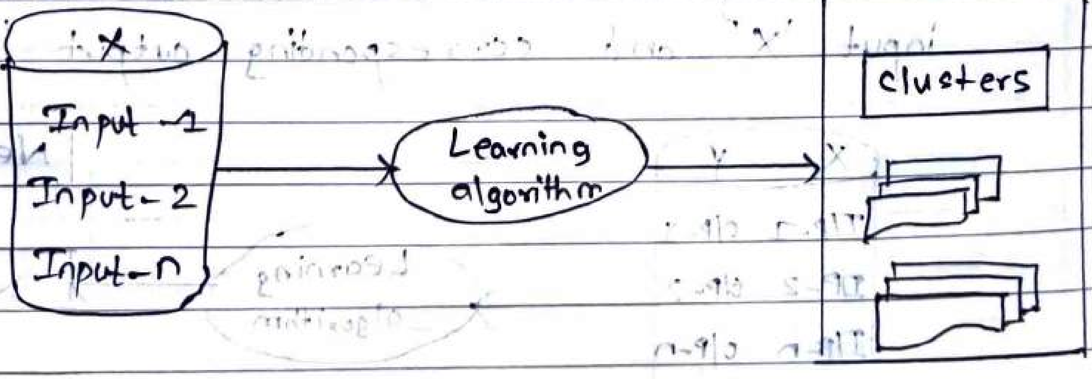
- For every instance of data we can have input 'X' and corresponding output 'Y'.



- From this ML system will build model so that given an observation 'x', for new observation 'x' it will try to find out what is corresponding 'Y'.
- For eg., Consider scenario where you have to build image classifier to differentiate between cats and dogs.
- If you feed the datasets of dogs and cats labelled images to the algorithm, the machine will learn to classify between a dog or cat from these labelled images.
- When we input new dog or cat images that it has never seen before, it will use the learned algorithm and predict whether it is dog or cat.

2) Unsupervised Learning

⇒ In unsupervised learning you are only given input 'x', there is no label to the data and given the data or different data points, you may want to form clusters or want to find some pattern.



- For e.g., consider that you have dataset that contains information about the purchases you made from the shop.
- Through clustering, the algorithm can group the same purchasing behaviour among you and other customers, which reveals potential customers without predefined labels.

3) Reinforcement Learning

⇒ In reinforcement learning you have an agent who is acting in an environment and you want to findout what actions the agent must take based on the reward or penalty that the agent gets it.
- In this an agent seeks to learn the optimal actions to take based on ~~custome~~ outcomes of past actions.
- For e.g., AI agent to play game like chess.

4) Semi-supervised learning

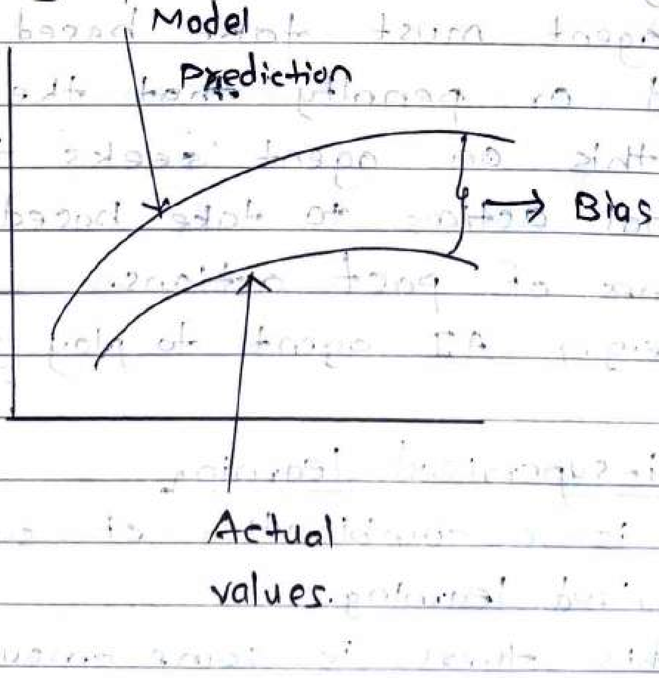
⇒ It is a combination of supervised and unsupervised learning.
- In this there is some amount of labeled training data and also you have large amount of unlabeled data and you try to come up with some learning algorithm that convert even when training data is not labeled.
- For eg, language translation model.

Q.2 Explain the following terms with the help of appropriate graph.

a) Bias

⇒ - Bias is defined as the inability of the model because of that there is some difference or error occuring betwoon the model's predicted value and actual value.

- Bias is a systematic error that occurs due to wrong assumption in the machine learning process.

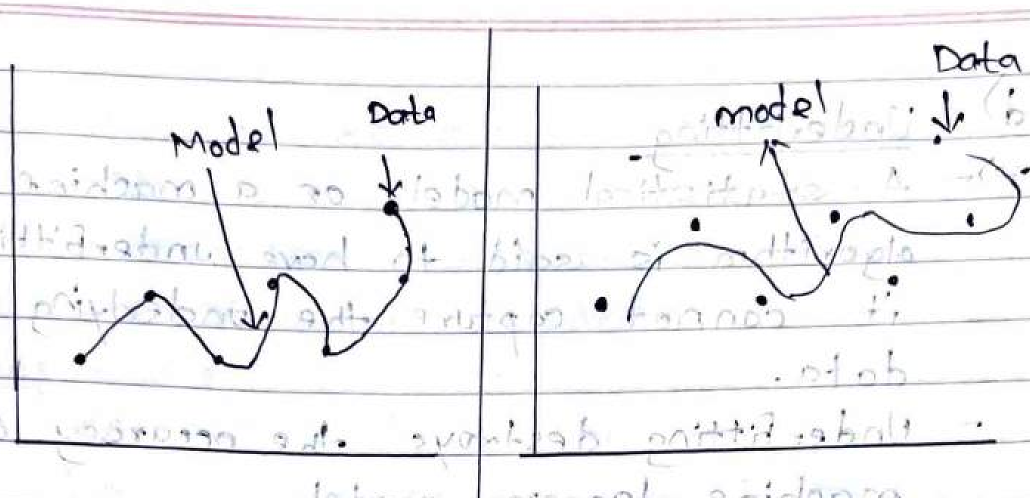


b) Variance

⇒ Variance is the measure of spread in data from its mean position.

- In ML, variance is the amount by which the performance of a predictive model changes when it is trained or different subsets of the training data.

classmate
Date
Page

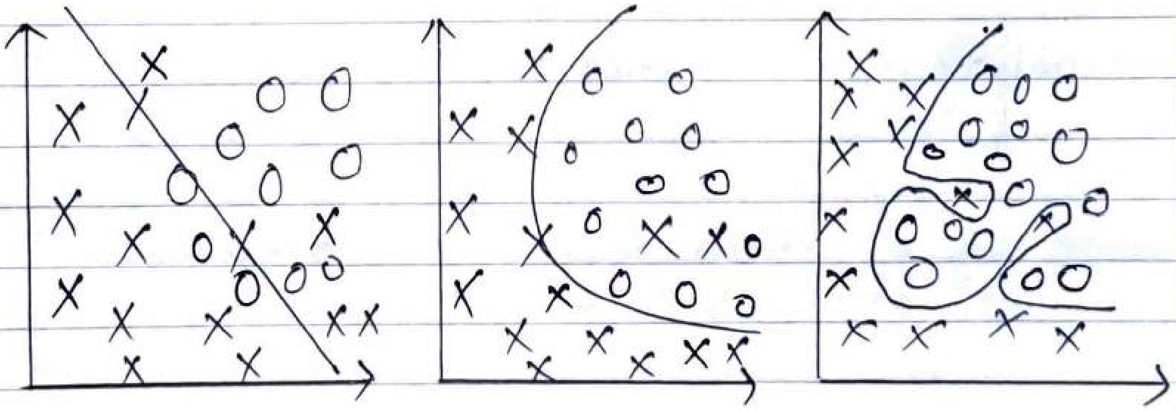


(training data) (new data)

c) Overfitting

⇒ - A statistical model is said to be overfitted, when we train it with lot of data.

- When model gets train with so much of data, it starts learning from the noise and inaccurate data entries in our data set.
- Then the model does not categorize the data correctly, because of too many details and noise.
- In a nutshell, Overfitting - High variance and low bias.



(Under-fitting) (Appropriate-fitting) (Over-fitting)

d) Underfitting

⇒ - A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of data.

- Underfitting destroys the accuracy of ~~data~~ our machine learning model.
- Its occurence simply means that our model or the algorithm does not fit the data well enough.
- In a nutshell, Underfitting - High bias and low variance.

Q.3 Discuss how to evoluate a ML model for overfitting and underfitting, explain using diagram What measures needs to be taken in case of overfitting and underfitting.

⇒ • <u>Underfitting</u>

⇒ A stastical model or ML algorithm is said to have underfitting when it cannot capture the underlying trend of the data.

- Underfitting destroys the accuracy of our ML model.
- In a nutshell, Underfitting - High bias and low variance.

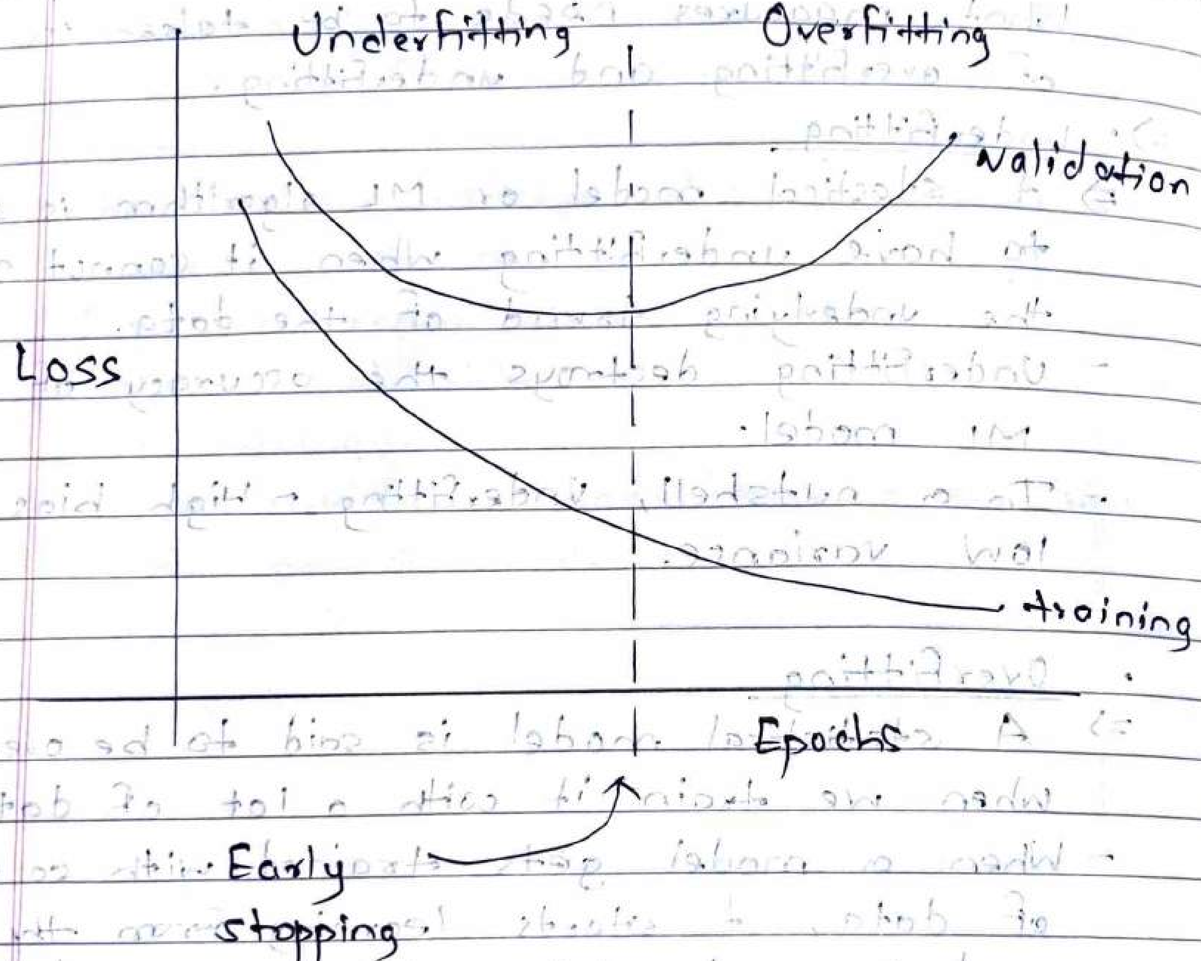
• <u>Overfitting</u>

⇒ A statistical model is said to be overfitted, when we train it with a lot of data.

- When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our dataset.
- Then model does not categorize the data correctly, because of too many details and noise.
- In a nutshell, Overfitting - High variance and low bias.

- To evaluate whether a model is overfitting or underfitting, we typically use learning curves.
- A learning curve plots the model's performance on the training and validation datasets as a function of the number of

training epochs, the size of training set or model complexity.



• Measures to address overfitting :

1) Regularization

⇒ Apply regularization techniques such as L1 and L2 regularization to penalize large coefficients in the model.

2) Cross-validation

⇒ Use techniques like k-fold cross-validation to ensure that the model generalizes well across different subsets of the data.

- Measures to address underfitting :

1) Increase Model Complexity

⇒ Use more complex model that can capture the underlying patterns in the data.

2) Reduce regularization

⇒ If regularization is too strong, it can contains the model too much, leading to underfitting. Consider reducing the regularization strength.

Q.4 Illustrate the process of learning with the gradient descent for a simple linear regression using bell shaped error curve.

⇒

- In simple linear regression, the goal is to find the best-fitting line through the data points, which minimizes the error between the predicted values and actual values.
- Gradient descent is an optimization technique used to minimize this error by iteratively adjusting the model parameters.
- The model in simple linear regression is represented as -

$$Y' = aX + b$$

where,

Y' = predicted value,
a = slope of the line,
b = intercept
X = input feature

- The error function typically used is the Mean Squared Error (MSE).

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

where,

n = no. of data points
$\hat{y}_i$ = predicted value
y_i = actual value

classmate
Date
Page

- Gradient descent aims to find the values of a and b that minimizes the MSE.
- The update rules for the parameters using gradient descent are :-

$$a = a - \alpha \cdot \frac{\partial MSE}{\partial a}$$

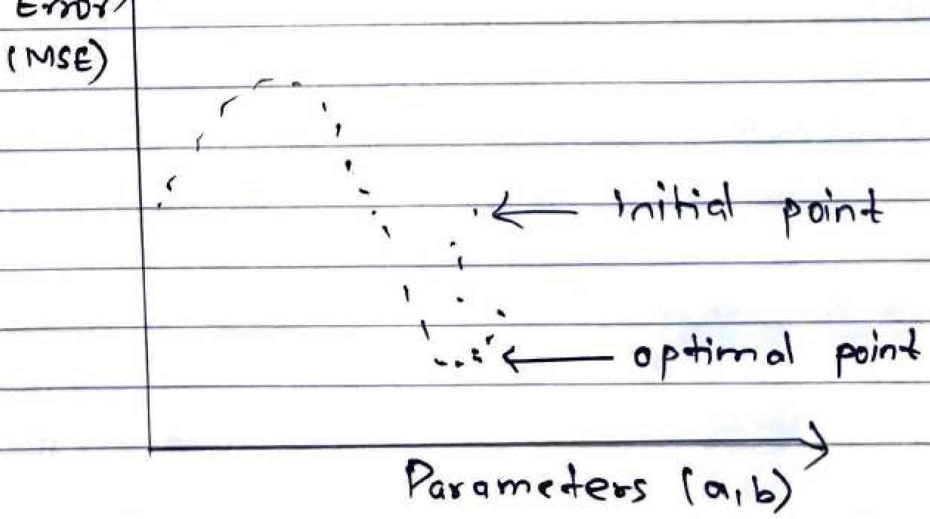
$$b = b - \alpha \cdot \frac{\partial MSE}{\partial b}$$

where, α is learning rate.

- Bell Shaped Error Curve

⇒ Imagine plotting the error (MSE) as a function of the model parameters a and b

- The error surface might look like a bell-shaped curve.
- The minimum point on this curve represents the optimal values of a and b, where the error is minimised.
- Gradient descent moves the parameters iteratively downhill on this surface until the minimum point is reached.



Q.5 Explain the steps in developing ML applications
⇒

1) Collection of Data
⇒ You could collect the samples from website and extracting data.
- From RSS feed or an API.
- From device to collect wind speed meaasurement
- Publicly available data.

2) Preparation of the input data
⇒ Once you have the input data, you need to check whether it's in a useable format or not.
- Some algorithms can accept target variables and Features as string; some need them to be integers.
- Some algorithms accepts features in special format.

3) Analyse the input data
⇒ Looking at the data you have passed in a text editor to check collection and preparation of input data steps are properly working and you don't have a bunch of empty values.
- Plotting data in 1, 2, or 3 dimensions can also help.

4) The importance of this step is that it makes you understand that you don't have any garbage value coming in.

5) Train the algorithm

⇒ Good clean data from the first two steps is given to the algorithm.

- The algorithm extracts the information or knowledge.
- This knowledge is mostly stored in format that is readily usable by machine for next 2 steps.

6) Test the algorithm

⇒ In this step the information learned in the previous step is used.

- When you are checking an algorithm, you will test it to find out whether it is works properly or not.
- If you are not satisfied with result, you can again go back to step 4, change some thing and test again.

7) Use it

⇒ In this step, real program is developed to do some task and once again it is checked if all the previous steps worked as you expected.

Q.6 Explain issues in ML

⇒

- Issues in ML :

1) Poor quality of data

⇒ Data plays a significant role in ML process.

- One of the significant issues that ML professionals face is the absence of good quality data.
- Unclean and noisy data can make the whole process extremely exhausting.
- We don't want our algorithm to make inaccurate or faulty predictions.
- Hence the quality of data is essential to enhance the output.

2) Underfitting of training data

⇒ This process occurs when data is unable to establish an accurate relationship between input and output variables.

- It signifies the data is too simple to establish a precise relationship.

3) Overfitting of training data

⇒ Overfitting refers to machine learning model trained with a massive amount of data that negatively affects its performance.

- This means algorithm is trained with noisy and biased data, which will affect its overall performance.

4) <u>Lack of training data</u>

⇒ The most important task you need to do in the ML process is to train data to achieve an accurate output.

- Less amount of training data will produce inaccurate or too biased predictions.

5) <u>Slow implementation</u>

⇒ This is one of the common issues faced by machine learning professionals.

- The machine learning models are highly efficient in providing accurate results, but it takes tremendous amount of time.
- Further, it requires constant monitoring and maintainance to deliver the best output.

Q.7 Write short note on applications of ML.

⇒

- Applications of ML :-

1) <u>Image Recognition</u>

⇒ Image recognition is one of the most common application of machine learning.

- It is used to identify objects, persons, places, digital images, etc.
- The popular use case of image recognition and face detection is Automatic friend tagging suggestion.

2) <u>Traffic prediction</u>

⇒ If we want to visit new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

- It predicts the traffic conditions such as whether traffic is cleared, slow-moving or heavily congested with the help of two ways:

i) Real time location of vehicle from google map and sensors.

ii) Average time has taken on past days at the same time.

3) <u>Product recommendations</u>

⇒ ML is widely used by various e-commerre and entertainment companies such as Amazon, Netflix, etc. for product recommendations.

classmate
Date
Page

to the users.

- Whenever we search some product on amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of ML.
- As similar, when we use Netflix, we find some recommendations for entertaiment series, movies, etc. and this is also done with the help of ML.

4) Self-driving cars

⇒ One of the most exfciting application of ML is self-driving cars.

- ML plays a significant role in self-driving cars.
- Tesla, the most popular car manufacturing company is working on self-driving cars.
- It is using unsupervised learning method to train the car models to detect peoples and objects while driving.

5) Speech Recognition

⇒ While using Google, we get an option of "search by voice", It comes under speech recognition and it's popular application of ML.

- Speech recognition is a process of converting voice instruction into the text, and it is also known as "Speech to text".
- Google assistant, Siri and Alexa are using speech recognition technology.

classmate
Date ______
Page ______

Q.8 Compare decision tree classification ~~and~~ with logistic regression classification.

=>

Decision tree classification	Logistic Regression classification
1) A decision tree is a type of supervised learning algorithm that is commonly used in ML to model and predict outcomes based on input data.	1) Logistic regression is also type of supervised learning algorithm used for classification tasks where the goal is to predict probability that an instance belongs to a given class or not.
2) Highly interpretable, as the tree structure clearly shows the decisions made at each node.	2) Coefficients can be interpreted to understand the influence of each feature on the outcome.
3) Prone to overfitting	3) Less prone to overfitting compared to decision tree.
4) ~~It is robust to noise~~	4) ~~It is majorly affect~~
4) It is majorly affected by noise.	4) It is robust to noise.
5) Can be trained on small training set.	5) Requires a large enough training dataset.

Q.9. Explain the properties of GINI Index.

⇒

- The Gini index is a measure used in decision trees to evaluate quality of a split.
- It quantifies the degree of impurity or disorder in a dataset.
- The Gini is used in Classification and Regreession Tree (CART)
- If a dataset T contains example from n classes, gini index, gini(T) is defined as -

$$Gini(T) = 1 - \sum_{j=1}^{n} (P_j)^2$$

- In the above equation, Pj represents the relative frequency of class j in T.
- After splitting T in two subsets T_1 and T_2 with sizes N_1 and N_2, gini index of split data is,

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

- The attribute with smallest $gini_{split}(T)$ is selected to split the node.
- The Gini index ranges from 0 to 0.5.
- 0 indicates perfect purity, meaning all instances belong to a single class.
- 0.5 represents maximum impurity, indicating that the instances are uniformly distributed across all classes.

Q.10 Discuss in brief pruning is Decision tree

⇒

- Pruning in decision tree is a technique used to reduce the size of a decision tree by removing sections of tree that provide little to no additional predictive power.
- The main goal of pruning is to prevent overfitting and improve the model's ability to generalize to unseen data.
- Decision tree can grow very large and complex, capturing noise and outliers in the training data, which leads to overfitting.
- Pruning helps this mitigate by simplifying the tree.
- By removing unnecessary branches, pruning reduces the complexity of the model, leading to better performance on new, unseen data.
- There are two main types of pruning in decision trees:-

1) Pre-pruning

⇒ This involves halting the growth of the tree, before it becomes too complex.

- It sets a limit on tree depth, minimum samples per leaf, or minimum information gain required for a split.

2) Post-pruning

⇒ This involves first allowing the tree to grow fully and then pruning back unnecessary branches.

classmate
Date
Page

- The pruning techniques are :-

1) Cost - Complexity Pruning

⇒ This method ~~gets~~ assigns a price to each subtree primarily based on its accuracy and complexity, then selects the subtree with the lowest fee.

2) Reduced Error Pruning

⇒ Removes the branches that do not significantly affect the overall accuracy.

• Advantages :-

1) Simplifies the model, improving its generalization to new data.
2) Results in smaller, more interpretable tree.
3) Reduces the size of the model, leading to facter predictions.

• Disadvantages :-

1) IF pruning is too agressive, it may remove important branches, leading to a model that is too simple to capture the underlying data patterns.
2) Not suitable for predictions of continuous attribute.
3) Computationally expensive to train.

Classmate
Date
Page

Q.11 Problem based on Simple Linear Regression

⇒ Following table shows the midterm and final exam grades obtained for student in a database course. Use the method of least squares using regression to predict the final exam ~~of~~ grade of a student who received 86 in the mid term exam.

Midterm exam(X)	72	50	81	74	94	86	59	83	86	33	88	81
Final exam (Y)	84	53	77	78	90	75	49	79	77	52	74	90

⇒

Soln:-

	X	Y	XY	X^2
	72	84	6048	5184
	50	53	2650	2500
	81	77	6237	6561
	74	78	5772	5476
	94	90	8460	8836
	86	75	6450	7396
	59	49	2891	3481
	83	79	6557	6889
	86	77	6622	7396
	33	52	1716	1089
	88	74	6512	7744
	81	90	7290	6561
Total	887	878	67205	69113

The equation for the regression line is -

$$Y' = aX + b$$

$$a = \frac{n\Sigma XY - \Sigma X \cdot \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{12(67205) - 887 \times 878}{12(69113) - (887)^2}$$

$$\boxed{\therefore a = 0.65}$$

$$b = \frac{1}{n}(\Sigma Y - a\Sigma X)$$

$$= \frac{1}{12}(878 - 0.65 \times 887)$$

$$\boxed{\therefore b = 25.12}$$

$$\therefore Y' = 0.65X + 25.12$$

The final grade of a student who received 86 in the mid term exam,

$$Y' = 86 \times 0.65 + 25.12$$

$$\boxed{\therefore Y' = 81.02}$$

classmate
Date
Page

Q.12 Problem based on decision tree algorithm

Suppose we want ID3 to decide whether the car will be stolen or not. The target classification is "car is stolen?" which can be Yes or No.

Car No.	Colour	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes