Worksheet 1 - submitted by Sukhpal Singh Int 34

MACHINE LEARNING

1. B
2. A
3. C
4. B
5. A
6. C
7. C
8. D, A
9. B, A
10. A
11. When dealing with categorical variables that have many unique values, using one-hot encoding may not be a good idea because it can create a lot of new columns and make the data too sparse. This can make it harder and slower to train a model. Instead, you can use target encoding or frequency encoding, which replace the categorical variable with a single value based on its relationship to the target variable or its frequency in the dataset. These encoding techniques can help simplify the data and make the model training process more efficient.
12. To balance a dataset with imbalanced classes, there are several techniques that can be used. Oversampling involves creating more instances of the minority class, while undersampling involves removing instances from the majority class. Cost-sensitive learning algorithms assign different misclassification costs to different classes, while ensemble methods like bagging and boosting combine multiple models to improve performance. However, undersampling should be done carefully to avoid losing important information.
13. Both SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are oversampling techniques used to address class imbalance in datasets. However, the main difference between the two is that ADASYN focuses more on generating synthetic samples in the harder-to-learn regions of the minority class distribution. In other words, ADASYN creates more synthetic samples near the decision boundary, while SMOTE creates synthetic samples in a more uniform way. This can make ADASYN more effective than SMOTE in certain scenarios, particularly when the decision boundary is complex or nonlinear.
14. GridSearchCV is a technique used to tune hyperparameters for a machine learning model. It allows us to specify different values for hyperparameters

and evaluates the model's performance with each combination of values. The purpose of using GridSearchCV is to find the best combination of hyperparameters that gives the highest performance score for the model. In general, it is not preferable to use GridSearchCV for large datasets because it can be computationally expensive and time-consuming. GridSearchCV can generate a large number of models and training each model can be very time-consuming, especially for large datasets. As an alternative, randomized search or Bayesian optimization can be used instead, as they are more efficient for large datasets.