1. What is ‘training Set’ and ‘test Set’ in a Machine Learning Model? Give examples.

The “training set” is used to train the model using some portion of the data from the available dataset. The “test set” is used to evaluate the model’s performance. The training set contains the input and output values to train the model regarding the relationship between the values provided.

For example, in a spam email classifier, the training set would contain emails labelled as spam or not spam along with their features. The test set on the other hand, is used to assess the performance of the model. It tells us how well the model generalizes to unseen data.

2. How missing and corrupted data is handled in dataset. Give suitable example to justify your answers.

Missing and corrupted data in a dataset are often handles through techniques such as data cleaning process. There are various methods offers in data cleaning part. Techniques such as finding null values, filling the missing values etc. Imputation is used to find out the missing values and fill those values with the help of remaining data by using the mean, mode, median etc. Deletion is also performed for removing the rows and columns with the missing or corrupted values. To exemplify, suppose let’s take a dataset of customer purchases, which has missing entries for certain transactions, here we can use the mean or median to fill in the missing values. But if the missing values are bigger in number or cannot do imputation, those rows and columns may be deleted to maintain data integrity.

3. What is the difference between precision and recall?

Precision and recall are both metrics used to evaluate the performance of the classification models. Precision calculates the accuracy of positive predictions, representing the ratio of true positive predictions to the total positive predictions generated by the model.

Recall measures the positive predictions by calculating the ratio of true positive predictions to the total actual positives in the data. For example, in a medical diagnosis system, precision measures the accuracy of correctly identifying patients

with a disease, while recall indicates how many of the actual classes were correctly identified.

4. What are Support Vectors in SVM?

Support vectors in SVM (Support Vector Machines) are critical data points that lies closest to the decision boundary between different classes. They define the margin of separation and play a major role in determining the optimal hyperplane that separates the classes. These support vectors involve in positioning and orientation of the decision boundary, which helps the SVM strong and effective in high-dimensional spaces and enables accurate classification in complex datasets.

5. What is the significance of hue, size, and style parameters in Seaborn plot.

In Seaborn plots, parameters like hue, size, and style enhances the visualization of the data. The hue parameter enables the user to provide different colors to different categories in the data. For example, if a user is plotting data about different types of fruits, user can use the hue to show each fruit with its own color. The size parameter allows to adjust the size of the plot elements, like dots or lines, mainly to highlight important parts of the data. The style parameter helps to change the appearance of elements like markers or lines based on the preference provided by the user. Overall, these parameters help a user to create cleaner and more engaging visualizations.

6. Explain how colors are effectively used in data Visualization.

Colors play a big role in making data visuals easy to understand and interesting. This is done by carefully selecting colors and employing appropriate color schemes, data visualizations can effectively highlight patterns, relationships, and insights within the data. For example, in a multiline graph providing different colors for each line which helps the user to understand the trend or patterns in it. By using colors carefully, user can make the data visualization much easier to understand.

7. What are the characteristics of effective data visualization?

Effective data visualization has several characteristics that enhances its communicative power and usefulness. Firstly, clarity ensures that the message conveyed by the visualization is easily understandable to the audience, avoiding ambiguity or confusion. Making the visualization simpler helps presenting

complex data in a clear and concise manner, avoiding unnecessary complexity helps not to distract from the main insights. Relevance ensures that the visualization focuses on what's important and provides insights that are useful to the audience. Interactivity allows users to play around with the data, so that they can explore it further and gain deeper understanding.

8. Explain by taking examples of e commerce or entertainment how recommendation system are working?

Recommendation systems in e-commerce or entertainment tracks user behavior and preferences to provide personalized recommendations. For example, in an e-commerce platform like Amazon uses past purchase history, browsing behavior, and other information to suggest product that match individual preferences and interests. Similarly, in entertainment platforms like Netflix uses viewing history, ratings and genre preferences to recommend movies or TV shows. These recommendation systems utilize various algorithms such as collaborative filtering, content-based filtering, or hybrid approaches to generate recommendations that are relevant and engaging, thereby enhancing user experience and satisfaction.

9. Give real world example for clustering.

A real-world example of clustering is customer segmentation in marketing. Companies often use clustering algorithms to group customers with similar characteristics or behavior together. For instance, a retail company may cluster customers based on factors such as purchase history, demographics, and browsing behavior to identify distinct customer segments such as budget shoppers or luxury buyers. Each cluster represents a group of customers with similar characteristics and behaviors. This segmentation enables targeted marketing strategies, personalized promotions, and tailored product recommendations, ultimately improving customer satisfaction. By effectively clustering customers, the company can improve customer satisfaction, increase their sales and so on.

10. How does the choice of 'k' impact the K-means algorithm?

The choice of the number of clusters, represented by 'k', significantly impacts the K-means algorithm's performance and clustering results. Selecting an appropriate 'k' value is crucial as it directly affects the interpretability and usefulness of the clustering solution. If 'k' is too small, clusters may be overly broad and fail to capture meaningful information in the data, leading to underfitting. On the other

side, if 'k' is too large, the algorithm may create unnecessary clusters, leading to overfitting and reducing the clarity and interpretability of the clustering solution. Therefore, selecting the optimal 'k' involves finding a balance between the simplicity of the model and its ability to accurately represent the underlying structure of the data.