Alexandria National University
Faculty of Computer and Data Science

Introduction to Machine Learning
Assigned: Friday, April 18, 2025
Due: Friday, April 25, 2025

Assignment 2
Data Classification

1 Objectives

1. Exploring different classification models and performing tuning of their parameters.
2. Exploring different techniques for evaluating classification models.

2 Problem Statement

Given the MAGIC gamma telescope dataset that can be obtained using this Link. This dataset is generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. The dataset consists of two classes; gammas (signal) and hadrons (background). There are 12332 gamma events and 6688 hadron events. You are required to use this dataset to apply different classification models such as **Decision Trees**, **Na¨ıve Bayes Classifier**, **Random Forests** and **AdaBoost** You are also required to tune the parameters of these models, and compare the performance of models with each other.

3 Lab session

1. **Data Balancing**
   Note that the dataset is class-imbalanced. To balance the dataset, randomly put aside the extra readings for the gamma “g” class to make both classes equal in size.
2. **Data Split**
   Split your dataset randomly so that the training set would form 70% of the dataset and the testing set would form 30% of it.

3. Classification

Apply the classifiers from the following models on your dataset, tune parameter(s) (if any), compare the performance of models with each other:

(a) Decision Tree
Parameters to be tuned: None

(b) AdaBoost
Parameters to be tuned: n estimators

(c) Random Forests
Parameters to be tuned: n estimators

(d) Na¨ıve Bayes
Parameters to be tuned: None

4. Model Parameter Tuning

Use cross-validation to tune the parameters of classifiers. Test the models trained with best obtained parameter values on the separate testing set.

5. Report Requirements

- For all the requirements mentioned above you should report the model accuracy, precision, recall and F-score as well as the resultant confusion matrix using the testing data.
- Your comments on all results and comparisons.

6. Notes

- You should write your code in python.
- You can use a third-party machine learning implementation like scikit-learn.
- You should work in groups of 3.

Good Luck