Machine Learning Worksheet-8

1. B) In hierarchical clustering you don't need to assign number of clusters in beginning

2. A) max_depth

3. B) RandomOverSampler

4. C) 1 and 3

5. D) 1-3-2

6. B) Support Vector Machines

7. 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).

8. A) Ridge will lead to some of the coefficients to be very close to 0, D) Lasso will cause some of the coefficients to become 0.

9. B) remove only one of the features, D) use Lasso regularization

10. A) Overfitting, B) Multicollinearity, D) Outliers

11. In case of high cardinal categorical columns, one-hot encoding must be avoided to prevent the phenomena of **curse of dimensionality**. In this case, simple **label encoding** can work fine, but for more effective encoding, one can use supervised encoding techniques like **mean** or **target encoding.**

12. Techniques used for imbalanced dataset are:

- **Under-Sampling:** It is used to remove majority class samples at random to balance the dataset.
- **Over-Sampling:** It is used to add duplicate samples of minority class at random to the existing data, so that the sample of minority class increases.
- **SMOTE:** Synthetic Minority Over Technique (SMOTE), as the name suggests, this technique artificially synthesizes data points from minority class samples based on the **k-nearest neighbors** for that particular point.

13. The main difference between **SMOTE** and **ADASYN** sampling techniques is that the latter adds random value to the synthetic data point thus making it more realistic. In other words, instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e., they are bit scattered.

14. GridSearchCV is a hyper-parameter tuning algorithm from **scikit-learn** library. It is used to search the optimum hyper-parameters of a learning algorithm in an exhaustive manner, in other words, it tests every combination of hyper-parameters to arrive at the optimum.

It is usually not preferable to use GridSearchCV for large datasets mainly because as it can be very computationally expensive in this scenario, particularly if many sets of hyper-parameters are being tested.

15. The following are the common evaluation metrics used for regression task:

Root Mean Squared Error (RMSE): RMSE is a very common evaluation metric. It can range between 0 and infinity. Lower values are better.

Mathematically it is denoted by:

square root of (1/n * (∑ (y -ŷ) ²))

Here,

n = Number of observations

y = Actual value

ŷ = Predicted value

RMSE is proportional to the squared error and is sensitive to outliers and can exaggerate results if there are outliers in the data set. Therefore, it may not be an appropriate metric for evaluation in order to tell how well the model is doing as it may lead to less interpretation of the final result. However, if we care about penalizing large errors, it's not a bad choice, indeed it is a great choice for a loss metric when hyperparameter tuning the model is concerned.

Mean Absolute Error (MAE): Mean Absolute Error (MAE) is simply the average of the absolute value of the errors.

Mathematically it is denoted by:

(1 / n) * (∑ |y - ŷ|)

MAE is the simplest evaluation metric and most easily interpreted. It's a great

metric to use if we don't want a few far-off predictions to overwhelm a lot of close ones. It's a less good choice if you want to penalize predictions that were really far off the mark.

R-Squared (R^2): R^2 represents the proportion of variance explained by the model. R^2 is a relative metric, so we can use it to compare with other models trained on the same data.

Mathematically it is denoted by:

1 - (SSE/SST)

Here,

SSE = Sum of squared errors; the sum of the squared differences between the actual values and predicted values.

SST = The sum of the squared differences between the actual values and the mean of the actual values.