Machine Learning

Q1-B) In hierarchical clustering you don’t need to assign number of clusters in beginning

Q2-A) max_depth

Q3-D)- ADASYN

Q4-C) 1 and 3

Q5-A) 3-1-2

Q6-B) Support Vector Machines

Q7-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)

Q8)-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

Q9)-B) remove only one of the features

C) Use ridge regularization

D) use Lasso regularization

Q10-A) Overfitting

C) Underfitting

Q11 – One hot encoding should be avoided when the target variable has more than 2 outputs like suppose weather rain, summer,winter. In such cases Label Encoding technique can be used.

Q12 – In case of data imbalance the following techniques can be:

- Change the performance metric
- Change the algorithm
- Oversample minority class
- Undersample majority class
- Generate synthetic samples

Q13--**SMOTE:** 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. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

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

Q14- GridSearchCV is a library function that is a member of sklearn's model_selection package.
It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible.

Q15-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). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)

MSE-Mean squared error gives the mean of squared difference between model prediction and target value. It can be used as the measure of the quality of an estimator.

RMSE-Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general purpose error metric for numerical predictions.

MAE-Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.