Type- Core

Course Code- CSET301	**Course Name-AIML**
Year- 2023	**Semester- Odd**
Date- 11-09-2023	**Batch- 5th Sem**

LAB ASSIGNMENT - #5 SET-1

Name	**CO1**	**CO2**	**CO3**
	✓	-	-

Objective: To provide hands-on experience **on** imbalance dataset, Decision Tree visualization, hyperparameter tuning. We will learn to find best hyper parameter combination for a given dataset using Grid Search Cross Validation method to compute best score for Accuracy.

Tasks1 : Visualize imbalanced dataset and use class weights to improve the ROC-AUC score

1) Download Dataset: creditcard,csv provided in LMS
2) Load the data (csv file), read the dataset into the data frame 'df' and print the different statistical values and shape of data.
3) Separate the features into X and Y and print the shape.
4) Understand/ Visualized the distribution of target variable by showing the value counts of two classes.(seaborn library)
5) Initialize Decision Tree Models without tunning any hyperparameter.
6) Apply Repeated Stratified K-Fold cross validator function (.RepeatedStratifiedKFold) with n_splits =10, n_repeats=1 and random_stae=1)
7) Evaluate a score by cross-validation of ROC-AUC by fitting the data in Decision tree
8) Since the dataset is imbalanced use (Hyper parameter : **class_weight**="balanced") in the estimator and repeat step 8 & 9.
9) Summarize your findings of non-tunning and tunning the hyperparameter of Decision tree. Give reasoning for your results.

Task 2: Visualize Decision tree and select the best hyperparameter combinations for the Iris dataset using GridSearchCV(Exhaustive search over specified parameter values for an estimator).

a) Import necessary library and function, download Iris dataset from sklearn.

b) Read the dataset into the data frame ‘df’ and create decision tree classifier.

c) Evaluate a score by cross-validation by specifying the number of folds =5.

d) Visualize the decision tree for the Iris dataset.

e) Define the hyperparameter to search over max_depth = [2,4,6,8], min_samples_split = [2,4,6,8], min_samples_leaf = [1,2, 3].

f) Find the best hyperparameter combination to achieve ‘Best Score’ for accuracy. Compute the value of the best score.

Further Fun (will not be evaluated)

- Explore Scikit-learn Train Test Split — random state and shuffle.
- Explore ways to deal with imbalanced dataset. Use different methods (such as eliminating outliers, under sampling, oversampling) to experiment with the given dataset.
- Analyze the performance of the model using other cross-validation methods such as Stratified K-Fold, Leave One Out and Leave P Out etc.

Useful links

1. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
2. https://towardsdatascience.com/what-is-k-fold-cross-validation-5a7bb241d82
3. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html