School of Computer Science Engineering and Technology

Course- BTech	Type- Core
Course Code- CSET301	Course Name-AIML
Year- 2022	Semester- Odd
Date- 31-11-2022	Batch- V Sem

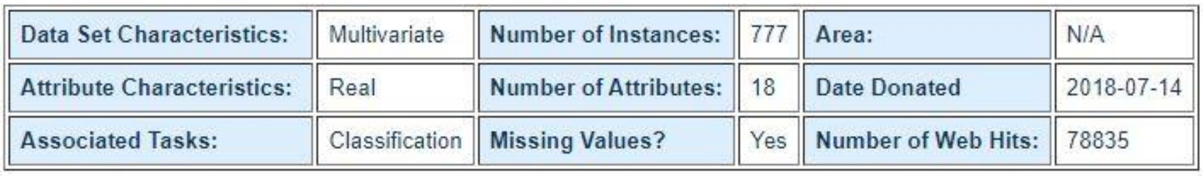
9 - Lab Assignment No. (9.1.1)

Exp. No.	Name	CO-1	CO-2	CO-3
9.1.1	AdaBoost Classifier	✓	✓	--

Objective: To implement AdaBoost classifier (using Scikit-learn) and perform binary classification after suitable pre-processing steps.

Download the dataset from:
https://archive.ics.uci.edu/ml/datasets/Audit+Data (10)

About Dataset:



Data Set Characteristics:	Multivariate	Number of Instances:	777	Area:	N/A
Attribute Characteristics:	Real	Number of Attributes:	18	Date Donated	2018-07-14
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	78835

Data Set Information:

The goal of the research is to help the auditors by building a classification model that can predict the fraudulent firm on the basis the present and historical risk factors. The information about the sectors and the counts of firms are listed respectively as Irrigation (114), Public Health (77), Buildings and Roads (82), Forest (70), Corporate (47), Animal Husbandry (95), Communication (1), Electrical (4), Land (5), Science and Technology (3), Tourism (1), Fisheries (41), Industries (37), Agriculture (200).

Attribute Information:

Many risk factors are examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm reputation summary, on-going issues report, profit-value records, loss-value records, follow-up reports etc. After in-depth interview with the auditors, important risk factors are evaluated and their probability of existence is calculated from the present and past records.

1. **Data Pre-processing step: (40)**
    a) Read audit_risk.csv using Pandas and display First 5 rows.
    b) Check the presence of Null Values/Missing Values. If present handle them with suitable approach.
    c) Display the data samples using scatter plot.
    d) Find and remove outliers if any using K-nearest neighbours approach (K = 5).

2. **Model Preparing and Evaluation:**
    a) Split the dataset into 80% for training and rest 20% for testing (sklearn.model_selection.train_test_split function) (5)
    b) Train ensemble classifier **Adaboost** using built-in function on the training set with default parameters (sklearn.ensemble.AdaBoostClassifier)(10)
    c) Evaluate the train model using test set with the help of confusion matrix, Accuracy, Precision and Recall. (10)
3. **Parameter Tuning:**
    a) Try with n_estimator as [50, 100, 150] (15)
    b) Learning Rate as [1.0, 0.01, 0.001] (15)

4. Compare the results and find the best suitable model

Suggested Platform: Python: Jupyter Notebook/Azure Notebook/Google Colab.