Company Name - Wipro
Role - Data Scientist

Q. What is the main Difference between where and having in SQL?

A. The main difference between them is that the WHERE clause is used to specify a condition for filtering records before any groupings are made, while the HAVING clause is used to specify a condition for filtering values from a group.

Q. Explain confusion matrix?

A. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

Q. Explain PCA?

A. The principal components are eigenvectors of the data's covariance matrix. Thus, the principal components are often computed by eigen decomposition of the data covariance matrix or singular value decomposition of the data matrix. PCA is the simplest of the true eigenvector-based multivariate analyses and is closely related to factor analysis.

Q. How do you cut a cake into 8 equal parts using only 3 straight cuts?

A. Cut the cake from middle first, then pile up the one piece on another, and then again cut it straight from the middle which will leave you with 4 pieces. Finally, put all the 4 pieces on one another, and cut it for the third time. This is how with 3 straight cuts, you can cut cake into 8 equal pieces.

Q. Explain k-means clustering?

A. K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart. Similarity of two points is determined by the distance between them.

Q. How is KNN different from k-means clustering?

A. K-means clustering represents an unsupervised algorithm, mainly used for clustering, while KNN is a supervised learning algorithm used for classification.

Q. Stock market prediction: You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). Would you treat this as a classification or a regression problem?

A. It is a classification problem.

Q. What graphs do you use for basic EDA?

A. Scatter Plot, Pair plots, Histogram, Box plots, Violin Plots, Contour plots etc.

Q. Equation to calculate the precision and recall rate?

A. Precision = True positives/ (True positives + False positives) = TP/ (TP + FP).

Recall = TruePositives / (TruePositives + FalseNegatives) = TP / (TP + FN).

Q. When do we use bagging and boosting?

A. Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.

Q. Difference between RNN and CNN?

A. CNN is considered to be more powerful than RNN .

Rnn includes less feature compatibility when compared to Cnn.

Cnn stands for Convolutional neural network while Rnn is Recurrent neural netwo

Q. What is Descriptive analytics?

A. Descriptive analytics is the interpretation of historical data to better understand changes that have occurred in a business. Descriptive analytics describes the use of a range of historic data to draw comparisons .

Q.What exactly is Central limit theorem?

A. the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

Q. What is residual plot?

A. A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.

Q. What is z-score?

A. A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.

Q. Image classification algorithms?

A. Image Classification algorithms are the algorithms which are used to classify labels for images using their characteristics. Example: Convolutional Neural Networks.

Q. How do you handle categorical data?

A. One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

Q. Scaling? It's advantage?

A. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

NOTE:
If you want to learn Data Science from scratch and become interview ready, check out our popular Data science course.

https://ai.cloudyml.com/Learn-Data-Science-from-Scratch