ML-WORKSHEET 8

In Q1 to Q7, only one option is correct, Choose the correct option:

1. What is the advantage of hierarchical clustering over K-means clustering?
   A) Hierarchical clustering is computationally less expensive
   B) In hierarchical clustering you don't need to assign number of clusters in beginning
   C) Both are equally proficient
   D) D) None of these

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
   A) max_depth
   B) B) n_estimators
   C) min_samples_leaf
   D) min_samples_splits

3. Which of the following is the least preferable resampling method in handling imbalance datasets?
   A) SMOTE
   B) B) RandomOverSampler
   C) RandomUnderSampler
   D) ADASYN

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?
   1. Type1 is known as false positive and Type2 is known as false negative.
   2. Type1 is known as false negative and Type2 is known as false positive.
   3. Type1 error occurs when we reject a null hypothesis when it is actually true.

   A) 1 and 2
   B) B) 1 only
   C) 1 and 3
   D) 2 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

   A) 3-1-2
   B) 2-1-3
   C) 3-2-1
   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?
   A) Decision Trees
   B) B) Support Vector Machines

C) K-Nearest Neighbors
D) Logistic Regression

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
   A) CART is used for classification, and CHAID is used for regression.
   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).
   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)
   D) None of the above

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?
   A) Ridge will lead to some of the coefficients to be very close to 0
   B) Lasso will lead to some of the coefficients to be very close to 0
   C) Ridge will cause some of the coefficients to become 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?
   A) remove both features from the dataset
   B) remove only one of the features
   C) Use ridge regularization
   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?
    A) Overfitting
    B) B) Multicollinearity
    C) Underfitting
    D) Outliers

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

Ans-One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.
One-Hot Encoding results in a Dummy Variable Trap as the outcome of one variable can easily be predicted with the help of the remaining variables.The Dummy Variable Trap leads to the problem known as multicollinearity. Multicollinearity occurs where there is a dependency between the independent features. Multicollinearity is a serious issue in machine learning models like Linear Regression and Logistic Regression. The categorical feature is ordinal (like Jr. kg, Sr. kg, Primary

school, high school) The number of categories is quite large as one-hot encoding can lead to high memory consumption Hence we use label encoding.

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

Answer-

Random Under-Sampling: Random Under sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out.

Random Over-Sampling: Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

Cluster-Based Over Sampling: K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.

Algorithmic Ensemble Techniques:

Bagging Based techniques for imbalanced data: Bagging is an abbreviation of Bootstrap Aggregating. The conventional bagging algorithm involves generating 'n' different bootstrap training samples with replacement. And training the algorithm on each bootstrapped algorithm separately and then aggregating the predictions at the end.

Boosting-Based techniques for imbalanced data: Boosting is an ensemble technique to combine weak learners to create a strong learner that can make accurate predictions. Boosting starts out with a base classifier / weak classifier that is prepared on the training data.

Adaptive Boosting- Ada Boost techniques for imbalanced data: Ada Boost is the first original boosting technique which creates a highly accurate prediction rule by combining many weak and inaccurate rules. Each classifier is serially trained with the goal of correctly classifying examples in every round that were incorrectly classified in the previous round.

Gradient Tree Boosting techniques for imbalanced data: In Gradient Boosting many models are trained sequentially. It is a numerical optimization algorithm where each model minimizes the loss function, y = ax+b+e, using the Gradient Descent Method.

XG Boost techniques for imbalanced data: XGBoost (Extreme Gradient Boosting) is an advanced and more efficient implementation of Gradient Boosting Algorithm discussed in the previous section.

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

Answer: 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 distributions.

SMOTE

1) Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space.
2) The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors.
3) The difference is multiplied by random number between (0, 1) and it is added back to feature.
4) SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

ADASYN

1) ADAptive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor.
2) The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data.
3) The algorithm uses Euclidean distance for KNN Algorithm. The latter generates the same number of synthetic samples for each original minority sample.

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

Answer: GridSearchCV is a function that comes in Scikit-learn's(or SK-learn) model_selection package. So an important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

We pass predefined values for hyperparameters to the GridSearchCV function. We do this by defining a dictionary in which we mention a particular hyperparameter along with the values it can take. Here is an example of it

```
{ 'C': [0.1, 1, 10, 100, 1000],
  'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
  'kernel': ['rbf','linear','sigmoid']}
```

Here C, gamma and kernels are some of the hyperparameters of an SVM model. Note that the rest of the hyperparameters will be set to their default values

GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

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

Answer: The various metrics used to evaluate are:

Mean Squared Error(MSE)
Root-Mean-Squared-Error(RMSE).
Mean-Absolute-Error(MAE).
R^2 or Coefficient of Determination.
Adjusted R^2

Mean Squared Error: MSE or Mean Squared Error is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model. As it squares the differences, it penalizes even a small error which leads to over-estimation of how bad the model is.

Root Mean Squared Error: RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors.

Mean Absolute Error: MAE is the absolute difference between the target value and the value predicted by the model. The MAE is more robust to outliers and does not penalize the errors as extremely as mse. MAE is a linear score which means all the individual differences are weighted equally. It is not suitable for applications where you want to pay more attention to the outliers.

R^2 Error: Coefficient of Determination or R^2 is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R^2 is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R^2 will always be less than or equal to 1.

Adjusted R^2: Adjusted R^2 depicts the same meaning as R^2 but is an improvement of it. R^2 suffers from the problem that the scores improve on increasing terms even though the model is not improving which may misguide the researcher. Adjusted R^2 is always lower than R^2 as it adjusts for the increasing predictors and only shows improvement if there is a real improvement.