MACHINE LEARNING ASSIGNMENT

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

Ans – B

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

   Ans – A

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

   Ans – A

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

   Ans – B

5. Arrange the steps of k-means algorithm in the order in which they occur:

   Ans – D

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

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

   Ans – B

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 and D

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

   Ans – B and D

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 and C

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

Ans - When the categorical features present in the dataset are ordinal i.e., for the data being like Junior, Senior, Executive, Owner. When the number of categories in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption. 4B/5B encoding technique is used in this case. In this type of encoding, double speed

clocks are not required. Instead, 4 bits of codes are mapped to 5 bits; having a minimum of 1-bit in the group.

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

Ans - Resampling Oversampling and Undersampling 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?

Ans - The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed.

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 technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. More Data = More Features The first and perhaps most obvious way in which more data delivers better results in data science is the ability to expose more features to feed your data, science models. In this case, accessing and using more data assets can lead to “wider datasets” containing more variables.

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

Ans - There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are:

1 Mean Squared Error (MSE).

2 Root Mean Squared Error (RMSE).

3 Mean Absolute Error (MAE)

1. The MSE is a measure of the quality of an estimator. As it is derived from the square of Euclidean distance, it is always a positive value that decreases as the error approaches zero.

2. Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

3. MAE evaluates the absolute distance of the observations (the entries of the dataset) to the predictions on a regression, taking the average over all observations.