LAB ASSIGNMENT-7

Title: Train any machine learning classifier on the imbalanced dataset. Then balance the dataset by using oversampling techniques. Compare the model performance before and after oversampling.

Objective: In this lab assignment, you will work with an imbalanced dataset and train a machine learning classifier on it. After that, you will apply oversampling techniques to balance the dataset and compare the model's performance before and after oversampling. The goal is to observe how oversampling affects the classifier's performance when dealing with imbalanced data.

Tasks:

1. Load the dataset from
2. Explore the dataset and analyze the class distribution to verify the imbalance.
3. Preprocess the data (handle missing values, convert categorical variables, etc.) if necessary.
4. Split the dataset into training and testing sets.
5. Choose a machine learning classifier of your choice. For example, you can use Logistic Regression, Random Forest, or Support Vector Machine (SVM).
6. Train the chosen classifier on the imbalanced dataset and evaluate its performance on the test set.
7. Apply oversampling techniques (e.g., Random Oversampling, SMOTE - Synthetic Minority Over-sampling Technique) to balance the dataset.
8. Train the same classifier on the balanced dataset obtained after oversampling and evaluate its performance on the test set.
9. Compare the performance metrics (e.g., accuracy, precision, recall, F1-score) of the classifier before and after oversampling.
10. Discuss your observations and insights into how oversampling affects the model's performance on the imbalanced dataset.

Submission Guidelines:

- Prepare a Jupyter Notebook (.ipynb) or a Python script (.py) to demonstrate your work.
- Include necessary comments and markdown cells to explain each step clearly.
- Present the performance metrics (e.g., confusion matrix, classification report) in a readable format.
- Provide your observations and conclusions in a separate section.
- If you used any external libraries, include instructions on how to install them using pip/conda.

Important Note: Remember to handle the class imbalance issue properly while training and testing the classifier. Pay attention to the performance metrics and not just accuracy, as accuracy alone can be misleading when dealing with imbalanced datasets.

Download Dataset from:
Dataset-1: https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv