Machine learning

1) B) In hierarchical clustering you don’t need to assign number of clusters in beginning
2) A) max_depth
3) A) SMOTE
4) C) 1 and 3
5) D) 1-3-2
6) B) Support Vector Machines
7) 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)
8) B) Lasso will lead to some of the coefficients to be very close to 0
9)
10) A) Overfitting

11) One Hot encoding has to be avoided in case it gives high cardinality. For such cases we should use label encoding.

12) Undersampling: Remove samples from majority set
Oversampling: Replicate sample in minority set
SMOTE: Synthesises new minority instances between existing minority instances

13) ADASYN 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 SMOTE generates the same number of synthetic samples for each original minority sample.

14) 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. For large datasets it becomes computationally expensive and takes longer time. For a large dataset we take a smaller subset and apply gridsearch to get best hyperparameters.

15) MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.
RMSE = sqrt(MSE), here the units of the RMSE are the same as the original units of the target value that is being predicted.
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.