Lab Assignment: Comparative Analysis of Classification Models on Synthetic Datasets

Objective

In this lab, you will explore the performance of different classification models on various synthetic datasets. You will create and analyze datasets where specific classifiers perform better than others. The goal is to understand the strengths and weaknesses of each classifier and the characteristics of the data that influence their performance.

Classification Models to Explore

1. **Parametric Methods:**
   - Naïve Bayes
   - Logistic Regression
   - Support Vector Machines (SVM) with a linear kernel
2. **Non-Parametric Methods:**
   - Decision Trees
   - Decision Rules
   - Nearest Neighbor

Task Overview

You are required to choose three pairs of classifiers from the list above. For each pair, you will create two synthetic datasets:

- **Dataset 1:** The first classifier from the pair performs better.
- **Dataset 2:** The second classifier from the pair performs better.

For each dataset, you must explain why the observed performance difference occurs. This will involve analyzing the characteristics of the datasets and understanding the nature of the classifiers.

Steps to Complete the Lab

1. Select Pairs of Classifiers

Choose three pairs of classifiers from the list provided. You can choose any combination of parametric and non-parametric methods.

2. Generate Synthetic Datasets

For each pair, generate two synthetic datasets. Use features such as linear separability, feature independence, class overlap, or non-linear boundaries to influence the performance of the classifiers. Ensure that the datasets are simple and illustrative, ideally with 2 or 3 features, so that they can be visualized and easily interpreted.

3. Train and Evaluate Classifiers

Split each dataset into training and testing sets (e.g., 70% training, 30% testing). Train both classifiers on the training set and evaluate their performance on the testing set. Record the accuracy (or other relevant metrics) of each classifier on both datasets.

4. Analyze Results

For each dataset, explain why one classifier outperforms the other. Consider aspects such as:

- Assumptions made by the classifiers (e.g., Naïve Bayes assumes feature independence).
- The linearity or non-linearity of decision boundaries.
- Sensitivity to overfitting or noise in the data.
- The impact of feature correlations, class distributions, or other characteristics of the data.

Submission Requirements

Please follow the guidelines below for submitting your assignment's solutions:

- Analytical Report: Submit a PDF file that serves as your Analytical Report. This should contain all your written answers, interpretations, and graphical visualizations for each exercise. Ensure that the graphics are clearly labeled and appropriately integrated into your explanations.
- Jupyter Notebook PDF: Additionally, submit a PDF version of your Jupyter Notebook that contains all the code used for data generation, analysis, and visualization. Make sure that the code is well-commented for readability.