MACHINE LEARNING
WORKSHEET

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

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

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?
Answer: 1 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur:
Answer: 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?
Answer: Support Vector Machines

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
Answer: 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?
Answer: A ,B & D

9. Which of the following methods can be used to treat two multi-collinear features?
Answer: B, C & 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?
Answer: Overfitting

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?
Answer : One-hot encoding must be avoided in situations where the categorical feature has high cardinality, i.e., a large number of unique values. One-hot encoding can result in a sparse matrix with a large number of columns, which can cause several issues, such as increased computational complexity, memory consumption, and overfitting. In extreme cases, one-hot encoding can also exhaust the available memory and crash the system.
An alternative encoding technique called Target Encoding can be used. Target encoding encodes the categorical variable using the mean target value for each category. Target encoding can provide a more compact representation of the categorical variable and can help capture the relationship between the categorical feature and the target variable.

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

Answer: Data imbalance problem occurs when one class in a binary or multi-class classification problem has a much smaller number of instances compared to the other classes. In such cases, the model may be biased towards the majority class and may not perform well in predicting the minority class.
To balance the dataset, the following techniques can be used:
Cost-sensitive learning: In this technique, the misclassification cost is adjusted to give more weight to the minority class instances. This technique can help improve the performance of the model on the minority class, but it may require a priori knowledge of the cost matrix.
Ensemble methods: In this technique, multiple models are trained on different subsets of the dataset and combined to make the final prediction. Ensemble methods, such as Bagging and Boosting, can help improve the generalization performance of the model and reduce overfitting, but they may require more computational resources.

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

Answer: SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are both sampling techniques commonly used in imbalanced learning problems, where the number of samples in one class is significantly lower than the other class.
The main difference between SMOTE and ADASYN is that ADASYN is an adaptive method that generates more synthetic samples for those minority class samples that are harder to learn, while SMOTE generates synthetic samples in a fixed proportion for all minority class samples.
In summary, while SMOTE generates synthetic samples in a fixed proportion for all minority class samples, ADASYN adaptively generates synthetic samples for minority class samples that are harder to learn

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

Answer: GridSearchCV is a technique used in machine learning to search for the optimal hyperparameters of a model. It is used to systematically evaluate a combination of hyperparameters to find the best configuration that produces the best performance of the model on the given data.
The purpose of using GridSearchCV is to automate the process of hyperparameter tuning and to ensure that the model is optimized for the given data. By searching through a specified set of hyperparameters, GridSearchCV can help identify the best combination of hyperparameters that result in the highest accuracy or other performance metric for a given problem. When it comes to large datasets, GridSearchCV can be computationally expensive and time-consuming. This is because it involves testing a combination of hyperparameters, which can result in a large number of model training and evaluation runs. As the size of the dataset increases, the computational cost of GridSearchCV can become prohibitively high.

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

Answer: Evaluation metrics are used to measure the performance of a regression model. Some of the commonly used evaluation metrics for regression models are:
Mean Squared Error (MSE): It is the average of the squared differences between the predicted and actual values. It measures how far off the predicted values are from the actual values.
Root Mean Squared Error (RMSE): It is the square root of the MSE. It is preferred over MSE as it has the same unit as the target variable and is easier to interpret.
Mean Absolute Error (MAE): It is the average of the absolute differences between the predicted and actual values. It measures the average magnitude of the errors in a set of predictions.