BIRZEIT UNIVERSITY
Electrical and Computer Engineering Department
Machine Learning and Data Science - ENCS5341
Assignment #3
Submission deadline: 21/12/2024

This assignment may be completed by a group of up to two students.

Topic: KNN, Logistic Regression, SVM, Kernel Methods, and Ensemble Methods (Boosting and Bagging)

Objective:
This assignment aims to deepen your understanding of core machine learning algorithms by implementing and evaluating K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM) with various kernels, and ensemble methods such as Boosting and Bagging. You will experiment with different configurations, compare their performance, and evaluate the models using classification metrics.

Instructions:

1. **Dataset:** Use a publicly available dataset such as the *Iris dataset*, *Breast Cancer dataset*, or any other classification dataset of your choice. Ensure the dataset has at least two classes.
2. **Tools:** Implement the algorithms using Python APIs. Use libraries such as NumPy, Pandas, Scikit-learn, and Matplotlib for preprocessing, analysis and visualization.
3. **Evaluation Metrics:**
    - Accuracy
    - Precision
    - Recall
    - F1-score
    - ROC-AUC

Assignment Tasks

Part 1: K-Nearest Neighbors (KNN)

1. Use APIs to implement the KNN algorithm.
2. Experiment with at least **three different distance metrics** (e.g., Euclidean, Manhattan, and Cosine distance). Compare their impact on the model's performance.
3. Use cross-validation to determine the optimal value of **K** (number of neighbors).
4. Analyze the results and discuss:
    - How do different distance metrics affect the classification performance?
    - What is the best value of K for your dataset, and why?

Part 2: Logistic Regression

1. Train a Logistic Regression model on the dataset.
2. Experiment with different regularization techniques (e.g., L1, L2).
3. Evaluate its performance using the classification metrics.
4. Compare the performance of Logistic Regression with KNN.

Part 3: Support Vector Machines (SVM)

1. Use APIs to implement SVM and train it using the dataset.
2. Try **at least three kernels** (e.g., linear, polynomial, radial basis function (RBF)).
3. Compare the performance of the kernels using classification metrics.
4. Discuss how the choice of the kernel impacts model accuracy and other evaluation metrics.

Part 4: Ensemble Methods

1. **Boosting:** Train a model using *AdaBoost*.
2. **Bagging:** Train a model using *Bagging* or *Random Forest*.
3. Compare the performance of Boosting and Bagging methods.
4. Discuss:
    - Which ensemble method performed better and why?
    - How do ensemble methods compare to individual models (KNN, Logistic Regression, SVM)?

Deliverables:

1. **Code:** Submit a well-documented Jupyter Notebook or Python script with your implementation.
2. **Report:** A concise report (7–9 pages) that includes:
    - A brief introduction to each method.
    - Your approach to experimenting with the algorithms.
    - A detailed analysis of the results, including metric comparisons.
    - Conclusions based on your findings.
3. Please compress your files, including both the code and the report, into a single zip file and submit it to the ritaj before the deadline. The file name should follow this format: "LastName_ID_Student1_LastName_ID_Student2.ZIP".

Submission Deadline:

Submit your assignment by **December 21st, 2024**. Late submissions will be accepted up to 3 days after the deadline, with a 10% deduction for each day delayed.

Group Work Policy

This project can be completed in groups of **two students at most**. Each student should clearly state their contribution to the project in the report.