1. **What is the advantage of hierarchical clustering over K-means clustering**?

   Ans: B) In hierarchical clustering you don't need to assign number of clusters in beginning

2. **Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?**

   Ans: A) max_depth

3. **Which of the following is the least preferable resampling method in handling imbalance datasets?**

   Ans: B) RandomOverSampler

4. **Which of the following statements is/are true about "Type-1" and "Type-2" errors?**

   Ans: C) 1 and 3

5. **Arrange the steps of k-means algorithm in the order in which they occur:**
   1. Randomly selecting the cluster centroids
   2. Updating the cluster centroids iteratively
   3. Assigning the cluster points to their nearest center

   Ans: D) 1-3-2

6. **Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?**

   Ans: B) Support Vector Machines

7. **What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?**

   Ans: C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

Ans: A) Ridge will lead to some of the coefficients to be very close to 0

D) Lasso will cause some of the coefficients to become 0.

9. Which of the following methods can be used to treat two multi-collinear features?

Ans: B) remove only one of the features
D) use Lasso regularization

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

Ans: A) Overfitting
B) Multicollinearity

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Ans :when the datasets are quite high and the categorical features present in the datasets are ordinal, one hot encoded must be avoided, as it can lead to high memory consumption. We can use ordinal encoder for this process, where the ordering sequence is unique, that can be defined and passed into the encoders.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Ans: When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling.

13. What is the difference between SMOTE and ADASYN sampling techniques?

SMOTE: Synthetic Minority Oversampling Technique (SMOTE) is **a statistical technique for increasing the number of cases in your dataset in a balanced way**. The component works by generating new instances from existing minority cases that you supply as input.

ADASYN (Adaptive Synthetic) is **an algorithm that generates synthetic data**, and its greatest advantages are not copying the same minority data, and generating more data for “harder to learn” examples

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible. One can shift to Random Search CV where the algorithm will randomly choose the combination of parameters.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square.
2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
3. Mean Absolute Error(MAE)

R square: It is statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable R-squared shows how well the data fit the regression model (the goodness of fit).

Adjusted R square: It is a corrected goodness-of-fit (model accuracy) measure for linear models

Mean Squared Error: The Mean Squared Error measures how close a regression line is to a set of data points. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

Root Mean square Error: It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).

Mean Absolute Error: Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.