MACHINE LEARNING ASSIGNMENT -8

ANSWERS:

1. B) In hierarchical clustering you don't need to assign number of clusters in beginning
2. A) max_depth
3. C) RandomUnderSampler
4. C) 1 and 3
5. A) 3-1-2
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. D) Lasso will cause some of the coefficients to become 0.
9. B) remove only one of the features, C) Use ridge regularization, D) use Lasso regularization
10. A) Overfitting, B) Multicollinearity
11. One-hot encoding should be avoided when dealing with high cardinality categorical features, where the number of distinct columns is very large. In such cases, an alternative encoding technique such as get_dummies can be used.
12. Data Balancing Techniques:
    I. **SMOTE** (when dataset is small)– SMOTE picks minority class nearest samples and creates new data.
    II. **Under Sampling** using Near Miss(when dataset is big) – Picks majority class and randomly eliminates data to match ratio with minority class.
    III. **Up sampling** using Normal Method – Increase size of minority class to match majority class data size.
    IV. **Down Sampling** using Normal Method – Decrease size of majority class randomly to match size with minority class.
13. The main difference between SMOTE and ADASYN lies in how they generate synthetic examples. SMOTE creates synthetic samples by interpolating between existing minority class samples, while ADASYN generates synthetic samples by adaptively increasing the difficulty of classification of minority class samples.
14. GridSearchCV is a method used to search for the best hyperparameters for a machine learning model. The purpose of using GridSearchCV is to automate the process of tuning hyperparameters and selecting the best combination of hyperparameters for a given model. GridSearchCV can be time-consuming and memory-intensive, especially for large datasets with a large number of hyperparameters to tune and RandomizedSearchCV is preferable to use in such cases.
15. Evaluation Metrics:

I. Mean Absolute Error (MAE): MAE is a commonly used metric to measure the absolute difference between predicted and actual values. It is calculated by taking the mean of the absolute differences between each predicted and actual value.
II. Mean Squared Error (MSE): MSE is another commonly used metric to measure the average squared difference between predicted and actual values. It is calculated by taking the mean of the squared differences between each predicted and actual value.
III. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and it provides a measure of the average magnitude of the error.
IV. R-squared (R^2): R-squared measures the proportion of variance in the dependent variable that is explained by the independent variable(s). It takes values between 0 and 1, with 1 indicating that all variance is explained and 0 indicating that none of the variance is explained.
V. Adjusted R-squared: Adjusted R-squared is a modified version of R-squared that takes into account the number of independent variables in the model. It penalizes the use of additional independent variables that do not contribute to the overall performance of the model.