MACHINE LEARNING

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) Random Over Sampler

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:
   Ans: A) 3-1-2

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: B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of 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: B) Lasso will lead to some of the coefficients to be very close to 0

9. Which of the following methods can be used to treat two multi-collinear features?
   Ans: B) remove only one of the features

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

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?
    Ans: One-hot encoding may not be the best choice when dealing with high cardinality categorical features. High cardinality features are those with a large number of unique categories. One-hot encoding can result in a large number of columns in the dataset, which can lead to the curse of dimensionality and cause computational issues.

In such cases, other encoding techniques such as target encoding, frequency encoding, or binary encoding can be used. These techniques can help to reduce the number of columns in the dataset while still capturing the information in the categorical feature.
Target encoding involves replacing each category with the mean of the target variable for that category. Frequency encoding replaces each category with its frequency in the dataset. Binary encoding converts each category into a binary code of 0s and 1s, with each digit representing the presence or absence of a particular category.
The choice of encoding technique will depend on the specific dataset and the nature of the categorical features. It is important to evaluate the performance of different encoding techniques and choose the one that works best for the particular problem at hand.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.
    Ans: When dealing with data imbalance in classification, where one or more classes have significantly fewer instances than others, the performance of the model can be affected. In such cases, various techniques can be used to balance the dataset and improve the performance of the classification model. Here are some commonly used techniques:

1. Resampling: Resampling involves either oversampling the minority class or undersampling the majority class. Oversampling can be done by duplicating instances of the minority class or by generating synthetic data using techniques such as SMOTE (Synthetic Minority Over-sampling Technique). Undersampling involves randomly removing instances from the majority class. Resampling can help to balance the dataset and ensure that the model has enough instances to learn from.
2. Class weighting: Class weighting involves assigning higher weights to the minority class and lower weights to the majority class. This can be done during model training to ensure that the model gives more importance to the minority class.
3. Ensemble methods: Ensemble methods such as bagging, boosting, and stacking can be used to improve the performance of the classification model. Ensemble methods combine multiple models to make predictions and can help to balance the dataset by giving more weight to the minority class.

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

Ans: SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are two popular techniques used for oversampling imbalanced datasets. The main difference between the two techniques is in how they generate synthetic samples.

SMOTE generates synthetic samples by interpolating between existing minority class samples, whereas ADASYN generates synthetic samples by creating new samples near the decision boundary of the minority class.

Specifically, SMOTE works by selecting a minority class sample, selecting one or more nearest neighbors of the same class, and then creating synthetic samples by linearly interpolating between the selected sample and its nearest neighbors. This process is repeated until the desired level of oversampling is achieved.

ADASYN, on the other hand, is an adaptive algorithm that generates synthetic samples based on the density distribution of the minority class near the decision boundary. The algorithm first computes

the density distribution of the minority class samples and then generates synthetic samples for each minority class sample based on the local density of the minority class. The algorithm gives more importance to regions with fewer minority class samples and generates more synthetic samples in those regions.

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?
    Ans: GridSearchCV is a function in scikit-learn that performs an exhaustive search over a specified parameter grid to find the best combination of hyperparameters for a machine learning model. The purpose of using GridSearchCV is to automate the process of hyperparameter tuning, which is an important step in building a well-performing model. GridSearchCV allows the user to define a parameter grid, which is a set of hyperparameters and their possible values. The function then trains and evaluates the model with each combination of hyperparameters in the grid and returns the combination that gives the best performance on a specified evaluation metric.
    GridSearchCV is a popular choice for hyperparameter tuning because it is simple to use and can quickly find the best combination of hyperparameters. However, it can be computationally expensive, especially when dealing with large datasets or complex models with many hyperparameters.
    In some cases, it may be necessary to use other techniques for hyperparameter tuning when dealing with large datasets. One option is to use randomized search, which randomly samples hyperparameters from a specified distribution, rather than searching over a fixed grid. This can be faster than GridSearchCV and can still find good hyperparameters. Another option is to use a Bayesian optimization approach, which iteratively builds a model of the hyperparameter space and evaluates the model to find the best hyperparameters.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief
    Ans: When evaluating a regression model, there are several metrics that can be used to measure the performance of the model. Here are some commonly used evaluation metrics:

1. Mean Squared Error (MSE): MSE is a widely used metric that measures the average squared difference between the predicted and actual values. It is calculated as the average of the squared residuals between the predicted and actual values. MSE penalizes large errors more heavily than small errors and is commonly used in linear regression.
2. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and is a popular metric for evaluating regression models. It is more interpretable than the MSE as it is in the same units as the target variable. RMSE is also commonly used in linear regression.
3. Mean Absolute Error (MAE): MAE is a metric that measures the average absolute difference between the predicted and actual values. It is calculated as the average of the absolute residuals between the predicted and actual values. MAE is less sensitive to outliers than MSE and RMSE and is commonly used in decision tree-based models.
4. R-squared (R2): R2 is a metric that measures the proportion of variance in the target variable that is explained by the model. It is calculated as the ratio of the explained variance to the total variance. R2 ranges from 0 to 1, with higher values indicating better performance.