Unit - 2

Supervised Learning

Que- Difference between classification and Regression.
⇒

classification	Regression
1] Predicts categorical labels or classes.	1] Predicts continuous numeric values.
2] Discrete labels	2] continuous values
3] Example :- Email spam detection, disease diagnosis	3] Example :- House Price prediction, Stock Price
4] Assigns input data to one of several Predefined categories.	4] Estimates a continuous value based on input data.
5] Evaluation metrics :- Accuracy, Precision, recall, F1 score, ROC AUC.	5] Mean squared Error (MSE), Mean Absolute Error (MAE), R-squared.
6] Algorithms :- Decision Tree, SVM, Naive Bayes	6] Linear Regression, Ridge Regression.
7] Class labels or Probabilities of class membership.	7] Real-valued outputs.
8] Minimize classification error	8] Minimize error between Predicted and actual continous values.

Que. Explain following terms :- ① underfitting ② overfitting ③ Bias ④ variance ⑤ Generalized model.

→

① Underfitting :-

- Definition :- Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in the data.
- It fails to learn the relationships between the input and output, leading to poor performance on both the training data and new, unseen data.
- Cause :- This is often caused by using models that are too simple or with insufficient training.
- Example :- A straight line trying to fit a curved dataset.

② Overfitting :-

- Definition :- overfitting happens when a model learns the training data too well, including noise and irrelevant patterns.
- While the model performs excellently on training data, it fails to generalize to new, unseen data.
- Cause :- This is usually due to an overly complex model with too many parameters relative to the data size.
- Example :- A model with high flexibility fitting even random fluctations in the training set, resulting in poor performance on test data.

3) Bias :-
- Definition :- Bias refers to the error introduced by approximating a complex real-world problem with a simplified model.
- High bias occurs when a model makes strong assumptions about the data, leading to underfitting.
- Example :-
  using a linear model to fit data that is clearly non-linear.

4) Variance :-
- Definition :- variance refers to the models sensitivity to small fluctuations in the training data.
- High variance models tend to overfit the training data, performing well on the training data but poorly on unseen data.
- Effect :- High variance causes the model to focus too much on the training data, including noise, resulting in poor generalization.
- Example :- A decision tree that grows too deep and fits the noise of the training data.

5) Generalized model :-
- Definition :- A generalized model is one that performs well on both training data and unseen data, meaning it neither underfits nor overfits.
- The goal of a generalized model is to capture the underlying patterns in the data while ignoring noise or irrelevant information.

•Effect:- A generalized model strikes the right balance between bias and variance, leading to good predictions on new data.
•Example:- A model that performs similarly well on both training and test datasets, indicating it has learned the true patterns in the data.

Que- Explain K-Nearest Neighbors (KNN) with example.

⇒ K-Nearest Neighbors (KNN):-
- K-Nearest Neighbors (KNN) is a simple, instance-based machine learning algorithm that is often used for classification and regression tasks.
- It works by finding the 'k' closest data points (neighbors) to a given query point and making predictions based on these neighbors.

How to choose the value of k for KNN:-
- The value of k in the k-nearest neighbors (KNN) algorithm should be chosen based on the input data.
- If the input data has more outliers or noise a higher value of k would be better.
- It is recommended to choose an odd value for k to avoid ties in classification.
- cross validation methods can help in selecting the best k value for the given dataset.

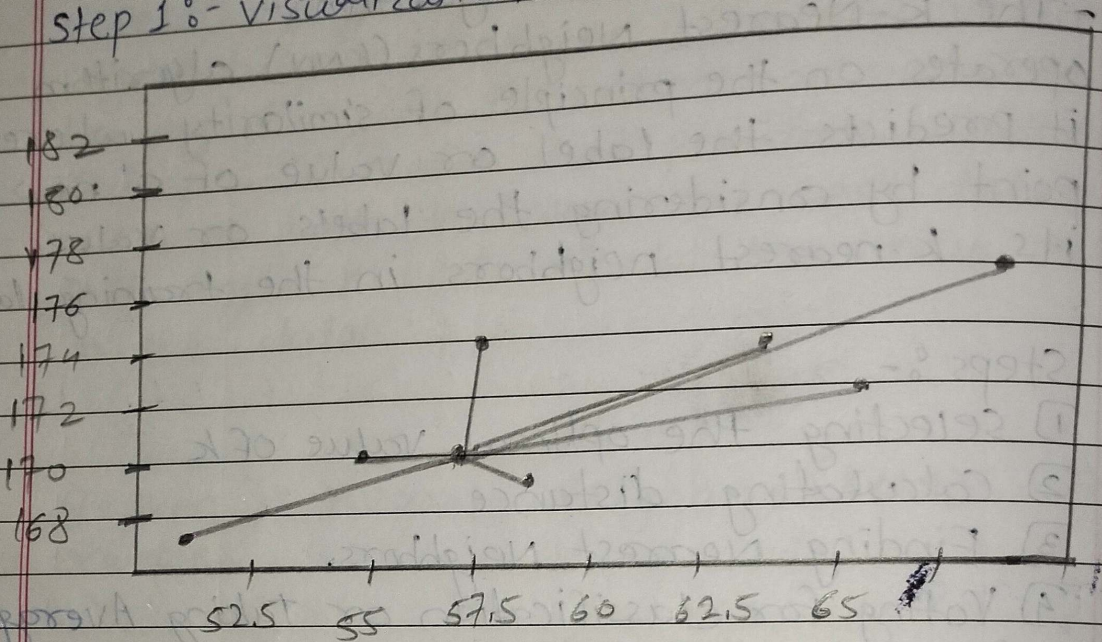
• Working of KNN algorithm :-
- The k-Nearest Neighbors (KNN) algorithm operates on the principle of similarity where it predicts the label or value of a new data point by considering the labels or values of its k-nearest neighbors in the training dataset.

Steps :-
1) Selecting the optimal value of k
2) Calculating distance
3) Finding Nearest Neighbors.
4) Voting for classification or taking Average for ~~eee~~ Regression.

• Example :-

Height	Weight	class
167	51	underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

step 1:- visualization



step 2:- using Encludine formula we going to Find distance between given data points with Prediction Data point.

$$\sqrt{[(x_1-x_2)^2+(y_1-y_2)^2}$$

here, $x_2 = 170$, $y_2 = 57$ — Prediction values

x_1 and y_1 — given values.

~~① $\sqrt{(167-170)^2+(57-51)^2} = 6.71$~~

②

① 167 (x_1) 51 (y_1)

170 (x_2) 57 (y_2)

$= \sqrt{(167-170)^2+(57-51)^2}$

$= \sqrt{(-3)^2+(-6)^2}$

$= \sqrt{9+36}$

$= \sqrt{45}$

$= 6.71$

② x_1 182, y_1 62; 170 x_2, 57 y_2

$= \sqrt{(182-170)^2 + (62-57)^2}$
$= \sqrt{(12)^2 + (5)^2}$
$= \sqrt{144 + 25}$
$= \sqrt{169}$
$= 13$

③ x_1 176, y_1 69; 170 x_2, 57 y_2

$= \sqrt{(176-170)^2 + (69-57)^2}$
$= \sqrt{(6)^2 + (12)^2}$
$= \sqrt{36 + 144}$
$= \sqrt{180}$
$= 13.4$

④ x_1 173, y_1 64; 170 x_2, 57 y_2

$= \sqrt{(173-170)^2 + (64-57)^2}$
$= \sqrt{(3)^2 + (7)^2}$
$= \sqrt{9 + 49}$
$= \sqrt{58}$
$= 7.6$

⑤ 172, 65; 170, 57

$= \sqrt{(172-170)^2 + (65-57)^2}$
$= \sqrt{(2)^2 + (8)^2}$
$= \sqrt{4 + 64}$
$= \sqrt{68} = 8.2$

⑥ x_1 174, y_1 56; 170 x_2, 57 y_2

$= \sqrt{(174-170)^2 + (56-57)^2}$
$= \sqrt{(4)^2 + (-1)^2}$
$= \sqrt{16 + 1}$
$= \sqrt{17}$
$= 4.1$

⑦ x_1 169, y_1 58; 170 x_2, 57 y_2

$= \sqrt{(169-170)^2 + (58-57)^2}$
$= \sqrt{(-1)^2 + (1)^2}$
$= \sqrt{1 + 1}$ ∴
$= \sqrt{2}$
$= 1.4$

⑧ x_1 173, y_1 57; 170 x_2, 57 y_2

$= \sqrt{(173-170)^2 + (57-57)^2}$
$= \sqrt{(3)^2 + (0)^2}$
$= \sqrt{9}$
$= 3$

⑨ x_1 170, y_1 55; 170 x_2, 57 y_2

$= \sqrt{(170-170)^2 + (55-57)^2}$
$= \sqrt{(0)^2 + (-2)^2}$
$= \sqrt{4}$
$= 2$

k = 2 or more
k value greter than 1

Height	weight	class	Distance	Rank
169	58	N	1.4	1
170	55	N	2	2
173	57	N	3	3
174	56	UW	4.1	4
167	51	UW	6.7	5
173	64	N	7.6	6
172	65	N	8.2	7
182	62	N	13	8
176	69	N	13.4	9
170	57	?		

∴ Take k = 3 and Predicted class for 170 and 57 data point is Normal.

- Advantages :-
1] Simplicity
2] No Assumptions
3] Versatility
4] Adaptability

- Disadvantages :-
1] computationly Expensive
2] Memory -Intensive
3] sensitive to Noise
4] Feature scaling required

- Applications :-
1] Pattern recognition
2] Recommendation system
3] Anomaly Detection
4] stock market prediction

Que- Explain support vector machine (svm) with example.

⇒

Support vector machine (svm):-
- support vector machine (svm) is a supervised machine learning algorithm commonly used for classification tasks, though it can also be applied to regression problems.
- svm aims to find the optimal hyperplane that best separates different classes in the feature space.

• key concepts of svm:-

1) Hyperplane:- A decision boundary that separates classes.
- In 2D, its a line; in 3D, its a plane.
- In higher dimensions, it generalizes similarly.

2) Support vectors:- The data points closest to the hyperplane.
- They influence its position and help maximize the margin.

3) Margin:- The distance between the hyperplane and the nearest data points from each class.
- svm aims to maximize this margin for better generalization.

4) Linear vs Non-Linear:-
• Linear:- separates data with a straight-line hyperplane.

- Non-linear :- uses a kernel trick to handle complex, non-linear separations.

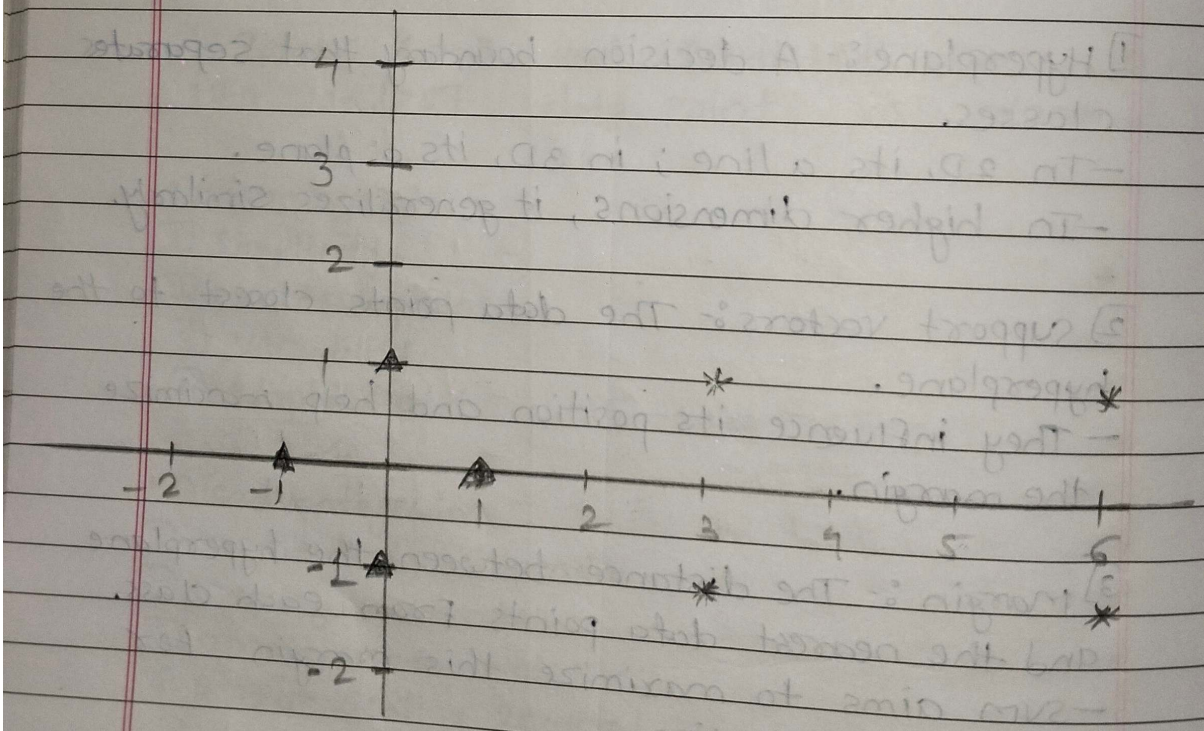
- Example :-
We are given the following positively labeled data points

$$\left\{\begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}3\\-1\end{pmatrix}, \begin{pmatrix}6\\1\end{pmatrix}, \begin{pmatrix}6\\-1\end{pmatrix}\right\}$$

and the following negatively labeled data points.

$$\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}\right\}$$

Step 1 :- Visualization



Step 2 :- Add bias
Each vector is augmented with a 1 as bias input so, $S_1 = \begin{pmatrix}1\\0\end{pmatrix}$ then $\hat{S}_1 = \begin{pmatrix}1\\0\\1\end{pmatrix}$

Similarly,

$S_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ then $\tilde{S}_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$

and

$S_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ then $\tilde{S}_3 = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$

step 3:-

$\alpha_1 \tilde{S}_1 \cdot S_1 + \alpha_2 \tilde{S}_2 \cdot S_1 + \alpha_3 \tilde{S}_3 \cdot S_1 = -1$

$\alpha_1 \tilde{S}_1 \cdot S_2 + \alpha_2 \tilde{S}_2 \cdot S_2 + \alpha_3 \tilde{S}_3 \cdot S_2 = +1$

$\alpha_1 \tilde{S}_1 \cdot \tilde{S}_3 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_3 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_3 = +1$

$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = -1$

$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} = 1$

$\alpha \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = 1$

$\alpha_1 (1+0+1) + \alpha_2 (3+0+1) + \alpha_3 (3-0+1) = -1$

$\alpha_1 (3+0+1) + \alpha_2 (9+1+1) + \alpha_3 (9-1+1) = 1$

$\alpha_1 (3+0+1) + \alpha_2 (9-1+1) + \alpha_3 (9+1+1) = 1$

$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$

$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1$

$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1$

$\alpha_1 = -3.5$

$\alpha_2 = 0.75$

$\alpha_3 = 0.75$

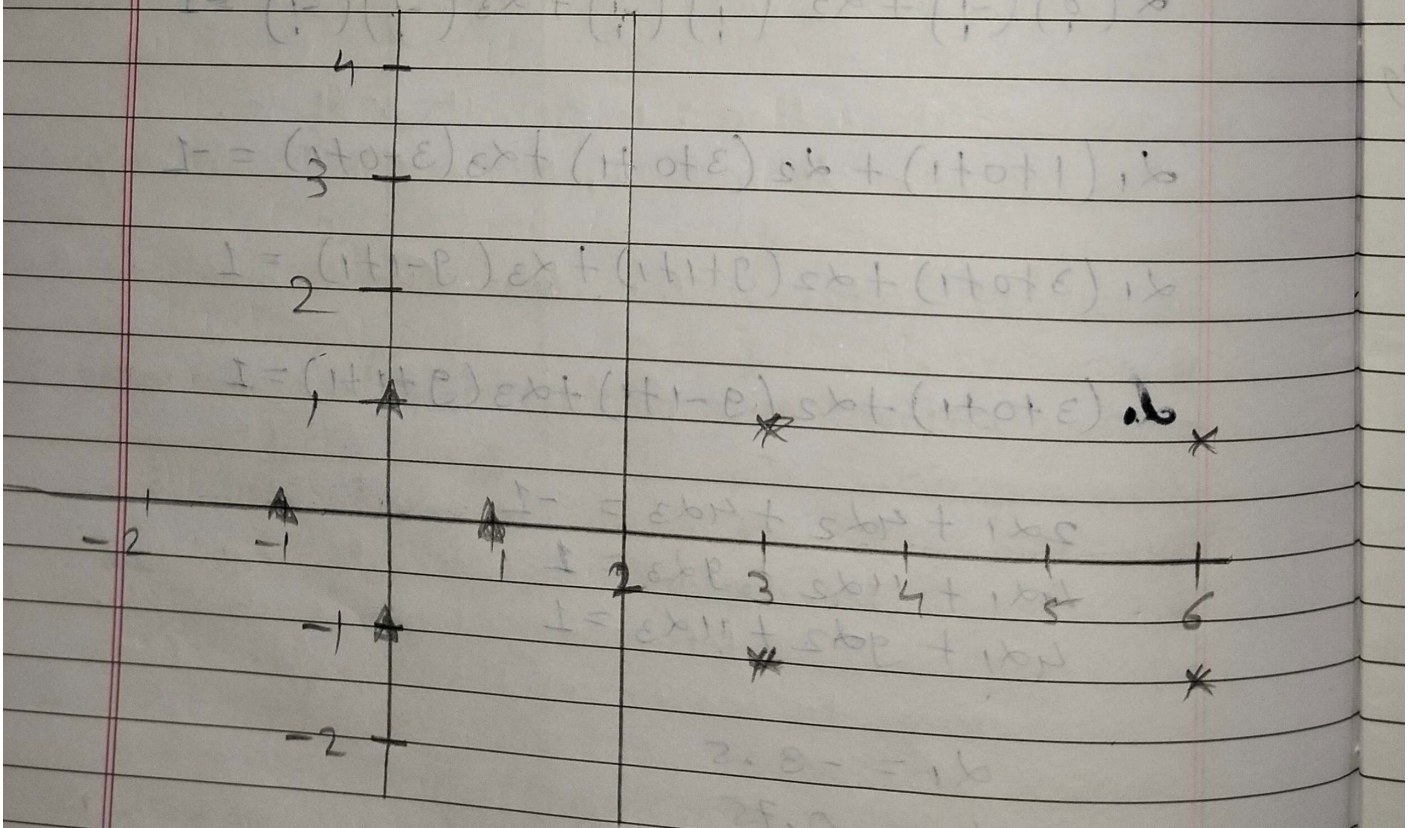
$$\tilde{\omega} = \sum \alpha_i \tilde{s_i}$$

$$= -3.5 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

Finally remembering that our vectors are augmented with a bias.

- We can equate the last entry in $\tilde{\omega}$ as the hyperplane offset b and write the separating.
- Hyperplane equation $y = \omega x + b$ with $\omega = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $b = 2$



• Advantages :-
1) Effective in high-dimensional spaces
2) Robust to overfitting
3) works with non-linear data.

• Disadvantages :
1) Computationally expensive
2) Difficult to interpret
3) sensitive to choice of kernel.

• Application :-
1) Image classification
2) Text categorization
3) Handwriting Recognition
4) medical Diagnosis.

Que- Explain Decision Tree with example.

⇒

Decision Tree :-

- A Decision Tree is a supervised machine learning algorithm used for classification and regression tasks.
- It works by splitting the data into subsets based on the values of input features, creating a tree-like structure. where each node represent a decision based on one of the features and each branch represents the outcome of that decision.

• structure of a Decision Tree :-

1) Root Node :- The top node of the tree, representing the entire dataset, which is split based on a feature that gives the best information gain or lowest Gini impurity.

2) Internal Nodes :- Each internal node represents a test or decision on a feature.

3) Leaf Nodes :- These are the final nodes that represent the predicted output or class.

• Example :-

consider a simple example of classifying whether a person will play tennis based on weather conditions.

Outlook	Temperature	Humidity	wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	strong	No
overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	cool	Normal	Weak	Yes
Rain	Cool	Normal	strong	No
overcast	Cool	Normal	strong	Yes
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes

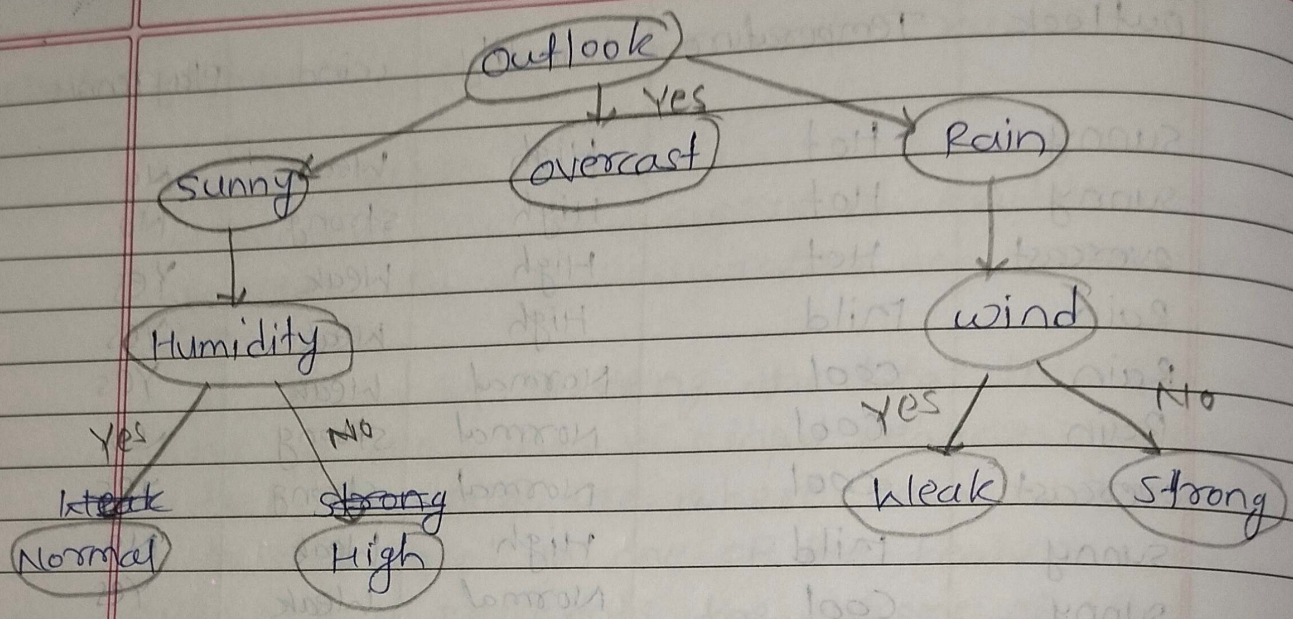
- Decision Tree construction :-

→ The algorithm chooses features that best split the data at each node using metric like Gini impurity or information Gain.

1) Root Node :- The first split is based on "outlook" feature because it provides the best information gain.

- if outlook = sunny, go to the next decision based on "Humidity".
  - if Humidity = High then playTennis = No
  - if Humidity = Normal, then playTennis = Yes

- if outlook = overcast, then playTennis = Yes
- if outlook = Rain, go to the next decision based on "wind".
  - if wind = weak, then playTennis = Yes
  - if wind = strong then playTennis = No



- Advantages :-
1) Easy to interpret and visualize
2) Handles both categorical and numerical data
3) can handle multi-output problems.

- Disadvantages :-
1) Prone to overfitting if the tree becomes too complex.
2) may not generalize well for unseen data if the tree is too deep

- Applications :-
1) Spam Detection
2) customer segmentation
3) Treatment Recommendation
4) credit card fraud
5) Loan Approval.

Que ⇒ Explain Linear Regression with example:

Linear Regression :-
- Linear Regression is a supervised machine learning algorithm used for predicting a continuous output variable based on one or more input features.
- It assumes a linear relationship between the input variables (independent variables) and the output variable (dependent variable).

• Formula :-

$$y = mx + b$$

where, y = dependent variable
x = independent variable
b = intercept
m = slope

• Assumptions of Linear Regression :-

1) Linearity :- The relationship between the features and target variable must be linear.
2) Independence :- The observation should be independent of each other.
3) Homoscedasticity :- The residuals (errors) should have constant variance.
4) Normality of Errors :- The errors should be normally distributed.

• Types of Linear Regression :-

1) simple Linear Regression :-
- It involves only one independent variable and one dependent variable.
equation :-

$$y = \beta_0 + \beta_1 x$$

2) multiple Linear Regression :-
- It involves more than one independent variable and one dependent variable.
equation :-

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

• Example :-
Pizza Prediction using Linear Regression

Diameter	Prize
8	10
10	13
12	16

Diameter (x)	Price (y)	mean (x)	mean (y)	Deviation (x)	Deviation (y)	Product of Deviation	Sum of Deviation	Square of Deviation (x)
8	10			2	3	6		4
10	13	10	13	0	0	0	12	0
12	16			−2	−3	6		4

By least square method

(1) Slope calculation

$$m = \frac{\text{sum of product of Deviation}}{\text{sum of square of Deviation}}$$

$$m = \frac{12}{8}$$

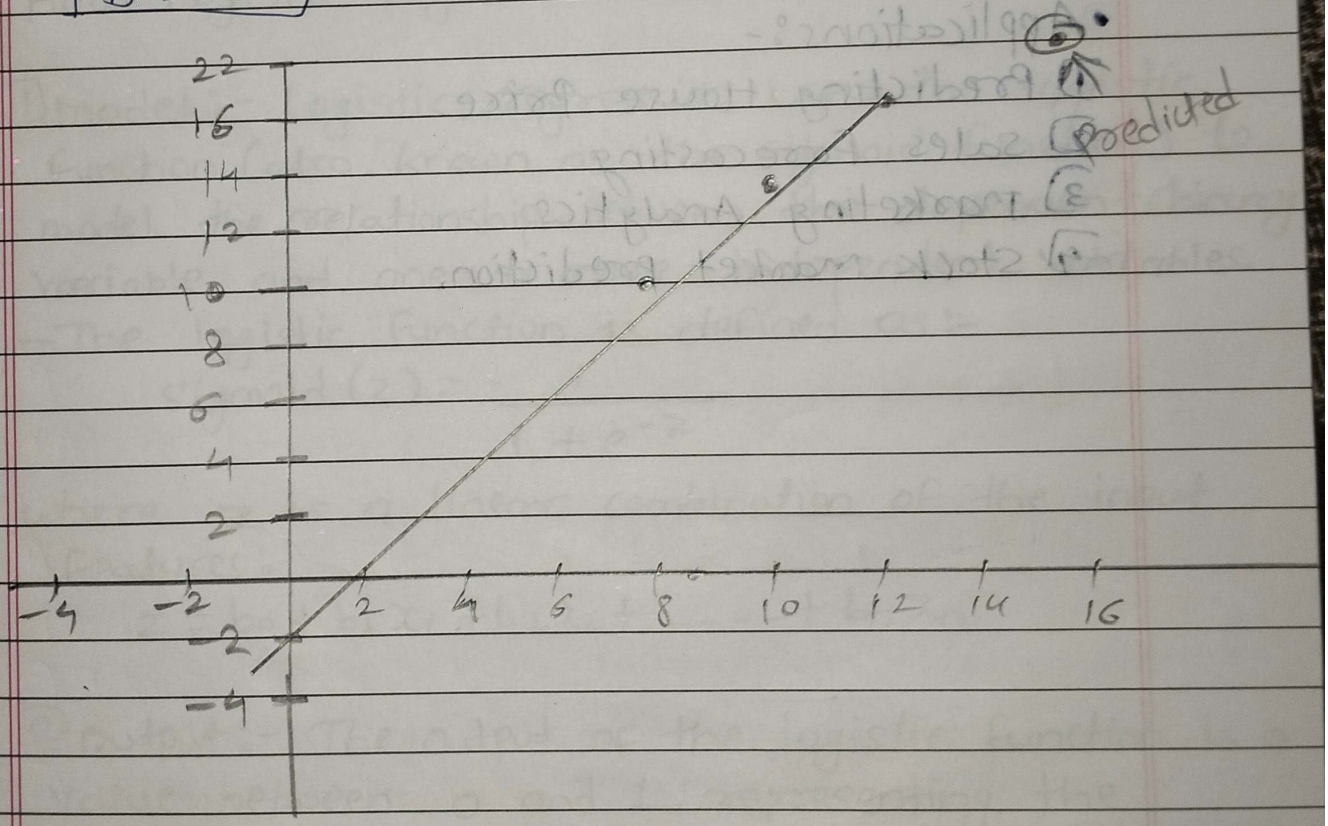
$$\boxed{m = 1.5}$$

(2) Intercept calculation:-

$$b = \text{mean of } y - [m * \text{mean of } x]$$

$$b = 13 - [1.5 * 10]$$

$$\boxed{b = -2}$$



(3) Prediction for new diameter Pizza $x = 16$

$y = ?$

$$y = mx + b$$

$$y = 1.5 \times 16 + (-2)$$

$$\boxed{y = 22}$$

- Advantages :-
1) simplicity
2) Efficiency
3) quick trainning
4) Low variance

- Disadvantages :-
1) Assumption of Linearity
2) Sensitive to outliers
3) overfitting in complex models
4) Multicollinearity.

- Applications :-
1) Predicting House Price
2) sales Forecasting
3) marketing Analytics
4) Stock market prediction.

Que- Explain Logistic Regression with example.

→

Logistic Regression :-
- Logistic Regression is a statistical method used for binary classification problems.
- Its used to Predict the probability of a binary outcome based on one or more predictor variables.
- logistic regression predicts a probility that is then mapped to a binary outcome (0 or 1).

How Logistic Regression works :-

1) model :- Logistic regression uses the logistic function (also known as the sigmoid function) to model the relationship between the dependent binary variable and one or more independent variables.
- The logistic function is defined as :-

$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

where z is a linear combination of the input features :-

$$z = b_0 + b_1 x_1 + b_2 x_2 + \text{----} + b_n x_n$$

2) output :- The output of the logistic function is a value between 0 and 1, representing the probability of the dependent variable being 1.

3) Threshold :- To make a classification decision, a threshold (0.5) is applied to this Probability.
- If the probability is greater than the threshold, the outcome is classified as 1; otherwise it is classified as 0.

$$P = \frac{1}{1+e^{-z}}$$

Example:-

Hours study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
3.9	1

1) Calculate the Probability of Pass for the student who studied 33 hours.
2) At least how many hours student should study that makes he will pass the course with Probability of more than 95 %

⇒

1)

$$z = -64 + 2 \times 33$$

$$\boxed{z = 2}$$

$$P = \frac{1}{1+e^{-z}}$$

$$P = \frac{1}{1+e^{-2}}$$

$$P = \frac{1}{1.135}$$

$$P = 0.88 = 88\ \%$$

∴ 33 hours studied 88 % Chance to Pass.

$\log(odds) = z = -64 + 2 * data$

2]

$$P = \frac{1}{1+e^{-z}}$$

$$0.95 = \frac{1}{1+e^{-z}}$$

$$0.95 * (1+e^{-z}) = 1$$

$$0.95 + 0.95\, e^{-z} = 1$$

$$0.95\, e^{-z} = 1 - 0.95$$

$$0.95\, e^{-z} = 0.05$$

$$e^{-z} = \frac{0.05}{0.95}$$

$$e^{-z} = 0.0526$$

$$e^{z} = \frac{1}{0.0526}$$

$$e^{z} = 19.0114$$

$$z = \ln(19.0114)$$

$$\boxed{z = 2.94}$$

$$\log(odd) = 2.94 = -64 + 2 * data$$

$$2.94 + 64 = 2 * data$$

$$data = \frac{2.94 + 64}{2}$$

$$\boxed{data = 33.4}$$

• Advantages :-
1) Simplicity
2) Efficiency
3) Interpretability
4) No Need for scaling.

• Disadvantages :-
1) Linearity Assumption
2) sensitivity to outliers
3) Limited to Binary classification
4) Requires Large sample size.

• Application :-
1) Medical Diagnosis
2) spam Detection
3) credit scoring
4) Voting Behavior.

Que- Explain Bayesian Linear Regression with example.

⇒

Bayesian Linear Regression:-

- Bayesian linear Regression is an extension of classical linear regression, where instead of finding point estimates for model parameters, we estimate their probability distributions.
- It is type of linear Regression that uses Bayesian statistic / theorem to estimate unknown parameters.
- The goal of Regression is to find the best estimation parameters of a linear model that describe relationship between dependent and independent variable.

Bayesian theorem

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Likelihood, Prior, Posterior, Marginal

- This theorem gives Relationship between prior Probability and posterior probability.

- When to use Bayesian theorem:-
  1) Data set is small
  2) Suitable for line based Learning
  3) It is try and test approach.

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

• Example :-

Student	marks	Grade
1	50	60
2	60	65
3	70	70
4	80	75
5	90	80

step-2 :-

$$y = \beta_0 + \beta_1 x$$

step-3 :-

$$\bar{x} = 70$$

$$\bar{y} = 70$$

step 4 :-

$$\beta_1 = \frac{(50-70)(60-70) + (60-70)(65-70) + (70-70)(70-70) + (80-70)(75-70) + (90-70)(80-70)}{(50-70)^2 + (60-70)^2 + (70-70)^2 + (80-70)^2 + (90-70)^2}$$

$$\beta_1 = \frac{500}{1000}$$

$$\boxed{\beta_1 = 0.5}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 70 - 0.5 \times 70$$

$$\beta_0 = 70 - 35$$

$$\boxed{\beta_0 = 35}$$

Predict Value For 85 marks :-

$$y = \beta_0 + \beta_1 x$$

$y = 35 + 0.5 \times 85$

$\boxed{y = 77.5}$

∴ if marks is 85 then Grade is 77.5

- Advantages :-
1) Incorporates prior information
2) uncertainty Estimation
3) Regularization
4) Handles small Datasets well.

- Disadvantage :-
1) computational complexity
2) choice of Prior
3) Slower for large Data.

- Applications :-
1) Risk Assessment
2) Decision making
3) Medical Research
4) Economics

Que- Explain Ridge Regression with example.

⇒

Ridge Regression :-
- Ridge regression is a type of regularized linear regression technique that addresses the issue of multicollinearity and overfitting by adding a penalty (regularization term) to the model.
- This penalty shrinks the regression coefficients, preventing them from becoming too large and making the model more generalizable.
- Ridge Regression adds a L2 penalty to the cost function.

• Why use Ridge Regression?

1) Multicollinearity :-
- When features are highly correlated the coefficients in ordinary linear regression can become unstable.
- Small changes in data can cause large changes in the coefficient estimates.
- Ridge regression reduces this sensitivity by imposing a penalty on the size of the coefficients.

2) Overfitting :-
- In high-dimensional datasets, linear regression tends to overfit, learning noise in the data.
- Ridge regression helps by constraining the model and reducing variance.

$$J(B_0 + B_1) = \sum_{i-1}^{n} (y_i - (B_0 + B_1 x_i))^2 + \lambda B_1^2$$

λB_1^2 ← Penalty term

Regularization

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2 + \lambda}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

- Example :-

Student	Marks	Grade
1	50	60
2	60	65
3	70	70
4	80	75
5	90	80

Step 2 :-

$$y = \beta_0 + \beta_1 x$$

Step 3 :-

$$\bar{x} = 70$$

$$\bar{y} = 70$$

$$\boxed{\lambda = 1}$$

Step 4 :-

$$\beta_1 = \frac{(50-70)(60-70) + (60-70)(65-70) + (70-70)(70-70) + (80-70)(75-70) + (90-70)(80-70)}{(50-70)^2 + (60-70)^2 + (70-70)^2 + (80-70)^2 + (90-70)^2}$$

$$\beta_1 = \frac{500}{1000 + 1} \rightarrow \lambda$$

$\beta_1 = 0.4995$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$

$\beta_0 = 70 - 0.4995\,(70)$

$\beta_0 = 35.0850$

step 5:- predict for 95 marks

$y = \beta_0 + \beta_1 x$

$y = 35.0350 + 0.4995\,(95)$

$y = 82.49$

∴ if marks 95 then grade 82.49

- Advantages:-
1) Reduce overfitting
2) Handles multicollinearity
3) Smoothes coefficients
4) works well with many Features

- Disadvantages:-
1) Does Not Perform Feature selection
2) Interpretability
3) Bias introduction

- Applications:-
1) Financial modeling
2) market research
3) Economic Forecasting
4) Disease prediction
5) Climate predictions.

que- Explain Model evaluation techniques.

⇒ For classification models :-

1) Accuracy :-
- The percentage of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{\text{Number of correct Predictions}}{\text{Total Number of predictions}}$$

2) Confusion matrix :-
- A matrix used to evaluate the performance of a classification algorithm by comparing actual vs predicted classes.

		Predicted	
		0	1
Actual	0	FN	FP
	1	TN	TP

3) Precision :-
- The proportion of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

4) Recall :-
- The proportion of correctly predicted positive observations to the all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

5) F1-score :-
- The harmonic mean of Precision and Recall.

$$F1\text{-score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}}$$

6) ROC curve :-
- A graphical representation of the performance of a classification model at all classification thresholds, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).

7) AUC (Area under the curve) :-
- The area under the ROC curve, providing a single scalar value to represent model performance.

• For Regression model :-

1) mean Absolute Error (MAE) :-
- The average of the absolute differences between the predicted and actual values.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

2) mean squared Error (MSE) :-
- The average of the squared differences between the predicted and actual values.

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

3] Root Mean squared Error (RMSE)
- The square root of MSE, bringing the error to the same scale as the original data.

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

4] R-squared :-
- Represents the proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where,
SS_{res} = sum of squares of residuals
SS_{tot} = total sum of squares.

5] Adjusted R-squared :-
- Adjused version of R^2 that penalizes the addition of irrelevant variables in the model

$$\text{Adjusted } R^2 = 1 - \left(\frac{1 - R^2}{n - k - 1}\right)$$

where,
n = number of observations
k = number of predictors.