

Data Science Interview Questions

Keywords used: data science, machine learning

It is impossible to ignore the importance of data and one's capacity to analyze, consolidate, and contextualize it. Data Science are relied upon to fill this need but is a serious lack of qualified candidates worldwide.



If one is moving down the path to become a data scientist, one must impress prospective employers with knowledge. In addition to explaining why data science is so important, one need to show technical proficiency with good concepts, framework, and applications.

Here's a list of the most popular data science interview questions that one can expect.

1Q: What is data science? List the differences between supervised and unsupervised learning.

Data science is a blend of various tools, algorithms, and machine learning principles with a goal to discover hidden patterns from the raw data.

Supervised learning uses trained data set whereas unsupervised uses the input data set. Supervised is used for prediction whereas unsupervised is used for analysis.

2Q: What is Selection Bias

It is a kind of error that occurs when the researcher decides who is going to be studies. It is usually associated with research where selection of participants isn't random.

It sometimes referred to as the selection effect. It is distortion of statistical analysis, resulting from the method of collecting samples.

3Q:What is bias-variance trade-off?

Bias: Bias is an error introduction in model due to oversampling of the machine learning algorithms. It can lead to underfitting. There are both low bias machine learning and machine learning.

Variance: Variance is error introduced in models due to complex machine learning algorithms, model learns noise also from the training data set and performs badly on test data set.

4Q: What is a confusion matrix?

The confusion matrix is a 2x2 table that contains 4 outputs provided by the binary classifier. Various measures, such as error rate, accuracy, specificity, sensitivity, precision, and recall are derived from it.

A data set used for performance evaluation is called a test data set. It should contain the correct labels and predicted label. The prediction labels will exactly the same if the performance of a binary classifier is perfect.

5Q: What is the difference between “long” and “wide” format data?

In the wide format, a subject's repeated responses will be in a single row, and each response is in a separate column. In the long format, each row is a one-time point per subject.

One can recognize data in wide format by the fact that columns generally represent groups.

6Q: Define normal distribution.

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution.

7Q: What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both correlation and covariance establish the relationship and also measure the dependency between two random variables.

8Q: what are the goals of A/B testing?

It is a hypothesis testing for a randomized experiment with two variables A and B. The goal of A/B testing is to identify any changes to the web page to maximize or increase the outcome of interest.

A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business.



Summing it Up!

Hope this set of Data Science interview Question and Answers will help one in preparing for the interviews.

