Company Name - Atos
Role - Data Scientist

Q. Choose algorithm if your data has noise?
A. Algorithms like Probabilistic Random Forest and DBSCAN can work well when you've noise in data.

Q. Is it fine if we don't do scaling before training model? Reason?
A. If the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.
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.

Q. Tricks you use to faster your model training?
A. Reduce Calculations by normalization or standardization. Use high computational memory while training(like GPU instead of local memory).

Q. What is SelectK best? how does it works?
A. The SelectKBest method selects the features according to the k highest score. By changing the 'score_func' parameter we can apply the method for both classification and regression data. Selecting best features is important process when we prepare a large dataset for training. The SelectKBest class just scores the features using a function (in this case f_classif but could be others) and then "removes all but the k highest scoring features".

Q. How will you handle heavy data?
A. 1. Allocate more memory 2. Work with smaller sample 3. Use a relational database like SQL 4. Use a big data platform such as Hadoop.

Q. What is the loss function SVM tries to minimize?
A. Although there is no "loss function" for hard-margin SVMs, the loss does exist when solving soft-margin SVMs. The hinge loss is a loss function used in machine learning to train classifiers. For "maximum-margin" classification, the hinge loss is utilised, most notably for support vector machines (SVMs).

Q. Detect heteroscedasticity in simple linear regression?
A. Heteroscedasticity refers to the situation where the spread of the residual's changes in a systematic way over the range of observed values. A fitted value vs.

residual plot is the simplest technique to determine heteroscedasticity. The "cone" form is a clear marker of heteroscedasticity if the residuals become significantly more spread out as the fitted values get greater. The Breusch-Pagan test is a more formal, mathematical method of determining heteroskedasticity.

Q. Explain ANOVA?

A. The analysis of variance (ANOVA) is a statistical technique for determining if the means of two or more groups differ significantly. One-way ANOVA, two-way ANOVA, and multivariate ANOVA are the three types. An ANOVA's null hypothesis is that there is no significant difference between the groups. The alternative hypothesis proposes that the groups have at least one substantial difference. The null hypothesis is rejected and the alternative hypothesis is validated if the p-value associated with the F is less than.05. If the null hypothesis is rejected, one concludes that the means of all the groups are not equal.

Q. Determine no. of neighbors in KNN?

A. The number of neighbors(K) in KNN is a hyperparameter that must be chosen during model construction. According to research, there is no ideal number of neighbors for all types of data sets. A small number of neighbors is the most flexible fit, resulting in low bias but high variation, whereas a big number of neighbors results in a smoother decision boundary, resulting in reduced variance but higher bias. If the number of classes is even, data scientists usually choose an odd number. You can also test the model's performance by creating it with different values of k and comparing the results. Elbow technique is another option.

Q. What do you mean by central trend?

A. The central trend is a description of a dataset represented by a single value that represents the data distribution's centre. The following measurements can be used to describe the central tendency of a dataset. The sum of all values in a dataset divided by the total number of values is the mean. The middle value in an ascending-ordered dataset is called the median. The most often occurring value in a dataset is

defined by the mode. Despite the fact that the measures listed above are the most generally employed to describe central tendency, there are others, such as geometric mean, harmonic mean, midrange, and geometric median.

Q. Is it possible to capture the correlation between continuous and categorical variable? How?

A. To capture the relationship between continuous and categorical data, we can utilise the ANCOVA (analysis of covariance) technique. It works in a similar way to factorial ANOVA in that it can tell you how much more information you can gain by focusing on one independent variable (factor) at a time and ignoring the others.

Q. What is Central Limit Theorem and why is it important?

A. Even if the data within each sample is not normally distributed, the Central Limit Theorem states that as sample sizes grow greater, the sampling distribution of the mean will become normally distributed. The Central Limit Theorem is significant in statistics because it allows us to assume that the mean sample distribution will be normal in the vast majority of circumstances. This means we can use statistical techniques based on the assumption of a normal distribution.

Q. What function of numpy will you use to find maximum value from each row in a 2D numpy array?

A. The amax function of numpy can be used to find the maximum value from each row in a 2D array. The axis of the array can be specified to get maximum number from each row.

Q. Consider a (5,6,7) shape array, what is the index (x,y,z) of the 50th element?

A. The answer is (1,1,1). This can be found out using the unravel_index function of numpy. Use "print (np.unravel_index(50, (5,6,7)))"

Q. How is standard deviation affected by the outliers?

A. Outliers will affect Standard Deviation since standard deviation is determined by subtracting the sample case from the mean. The presence of outlier changes the mean of the data and hence affect the entire standard deviation and makes the distribution skewed. Because the number of outliers in a sample is expected to be unusual, this effect diminishes as sample size grows.
