1)What is the purpose of grid search cv in machine learning, and how does it work?

Ans- GridSearchCV is a technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.

Grid search is a process that searches exhaustively through a manually specified subset of the hyperparameter space of the targeted algorithm. Random search, on the other hand, selects a value for each hyperparameter independently using a probability distribution.

2) Describe the difference between grid search cv and randomize search cv, and when might you choose one over the other?

Ans- The only difference between both the approaches is in grid search we define the combinations and do training of the model whereas in RandomizedSearchCV the model selects the combinations randomly. Both are very effective ways of tuning the parameters that increase the model generalizability.

In Grid Search, we try every combination of a preset list of values of the hyper-parameters and choose the best combination based on the cross-validation score. Random search tries random combinations of a range of values (we have to define the number iterations).

3) What is data leakage, and why is it a problem in machine learning? Provide an example.

Ans- Data leakage (or leakage) happens when your training data contains information about the target, but similar data will not be available when the model is used for prediction. This leads to high performance on the training set (and possibly even the validation data), but the model will perform poorly in production.

Data leakage is one of the major problems in machine learning which occurs when the data that we are using to train an ML algorithm has the information the model is trying to predict. It is a situation that causes unpredictable and bad prediction outcomes after model deployment.

4) How can you prevent data leakage when building a machine learning model?

Ans- One of the best ways to get rid of data leakage is to perform k-fold cross validation where the overall data is divided into k parts. After dividing into k parts, we use each part as the cross-validation data and the remaining as training data.

5) What is a confusion matrix, and what does it tell you about the performance of a classification model?

Ans- A confusion matrix is a table that allows you to visualize the performance of a classification model. You can also use the information in it to calculate measures that can help you determine the usefulness of the model. Rows represent predicted classifications, while columns represent the true classes from the data.

6) Explain the difference between precision and recall in the context of a confusion matrix.

Ans- Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance.

7) How can you interpret a confusion matrix to determine which types of errors your model is making?

Ans- A confusion matrix is a table that allows you to visualize the performance of a classification model. You can also use the information in it to calculate measures that can help you determine the usefulness of the model. Rows represent predicted classifications, while columns represent the true classes from the data.

8) What are some common metrics that can be derived from a confusion matrix, and how are they calculated?

Ans- Confusion matrices can be used to calculate performance metrics for classification models. Of the many performance metrics used, the most common are accuracy, precision, recall, and F1 score.

The two main metrics that can be derived from the confusion matrix (aside from accuracy) are Precision and Recall.

1. You need a test dataset or a validation dataset with expected outcome values.
2. Make a prediction for each row in your test dataset.
3. From the expected outcomes and predictions count: The number of correct predictions for each class.

9) What is the relationship between the accuracy of a model and the values in its confusion matrix?

Ans- Here are some of the most common performance measures you can use from the confusion matrix. Accuracy: It gives you the overall accuracy of the model, meaning the fraction of the total samples that were correctly classified by the classifier. To calculate accuracy, use the following formula: (TP+TN)/(TP+TN+FP+FN)

10) How can you use a confusion matrix to identify potential biases or limitations in your machine learning model?

Ans- A confusion matrix is a table that allows you to visualize the performance of a classification model. You can also use the information in it to calculate measures that can help you determine the usefulness of the model. Rows represent predicted classifications, while columns represent the true classes from the data.