**Data Science Interview Questions: Basic to Technical**

1. What is the difference between supervised and unsupervised learning?
2. How is logistic regression done?
3. Explain the steps in making a decision tree?
4. How do you build a random forest model?
5. How can you avoid overfitting your model?
6. Difference between univariate, bivariate, and multivariate analysis?
7. What feature selection methods are used to select the right variable?
8. In your choice of language, write a program that prints the number ranging from 1 to 50.
9. You are given a dataset consisting of variables or more than 30 % missing values. How will you deal with them?
10. For the given points, how will you calculate the Euclidean distance in Python?
11. What are dimensionality reduction and its benefits?
12. How will you calculate eigenvalues and eigenvectors of the following 3\*3 matrix?
13. How should you maintain a deployed model?
14. What are recommender systems?
15. How do you find RMSE and MSE in a linear regression model?
16. How can you select k for k-means?
17. What is the significance of the p-value?
18. How can outlier values be treated?
19. How can time series data be declared stationary?
20. How can you calculate accuracy using a confusion matrix?
21. Write an equation and calculate the precision and recall rate?
22. People who bought this also bought recommendations seen on Amazon are a result of which algorithm?
23. Write a basic SQL query that lists all the orders with customer information?
24. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 per cent. Why shouldn’t you be happy with your model’s performance? What can you do about it?
25. Which of the following machine learning algorithms can be used for imputing missing values of both categorical and continuous variables?
26. We want to predict the probability of death from heart diseases based on three risk factors: age, gender and blood cholesterol level. What is the most appropriate algorithm for this case?
27. After studying the behaviour of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each type. Which algorithm is most appropriate for this study?
28. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true?
29. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should be used?
30. What do you understand about true and false positive rates?
31. What is the ROC curve?
32. What are the feature vectors?
33. What are the steps in making a decision tree?
34. What is root cause analysis?
35. What is logistic regression?
36. What are recommendation systems?
37. Explain cross-validation?
38. What is collaborative filtering?
39. Do gradient descent methods always converge to similar points?
40. What is the goal of A/B testing?
41. What are the drawbacks of the linear model?
42. What is the law of large numbers?
43. What are the confounding variables?
44. What is star schema?
45. How regularly must an algorithm be updated?
46. What are eigenvalue and eigenvector?
47. Why is resampling done?
48. What is selection bias?
49. What are the types of biases that can occur during sampling?