Q1) a) ① Lasso regression stands for Least Absolute Shrinkage and Selection Operator

② It is a type of linear regression that includes a regularization term.

③ The regularization term is L1 penalty.

④ The goal of lasso regression is to minimum the sum of RSS i.e. Residual sum of squares with the absolute values of coefficients by tuning parameter λ

$$\text{minimize} \left(\sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where,

y_i = observed response
$\hat{y_i}$ = predicted response,
β_j = model coefficients
λ = regularization parameter.

⑤ Feature Selection in Lasso :-
Lasso performs automatic feature section by forcing coefficients of some features to be exactly zero.
The process is as :-
a) L1 penalty added to cost function encourages sparsity.

b) feature with coefficients shrinking to zero are removed model

c) This allows lasso to select only the most important features, making the model more efficient and reducing the multicolinearity.

Q1) b) ① Linear Regression :-
Models the relationship between a dependent variable and independent variable using a straight line.

② Ridge Regression :-
A variant of linear regression with L2 regularization to prevent overfitting

③ Lasso Regression :-
Similar to Ridge but uses L1 regularization, which shrink some zero and performing feature selection.

④ Elastic Net Regression:
Combine L1 and L2 by balancing feature selection.

(5) Polynomial Regression:
Extends linear regression by modelling non-linear relationships using polynomial terms.

(6) Logistic Regression:
Used for binary classification, models and outcomes of 0 or 1.

(7) KNN :-
K-Nearest Neighbors Regression. It predicts the value based on average of k nearest neighbors.

Q1)(c) (1) Bias Variance Trade off

(1) Bias :-
The error is by or comes by assuming a model that don't capture the underlying models of data.
high bias leads to overfitting

(2) Variance :-
The error is due to models sensitivity to the data. High variance leads to overfitting, where the model becomes too complex.

* Relationship to Underfitting & Overfitting

(1) Underfitting occurs when the model has high bias and poor performance

(2) Overfitting occurs when the model has high variance and gives good performance.

(3) a) (1) Kernel methods in SVM are used to transform non-linearly data into a higher-dimensional space where it becomes linearly separable.

(2) This is achieved without computing the coordinates of data in this higher-dimensional space.

* Common Kernel functions

(1) Linear kernel :-
No transformation is applied, and the data is assumed to be linearly separated.

(2) Polynomial Kernel :-
Maps the data into a higher-dimensional polynomial feature

(3) RBF
Maps the data infinite dimensional space

(3) The Kernel method allow SVM to learn non-linear decision boundaries by working in higher-dimensional space, without the need to compute the transformed data.

(4) This helps SVM to classify the complex data.

(Q3) 2) Advantages of KNN :-

(1) Simple and Easy :
It is easy to understand.

(2) No Training Phase :-
KNN is lazy learner it doesnot require a training phase

(3) Non-Parametric :-
It makes no assumptions about data distribution, useful for complex and non-linear data.

(4) Flexible :-
Can be used for both classification

disadvantages of KNN

(1) Computationally Expensive :-
The prediction phase is slow, it requires calculating distances to all training samples

(2) Sensitive to Irrelevant features :-
Performance can degrade with high dimensional or noisy data.

(3) Memory Intensive :-
Requires storing the dataset for prediction.

(4) Choosing K :-
The choice of the right value K can affect the performance

3.)c.) Matrics of KNN.

(1) Euclidean distance

The most widely used metric, measures the straight-line distance between two points.

(2) Manhattan distance

Measure the sum of absolute differences of coordinates useful in grid-like data.

(3) Minkowski Distance :-
Generalizes Euclidean and Manhattan distances by introducing a parameter p

(4) Cosine Similarity :-
measures the cosine similarity of data.

(5) Chebyshev Distance
Measures the maximum absolute difference along any dimension

Q6.)a) Isolation forest

(1) Isolation forest model is an anomaly selection algorithm that works by isolating the selected splits

(2) How it works :-

1) Isolation :-
The algo builds multiple random trees by selecting random features and splitting the data at random values.

2) Isolation of Anomalies :-
Anomalies are isolated in some steps because they are different

(3) Scoring:
The number of steps required to isolate a data point is used as a measure of its anomaly score.

(4) The advantages of Isolation forest are Efficient and No assumptions.

Q. 6 (1) Hierarchical Clustering is a method of cluster analysis that seeks to build a hierarchy of clusters :-

Two types :-

(1) Bottom - Up :-
Starts with each data point and merges the clusters.

(2) Top - Down :-
Starts with all points in a single cluster and splits into smaller clusters.

* The process of Clustering is :
(1) Start
(2) Merge
(3) Repeat.

Eg:-
Consider a dataset of 5 points A, B, C, D, E

Step 1:-
Start with individual
{A}, {B}, {C}, {D}, {E}.

Step 2:-
Merge the closest pair
{A, B}.

Step 3:-
calculate distance between
{A, B}, {C}, {D}, {E} and merge

Step 4:-
Repeat the process until one cluster remains.

Q.) (c) (i) Micro - Average Precision and Recall:-

(1) Micro - Avg Precision :-
Calculated by considering the total true positives, false positives and false negatives than calculate precision.

formula :- Micro precision = $\frac{\Sigma \text{ True positives}}{\Sigma (\text{True positives} + \text{false positives})}$

(2) Micro- Average Recall :-
Similer to precision, it aggregates true positives and false negatives

formula :- Micro Recall = $\frac{\Sigma \text{ True Positives}}{\Sigma (\text{True positives} + \text{false negatives})}$

(3) Micro- Average F-score.
(1) The harmonic mean of micro-average precision and recall, offering a single score that balance both metrics

formula :- Micro-F1 = $2 \times \frac{\text{Micro-Precision} \times \text{Micro Recall}}{\text{Micro-Precision} + \text{Micro Recall}}$

Q7.) a) (1) RNN stands for Recurrent Neural Networks.

(2) They are a type of neural network used in data

(3) They allow them to maintain a memory of previous inputs.

(4) This makes them particularly useful for tasks like time series prediction.

Eg:-
Predicting the Next Word in Sentence.

① Input sequence:-
"I love programming"

② The RNN takes "I" as input and processes it.

③ Then takes "love" as input, updates the state again and so on.

④ At every step, the RNN uses the memory from previous words to predict the next word

(Q7.) b.) The activation functions.

① Sigmoid:-
Output value between 0 and 1, it is a binary classification.

formula: $\sigma(x) = \frac{1}{1+e^{-x}}$

② Tan h:-
Output value between -1 and 1, a scaled version of sigmoid

formula :- $\tanh(x) = \frac{2}{1+e^{-2x}} - 1$

(3) ReLU (Rectified linear Unit)

Outputs the input if positive, otherwise zero, its mostly used in activation function.

formula :- $ReLU(x) = \max(0, x)$

(4) Leaky ReLU) :-

It allows a small or non-zero gradient for negative inputs.

formula :- Leaky ReLU $(x) = \max(\alpha x, x)$

(5) Softmax :-

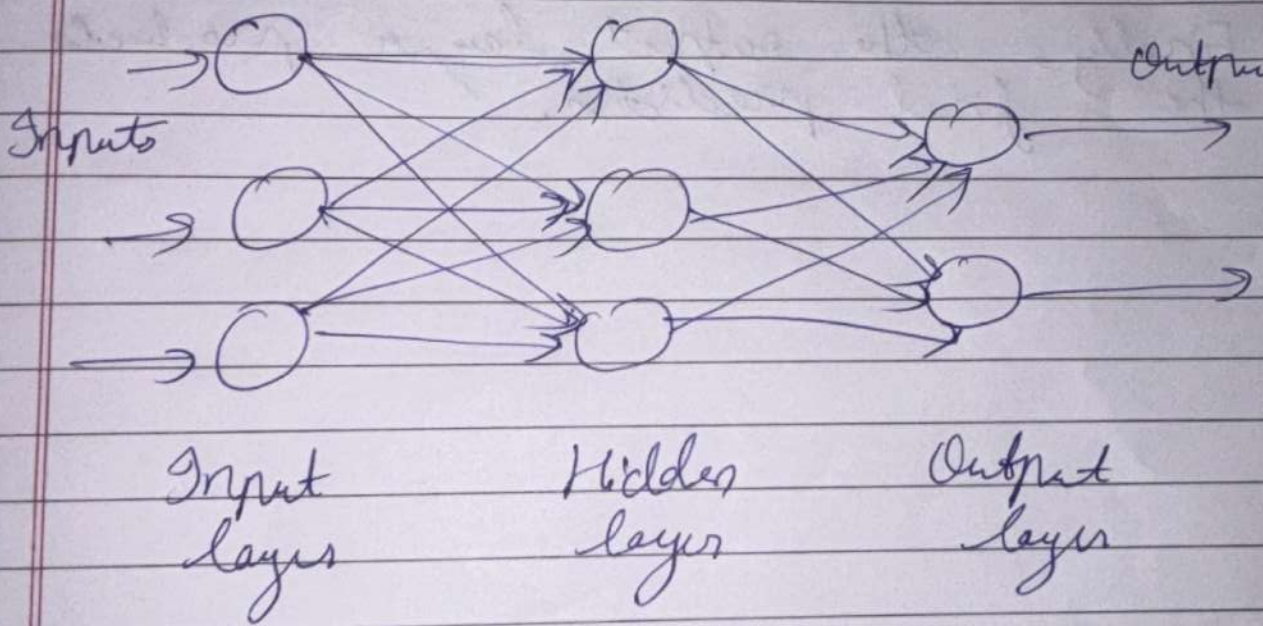
Used for multi-class classification, it converts logits into probability distributions over multiple classes

formula :- $\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Q7.) (1) MLP (Multilayer Perceptron)

(2) It is a type of feedforward neural network consisting of multiple layers of neurons.

(3) It has input layer, hidden layer, output layer



i) Structure of MLP:-

(1) Input layer:-
The first layer that get the input data.

(2) Hidden layer:-
One or more layer where the activation functions are performed

(3) The output layer :-
The produced output typically uses activation function. like softmax.

(4) How it works :-

(1) Data is passed from input layer to the first hidden layer then proce to the output layer.

(2) Finally, the output layer produces the final prediction.