MACHINE LEARNING - 8

1. b
2. b
3. c
4. c
5. d
6. b
7. c
8. a, d
9. b, d
10. a, b

11.

One hot encoding is used when we have to transform categorical data to numerical data for the model to understand. We must avoid one hot encoding when the categorical data is of ordinal type. This means OHE should not be used when the values of the categorical feature have a linear relation. For eg, Outstanding ⍰Good ⍰ Bad ⍰ Worse, in this case we can use OHE as it will provide equal weightage to all the feature and in short Outstanding and Worse would be same. This will cause a bias in the model. Thus we should avoid OHE. To counter this, we can use the ordinal encoding techniques/label encoding techniques which will preserve the relation between the values.

12.

The techniques widely used are: **1. Random Under Sampling** : Random Undersampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out. **2. 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. **3. Cluster based Over Sampling** : In this case, the 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. **4. Synthetic Minority Over-sampling Technique for imbalanced data (SMOTE)** : This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models. 5. **Modified SMOTE** : It is a modified version of SMOTE. SMOTE does not consider the underlying distribution of the minority class and latent noises in the dataset. To improve the performance of SMOTE a modified method MSMOTE is used.

13.

ADASYN uses density distributions 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 whereas SMOTE generates the same number of synthetic samples for each original minority sample.

14.

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. It 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. NO, For larger datasets the high dimensions will greatly slow down computation time and be very costly. In this instance, it is advised to use Randomized Search over Grid Search.

15.

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: **Mean Squared Error (MSE)** : The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset. **Root Mean Squared Error** : RMSE is an extension of the mean squared error where the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted. **Mean Absolute Error (MAE)** : The MAE score is calculated as the average of the absolute error values. Absolute or abs() is a mathematical function that simply makes a number positive. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE.