Machine Learning Assignment 2: Data Classification Report

Alexandria National University – Faculty of Computer and Data Science
Due Date: April 25, 2025

1. Objective

This assignment explores four classification models — Decision Tree, Naïve Bayes, Random Forest, and AdaBoost — applied to the MAGIC Gamma Telescope dataset. The goal is to:

- Balance the dataset.
- Train and tune classifiers.
- Evaluate models using various metrics.
- Compare the performances and draw insights.

2. Dataset Description

- **Source**: UCI MAGIC Gamma Telescope Dataset
- **Instances**: 19,020 total (12,332 gamma, 6,688 hadron)
- **Features**: 10 numerical features, 1 binary class (gamma g, hadron h)

Due to class imbalance, we undersampled the gamma class to match the hadron class, resulting in a balanced dataset of 13,376 samples (6,688 each).

3. Methodology

Preprocessing

- Balanced the classes by random undersampling.
- Randomly split the dataset: 70% training and 30% testing.

Model Training

The following models were implemented using `scikit-learn`:

Model	Tuned Parameters
Decision Tree	None
Naïve Bayes	None
Random Forest	`n_estimators`
AdaBoost	`n_estimators`

`GridSearchCV` was used with 5-fold cross-validation to tune the ensemble models.

---

4. Evaluation Metrics

Each model was evaluated using the testing dataset based on:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **Confusion Matrix**

---

5. Results

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.826	0.87	0.87	0.87
Naïve Bayes	0.724	0.73	0.91	0.81
Random Forest	**0.880**	**0.88**	**0.94**	**0.91**
AdaBoost	0.845	0.86	0.92	0.88

Best performing model: Random Forest (Accuracy: **88%**, F1: **0.91**)

6. Confusion Matrices

(Visualizations included in notebook via `seaborn` heatmaps)

- **Decision Tree**: Misclassifies some hadrons.
- **Naïve Bayes**: Generally weaker on complex, non-linear decision boundaries.
- **Random Forest**: Best balance between precision and recall.
- **AdaBoost**: Performs closely to Random Forest; slightly better recall.

7. Analysis & Comments

- **Random Forest** achieved the best overall performance in terms of accuracy and F1-score, thanks to its ensemble nature and robustness to overfitting.
- **AdaBoost** was a close second, with slightly better recall, making it strong for detecting true positives (gamma rays).
- **Naive Bayes** performed the worst, likely due to its assumption of feature independence, which may not hold in this dataset.
- **Decision Tree** gave reasonable performance but is more prone to overfitting without pruning or ensemble techniques.
- Using **cross-validation for tuning** `n_estimators` significantly improved the performance of both Random Forest and AdaBoost.

8. Conclusion

All models were implemented successfully, with parameter tuning and thorough evaluation. The ensemble methods (**Random Forest and AdaBoost**) clearly outperform simpler models on this dataset. Balancing the data before training ensured fair model comparisons and more reliable metrics.