Machine Learning - Term 242
Assignment 2: Classification

1. Dataset

The dataset utilized for this assignment, titled **Diabetes Prediction**, is publicly available on Kaggle. It comprises data for the classification of 13 distinct types of diabetes. The dataset can be accessed at the following link: https://www.kaggle.com/code/renjiabarai/diabetes-prediction-13-classes/input

This comprehensive dataset includes medical data associated with various forms of diabetes, such as Steroid-Induced Diabetes, neonatal diabetes mellitus (NDM), Prediabetes, Type 1 Diabetes, Wolfram Syndrome, and more. It contains diverse features, including genetic markers, environmental and lifestyle factors, and medical indicators. The dataset provides a valuable opportunity to explore multi-class classification in a real-world healthcare context.

2. Objective

The objective of this assignment is to apply supervised machine learning techniques to develop a multi-class classification model capable of predicting the type of diabetes based on patient data.

3. Tasks

3.1 Model Development (50 points)

You are required to implement and evaluate the following classification models:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- A bagging ensemble method (e.g., Random Forest)
- A boosting ensemble method (e.g., AdaBoost)

Note: Apply appropriate feature selection techniques to enhance model performance if needed.

3.2 Model Evaluation (20 points)

- Cross-validation (10 points): Use 5-fold cross-validation to ensure robustness of your evaluation.
- Metrics (10 points): Evaluate and compare your models using Accuracy and F1-Score

4. Submission Guidelines

- Deadline: 11:59 PM, April 26, 2025
- Submission Platform: **Blackboard**
- Submission Format: A single ZIP file named with your student ID, containing the following items:
  - ✓ Source Code (5 points): Clean, well-documented, and executable code
  - ✓ Report (20 points) – PDF format (2–3 pages):

- Experimental Setup (10 points): Describe the dataset, preprocessing steps, feature selection techniques, and model configurations.
- Model Comparison (10 points): Present a comparative performance analysis of the implemented models, supported by tables or figures, and justify your findings.

- Timeliness (5 points): Late submissions may incur penalties

Ensure that your report is well-structured, concise, and professionally formatted

Good luck!