2 APRIL

Q1. Purpose of Grid Search CV:

Grid Search CrossValidation (Grid Search CV) is a technique used in machine learning to systematically search for the optimal hyperparameters of a model.

It works by defining a grid of hyperparameter values to explore, and it trains and evaluates the model with each combination using crossvalidation.

The purpose is to find the hyperparameters that result in the best model performance on the validation data, helping to tune and optimize the model.

Q2. Difference Between Grid Search CV and Randomized Search CV:

Grid Search CV: It exhaustively searches through all possible combinations of hyperparameter values within a predefined grid. It's suitable when you have a limited number of hyperparameters to search and computational resources.

Randomized Search CV: It randomly samples hyperparameter values from predefined distributions. It's suitable when you have a large hyperparameter space or limited computational resources. Randomized search may not explore every possible combination but can find good values more efficiently.

Choice between them: Use Grid Search when you have a reasonable number of hyperparameter combinations to explore and want to ensure exhaustive search. Choose Randomized Search when the hyperparameter space is vast, and you want to balance exploration and computational efficiency.

Q3. Data Leakage in Machine Learning:

Data leakage occurs when information from the validation or test data inadvertently influences the training process, leading to overly optimistic model performance.

Example: Using future information (e.g., target values) in the training set that would not be available at the time of prediction. For instance, including a stock price from the future to predict past stock prices.

Q4. Preventing Data Leakage:

Ensure strict separation of training, validation, and test datasets.

Avoid using future information, and be cautious with feature engineering and transformations.

When using crossvalidation, make sure that preprocessing steps (e.g., scaling) are applied within each fold.

Q5. Confusion Matrix:

A confusion matrix is a table used to evaluate the performance of a classification model. It summarizes the counts of true positives, true negatives, false positives, and false negatives.

Q6. Precision and Recall:

Precision is the ratio of true positives to the total predicted positives. It measures the model's ability to avoid false positives.
Recall (Sensitivity or True Positive Rate) is the ratio of true positives to the total actual positives. It measures the model's ability to find all relevant instances.

Q7. Interpreting a Confusion Matrix:

Analyzing the confusion matrix allows you to identify which types of errors your model is making. For example, you can see if the model is frequently misclassifying a specific class or if it has a problem with false positives.

Q8. Common Metrics from a Confusion Matrix:

Accuracy: (TP + TN) / (TP + TN + FP + FN)
Precision: TP / (TP + FP)
Recall (Sensitivity): TP / (TP + FN)
F1score: 2 * (Precision * Recall) / (Precision + Recall)

Q9. Relationship Between Accuracy and Confusion Matrix:

Accuracy is the overall correctness of the model, but it doesn't provide insights into the type of errors made. Values in the confusion matrix (TP, TN, FP, FN) are used to calculate accuracy.

Q10. Using Confusion Matrix for Bias and Limitation Analysis:

By examining the confusion matrix for each class, you can identify if the model has biases or limitations. For example, disproportionate false negatives in one class may indicate bias against that class. It helps in understanding model behavior and potential areas of improvement.