Project Title: Exploring Classification Techniques on Small and Imbalanced Datasets: A Case Study with the Wine Recognition Dataset

Project Goal: This project aims to investigate the behavior and performance of various classification algorithms on a small and moderately imbalanced dataset, using the Wine Recognition Dataset from the UCI repository (available via scikit-learn). The dataset contains only 178 instances, offering a unique opportunity to explore the limitations and strengths of machine learning models in low-data settings.

The primary research question is: How do different classification algorithms perform on small datasets with imbalanced classes, and how does binary classification compare to multi-class classification in such settings?

We will apply and compare the following classification methods:

- Naive Bayes
- K-Nearest Neighbors (KNN)
- Gaussian Process Classification
- Support Vector Machine (SVM)

Our exploration will be structured into four key aims:

Aim 1 (Performance on Small Datasets): We will analyze how well these algorithms generalize with limited training data and whether specific models show better stability and accuracy in small-sample scenarios.

Aim 2 (Binary vs. Multi-Class Classification): We will reframe the dataset to form a binary classification task by merging class_0 (59 instances) and class_1 (71 instances) into a new label, class_a, and keeping class_2 (48 instances) as class_b. This will allow us to compare performance metrics between multi-class and binary setups using the same dataset.

Aim 3 (Model Behavior Under Class Imbalance): Given the imbalance in the class distribution (e.g., class_2 having fewer samples), we will evaluate which classification models are more robust against imbalanced data and how different evaluation metrics (accuracy, precision, recall, F1-score, and ROC-AUC) reflect this robustness.

Aim 4 (Comparative Evaluation): All models will be evaluated and compared using cross-validation and relevant performance metrics. The goal is to identify which classifier provides the best trade-off between accuracy and generalization across balanced and imbalanced settings.

This project will provide insights into effective model selection and data preprocessing strategies when dealing with small and imbalanced datasets in real-world scenarios.

Submitted by: Sajjad Islam, Tanjina Zaman, Dewan Imran