Name → Eklavya Yadav

AI/ML Test 3.

Q.1. What is 'training set' and 'test set' in a Machine learning Model ? give examples.

Ans ① In machine learning Model, when we are going to prepare Model than we must train and test our dataset for further predictions.

Training Set → It is just like subset of Dataset on which our model will train. we always take larger part of Dataset in training data.

Test Set → It is also a small subset of Dataset which is smaller than training dataset to test our data / model and see the accuracy of the model i.e; whether the model is underfit or overfit or just a good model.

examples → Suppose we have any Dataset of an Employee salary with features like 'Age', 'Experience', 'Salary'. we can split our dataset into 60:40 ratio or 80:20 ratio in which larger part will always goes with training Dataset. and then we train model and test our model by applying Maching Learning algorithms.

Q.2 How missing and corrupted data is handled in dataset. give suitable example to justify your answer.

Ans (2) . So, if we have many missing and corrupted data than it is very difficult to prepare our Model and it will gives error while we run our program. So it is very necessary to handle those situations in our dataset and it is come under the "Data Analysis and Data cleaning pracess".

→ If our Data set has some missing values than we have two options two handling it :-

① Drop that missing values → It can be used when our Dataset is huge in numbers so that will not affect our model.
function used → . dropna () .

② fill missing value with Average / mean value with the taking whole column's / variable's mean → It can be done when our dataset is small in numbers.
function used → . fillna () .

eg → If we dataset like :-

x	y	z
3	3	10
4	7	12
5	N/A	13
6	8	14

we handled this like :
→ a = y.mean()
→ y.fillna(a) .

→ If we have corrupted data than we have to drop that whole data row/column.

→ These all steps come under "Data preprocessing". to ensure the quality of the data.

Q.3. What is difference b/w precision and recall?

Ans (2) Precision and recall are come under the metrics of the model for evaluate the performance of the prepared model in machine learning.
They are defined as based on the principals of TP (true positives), FP (false positive), FN (False Negatives).

$$\text{precision} = \frac{\text{True Positives}}{(\text{True positive}) + (\text{False positives})}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True positives} + \text{False Negatives}}$$

precision → precision focuses on the accuracy of the positive prediction made by model.

Recall → Recall focuses on the ability of the model to find all the positive instances.

Q.4 What are Support Vectors in SVM?

Ans (4) SVM stands for Support Vector Machine. Support vectors are the data points that lie nearest to the plane in SVM. the SVM goal is to find the best optimal plane that best separates the data into classes.

→ It one of the major work is reduce the Dimensionality of the features we are taking in large dataset and pick up the optimal / efficient features only for gaining better accuracy of the model.

Q.5 What is the significance of hue, size and style parameter in seaborn plot.

Ans (5) Seaborn is one of the most important module in python. Machine learning for visualization and it has many graphs with additional features in it to make it stand out from other modules.

It has many parameters for plotting a Dynamic graphs but we will talk about the only 3 three:

① hue → 'hue' parameter determines which column in data frame should be used for color encoding.

② size → 'size' parameter determines the which column in data frame should be used to set

the size of the data point in graphs.

(3) style → 'style' parameter determines which column in data frame should be used to set the style of the data points.

Q.6. Explain how colors are effectively used in data visualization?

(6) color is also a parameter for plotting graphs and it is one of the most important parameters for visualization.

→ It is a critical element in graphs that can greatly enhanced the clarity, impact and accessibility of the visualization. By carefully selecting and applying color schemes for highlighting key data points for better understanding to the viewers.

→ use color to encode data i.e, if we want to show a Dense data point than we use darker color and for less dense data points we can use light color.

→ we should avoid too many colors, it will create confusion to the viewers.

→ some examples of effective use of color:

(1) Heat map (2) Pie charts and Bar graphs

(3) Scatter plots and line graphs.

Q.7 → What are the characteristics of effective data visualization ?

Ans(7) Some effective way for Data Visualization are given under below :-

(1) Clarity → information should be easy to understand.

(2) Relevance → Avoid including irrelevant Sights.

(3) Accuracy → Ensure data is presented truthfully & accurately.

(4) Engagement → use engaging values to capture viewer's attention.

Q.8 Ans(8.) Recommendation System are the most important part of machine learning. In today's era each and every e-commerce or entertainment company website using recommendation system because it helps in their company growth and helps in to boost their profit.

Recommendation system → It is a system in which it recommends those products or data on which user or customer is frequently used it and go through those products.

for eg → (1) On flipkart, if we see any product then it automatically recommends and giving push notification about that product similar to that.

(8) on Netflix, if we watch horror movie than it recommends you another horror movie in your interface.

Q.9 (9) → Clustering is a unsupervised machine learning techniques used to group similar data points together based on certain features.

(1) Customer Segmentation in market → marketing department can create personalized marketing campaigns for each customer segment.

(2) Recommendation system in E-commerce → recommend product to a user based on what similar users have purchased or viewed

Q.10 (10) K determines the number of clusters in the dataset. The K-mean algorithm seeks to find the cluster centers that minimize the within-cluster sum of squares.
It is essential to choose an optimal K value to obtain meaningful cluster.
Data points move towards the centroids & forms the clusters.