Akash Kamerkar

Must-Known 20 Statistics Medium level Questions for Data science Interview Preparation : A Freshers' Guide

Akash Kamerkar

1.What's the difference between Likelihood and Probability?

Likelihood refers to the probability of observing the data given a specific parameter value, while probability refers to the likelihood of an event occurring.

Akash Kamerkar

2.What is the difference between Type I and Type II errors in hypothesis testing?

Type I error occurs when we reject the null hypothesis when it is true, while Type II error occurs when we fail to reject the null hypothesis when it is false.

Akash Kamerkar

3.What is the bias-variance trade-off in machine learning?

The bias-variance trade-off refers to the trade-off between the model's ability to fit the training data well (low bias) and its ability to generalize to new, unseen data (low variance).

Akash Kamerkar

4.Explain the concept of multicollinearity in regression analysis.

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can lead to unstable or unreliable coefficient estimates.

Akash Kamerkar

5.What is the difference between parametric and non-parametric statistics?

Parametric statistics make assumptions about the population distribution, while non-parametric statistics do not rely on specific distributional assumptions.

Akash Kamerkar

6.Explain the concept of overfitting in machine learning.

Overfitting occurs when a model performs well on the training data but fails to generalize to new, unseen data. It usually happens when a model is too complex and captures noise or random fluctuations in the training data.

Akash Kamerkar

7.What is the purpose of feature scaling in machine learning?

Feature scaling is used to bring different features or variables onto a similar scale to prevent certain features from dominating others and to ensure fair comparisons during model training.

Akash Kamerkar

8.What is the difference between stratified sampling and cluster sampling?

Stratified sampling involves dividing the population into homogeneous groups and then randomly selecting samples from each group, while cluster sampling involves dividing the population into heterogeneous groups (clusters) and randomly selecting entire clusters for sampling.

Akash Kamerkar

9.What is the difference between precision and recall in binary classification?

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances among all actual positive instances.

Akash Kamerkar

10.What is the purpose of regularization in machine learning models?

Regularization is used to prevent overfitting by adding a penalty term to the loss function, which encourages the model to have smaller parameter values and simpler representations.

Akash Kamerkar

11.Explain the concept of feature importance in a machine learning model.

Feature importance refers to the relative importance or contribution of each feature in a model's predictions. It helps identify the most influential features and understand their impact on the model's performance.

Akash Kamerkar

12.What is the difference between unsupervised and supervised learning?

In unsupervised learning, the model learns patterns and relationships in the data without explicit target labels, while in supervised learning, the model is trained using labeled data to predict or classify new instances.

Akash Kamerkar

13.What is the purpose of cross-validation in model evaluation?

Cross-validation is used to assess the performance and generalization ability of a model by splitting the data into multiple subsets, training the model on some subsets, and evaluating it on the remaining subsets.

Akash Kamerkar

14.Explain the concept of the bias-variance decomposition in machine learning.

The bias-variance decomposition decomposes the expected prediction error of a model into bias, variance, and irreducible error components, providing insights into the model's overall error and its sources.

Akash Kamerkar

15.What is the difference between precision and accuracy in classification metrics?

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, while accuracy measures the proportion of correctly predicted instances (both positive and negative) among all instances.

Akash Kamerkar

16.What is the purpose of the Receiver Operating Characteristic (ROC) curve?

The ROC curve is used to evaluate the performance of a binary classifier by plotting the true positive rate against the false positive rate at various classification thresholds.

Akash Kamerkar

17.Explain the concept of the curse of dimensionality in machine learning.

The curse of dimensionality refers to the challenges and issues that arise when working with high-dimensional data, such as increased computational complexity, sparsity of data, and the need for more data to maintain model performance.

Akash Kamerkar

18.What is the difference between a random forest and a gradient boosting model?

Random forest is an ensemble model that combines multiple decision trees with random feature selection, while gradient boosting builds an ensemble model iteratively by focusing on the samples with higher prediction errors.

Akash Kamerkar

19.Explain the concept of hypothesis testing using p-values.

Hypothesis testing involves making inferences about a population based on sample data. The p-value measures the strength of evidence against the null hypothesis and helps determine whether the observed result is statistically significant.

Akash Kamerkar

20.What is the difference between a chi-square test and a t-test?

A chi-square test is used to compare observed frequencies with expected frequencies in categorical data, while a t-test is used to compare means between two groups in numerical data.