Akshata Atpadkar - 23055588
Group NO:24
Github_link: Classification of Gamma Rays

Classification of Gamma Rays

Introduction

The MAGIC Gamma Telescope dataset provides data on cosmic particles detected by a telescope. Classifying hadronic showers (background noise) and gamma rays (signal) using machine learning techniques is the main objective. This classification is crucial for astrophysical research because it helps separate important observations from irrelevant background noise.

This study focuses on two classification methods: Support Vector Machine (SVM) and Random Forest; a Decision Tree is employed as a reference model. The best strategy for this dataset is determined by evaluating these models.

Data Preprocessing

The dataset consists of ten numerical properties related to the distribution, shape, and concentration of the particles detected. Before being converted into binary values, the target variable (class) was categorical (g for gamma-ray, h for background), with 1 representing gamma and 0 representing background.

MinMax Scaling was utilized because SVM needs normalized input. Neither Random Forest nor Decision Tree requires scaling.
Data Splitting: To assess the model, the dataset was split into 80% training and 20% testing.

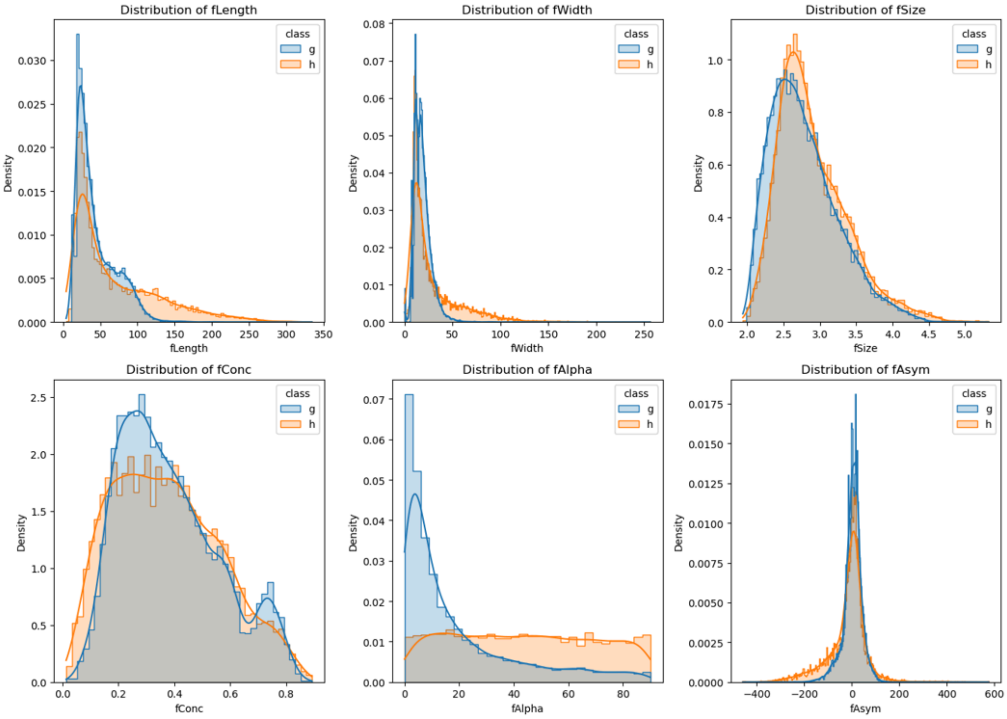
Model Comparison and Results: Model performance was measured using F1-score, recall, accuracy, and precision.

Key Observations:

- SVM showed strong performance but struggled with outliers, which affected its decision boundary.
- Decision Tree performed well but overfitted slightly, reducing its generalization ability.
- Random Forest achieved the highest accuracy, handling outliers better and reducing overfitting.

Feature Distribution Analysis:

The plots display the differences between each feature in hadrons (h) and gamma rays (g). Features such as fAlpha and fConc are helpful for classification because they exhibit distinct separation. Others, such as fSize and fAsym, are complementary despite their overlap. The model's emphasis on the most instructive features is supported by this.



Model name	Accuracy	Precision	Recall	F1 Score
Random Forest	0.878023	0.881498	0.937398	0.908589
SVM	0.876183	0.862295	0.962195	0.909510
Decision Tree	0.813617	0.819876	0.912195	0.863575

Conclusion & Recommendations

Random Forest achieved the highest accuracy (88.1%), making it the best choice for this dataset.
SVM was affected by outliers and required extensive preprocessing, making it less practical in this case.
Feature importance analysis showed fAlpha and fDist were the most critical variables.

SVM works best in well-separated data but struggles with noisy datasets like this one. Random Forest provides a more generalized solution and consistently outperforms SVM