Data Mining for Particle Classification: Analysing the Magic Gamma Telescope Dataset

CPS844 **April 2025**

By Aseef Nazrul

Table of contents

I. Introduction

In this project, we analyze data from the MAGIC Gamma Telescope to classify high-energy particle events. The MAGIC telescope detects gamma-ray events from space by observing Cherenkov radiation, light emitted when gamma rays interact with the Earth's atmosphere [1]. However, distinguishing true gamma-ray signals from hadronic (cosmic-ray) backgrounds remains a significant challenge, as incorrect classification can negatively impact scientific results and resource allocation.

To address this classification problem, we employ various data mining methods, including Decision Trees, Naive Bayes, Support Vector Machines (SVM), Random Forests, and XGBoost. We use a dataset from the UCI Machine Learning Repository containing features that describe the characteristics of detected particle showers. These characteristics, such as the shape, size, and asymmetry of particle events, are essential for identifying patterns that distinguish gamma rays from background particles.

Our approach involves performing exploratory data analysis to understand the dataset thoroughly, followed by preprocessing steps like normalisation and feature scaling. We then compare the performance of multiple classification algorithms using measures like accuracy, precision, recall, and area under the ROC curve (AUC). Additionally, we investigate feature selection techniques to identify which features contribute most effectively to accurate predictions.

The primary goal of this project is to identify the most effective model for accurately classifying gamma-ray events. By achieving this, we aim to provide valuable insights for future particle physics research and observational techniques, ensuring better scientific outcomes and more efficient data processing strategies.

II. Dataset Description

For our project, we used the MAGIC Gamma Telescope dataset from the UCI Machine Learning Repository [2]. This dataset contains simulated data created using the Monte Carlo program CORSIKA, which models particle showers detected by atmospheric Cherenkov telescopes. These simulations help researchers study how high-energy gamma rays interact with the Earth's atmosphere.

The dataset includes 19,020 events, each representing a single particle shower. Each event has 10 continuous features that describe the shape, size, and orientation of the shower. These features include fLength and fWidth (axis lengths), fSize (overall brightness), fConc and fConc1 (light concentration), fAsym (asymmetry), fM3Long and fM3Trans (third moments), fAlpha (orientation angle), and fDist (distance to the camera center). All values are numeric, and there are no missing entries.

Each event in the dataset is labeled as either a gamma-ray signal or a hadronic background. While there are more gamma events than hadronic events, the difference in class sizes is significant enough to impact model performance. Since imbalanced datasets can cause models to favor the majority class, it is important for us to understand the distribution before training and evaluating classifiers. We discuss the class imbalance and its implications in more detail in the other sections.

We chose this dataset because it is well-suited for classification tasks and allows us to test different data mining models. The goal is to accurately identify gamma-ray events, which are important for astrophysics research. Because of the class imbalance and scientific importance, we used evaluation metrics like AUC-ROC in addition to accuracy to compare model performance.

III. Exploratory Data Analysis

We first explored the dataset to better understand the features and their distributions. The dataset contains 19,020 events, with about 65% gamma-ray signals and 35% hadronic backgrounds.

Each event has ten numeric features, describing characteristics like the length, width, and shape of the observed particle showers. We noticed that most features varied widely; for example, the major axis length (fLength) ranged from 4.3 mm to 334.2 mm, and the minor axis width (fWidth) ranged from 0 mm to 256.4 mm. Another feature, fSize, measuring the intensity of showers, varied between 1.94 and 5.3.

Initially, we were confused by negative values in certain features, such as fAsym, which ranged from -457.9 to 575.2. After further review, we understood that negative values simply indicate direction along the ellipse axis, meaning the highest intensity pixel is located opposite the positive reference direction along that axis. Similarly, negative values in features like fM3Long (from -331.8 to 238.3 mm) and fM3Trans indicate asymmetrical distributions of the shower intensity.

We also examined correlations between features using a heatmap, noting several strong relationships. Understanding these feature distributions and their relationships helped us decide on suitable preprocessing methods and guided our choice of data mining models for accurate classification.

IV. Data Preprocessing

Before building our data mining models, we performed some preprocessing on the MAGIC Gamma Telescope dataset. We began by splitting the dataset into a training set and a test set to evaluate our models fairly. We used a 70/30 split, which gave us 13,314 training samples and 5,706 test samples. To keep the proportion of gamma-ray and hadronic events consistent in both sets, we applied stratified sampling based on the class labels, ensuring balanced representation during training and evaluation.

Next, since our dataset contains ten continuous features with different ranges and scales, we standardized the data using StandardScaler. This step ensured that each feature had a mean of 0 and a standard deviation of 1. Standardization helped make all features equally important to our models and prevented any single feature from dominating due to its scale. Without this step, models like SVMs and k-NN could have been biased toward larger-valued features.

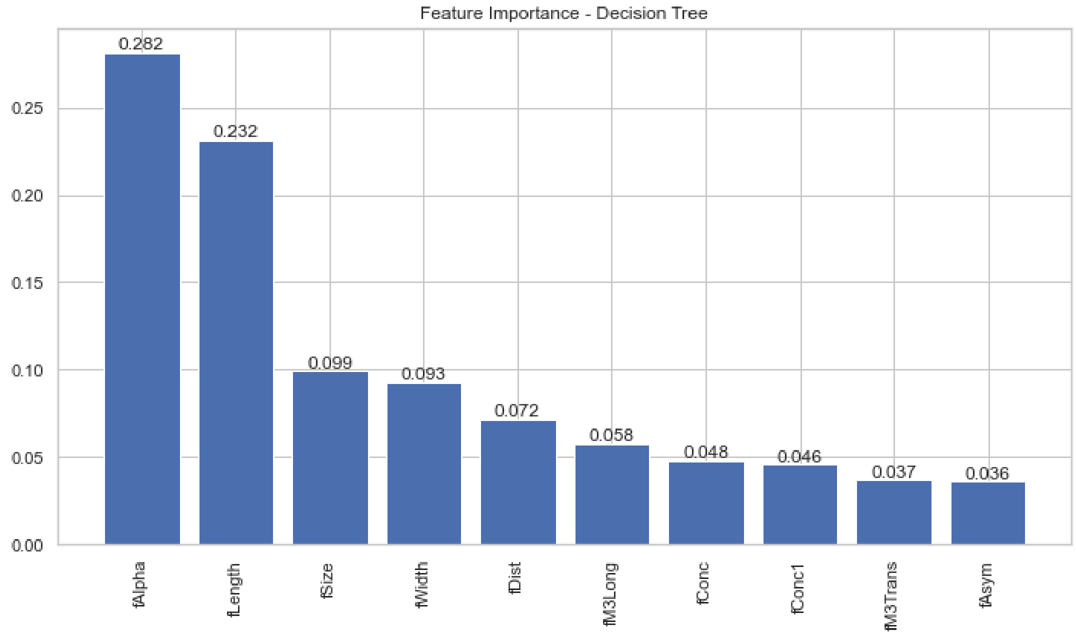
At this stage, we intentionally avoided other preprocessing techniques like aggregation, discretization, dimensionality reduction, feature creation, or feature selection. We chose to keep the preprocessing simple so we could clearly observe how different algorithms performed with minimal manipulation of the original data. By doing this, we aimed to establish a reliable baseline and better understand the raw predictive power of each model before exploring further improvements.

After completing these preprocessing steps, our dataset was fully prepared for training and evaluating different data mining models.

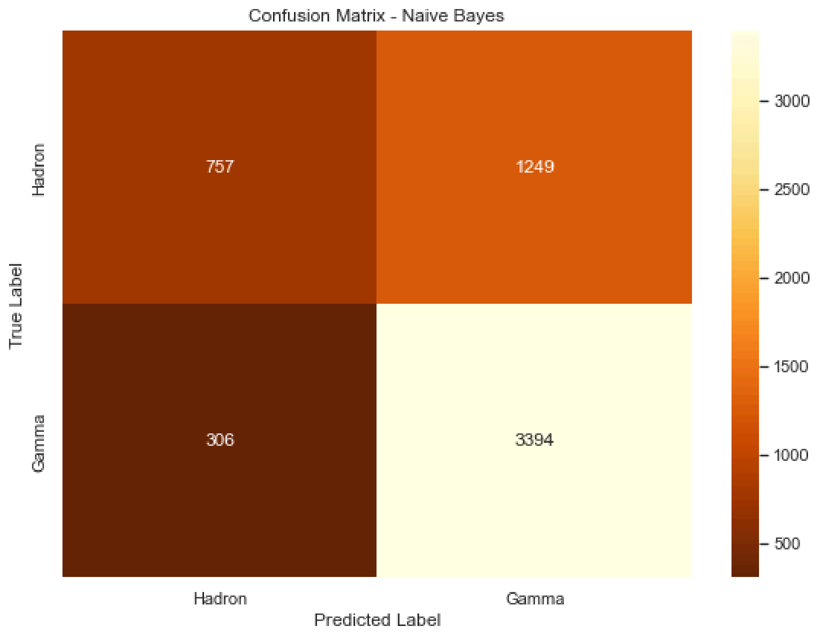
V. Data Mining Models

In this section, we evaluated five different data mining models: Decision Tree, Naive Bayes, Support Vector Machine (SVM) with an RBF kernel, Random Forest, and XGBoost. Each model was trained using 5-fold cross-validation on the training dataset and tested on a separate set of 5,706 samples to assess performance.

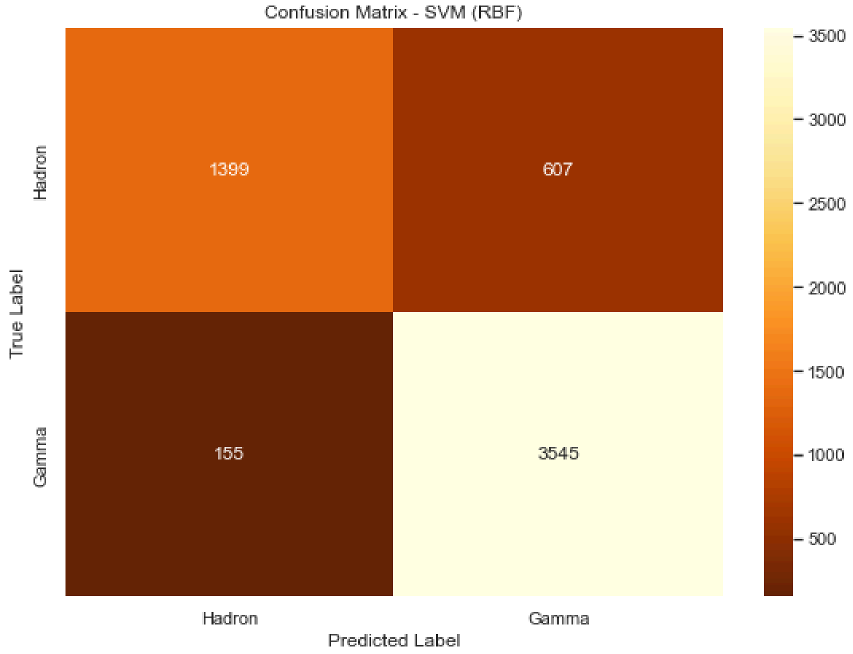
The ***Decision Tree*** model provided a straightforward and interpretable baseline for classification. It achieved a cross-validation accuracy of 80.79% and a test accuracy of 82.09%. For Gamma events, it performed well, with precision and recall both above 85%, showing good reliability. However, it had more difficulty with Hadron events, leading to 495 false positives and 527 false negatives. This imbalance suggests the model was better at detecting Gamma events than Hadron events. The most important features driving the model's predictions were fAlpha and fLength, which contributed significantly to the classification decisions.



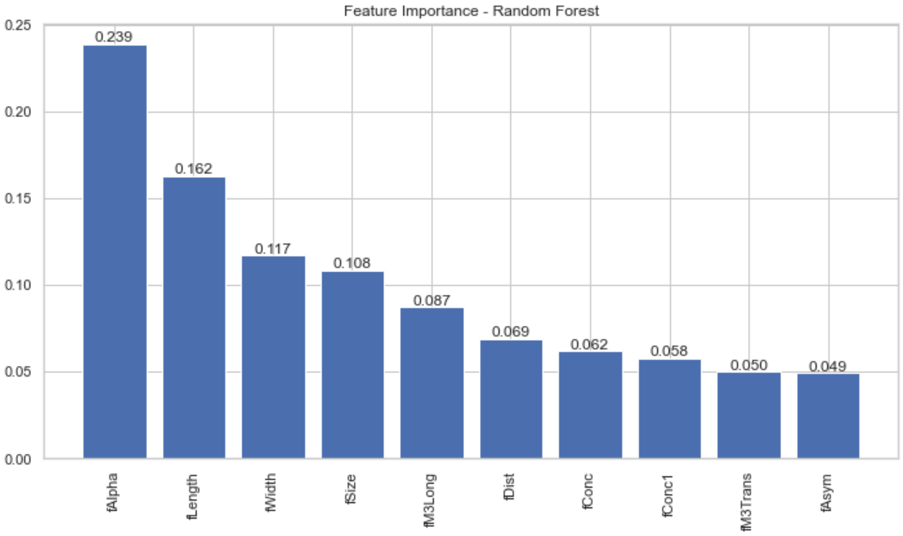
The ***Naive Bayes*** model, known for its simplicity and speed, showed lower performance compared to the other models, achieving a test accuracy of 72.75%. It struggled particularly with Hadron events, often misclassifying them as Gamma events. This weakness resulted in 1,249 false positives, significantly impacting the model's reliability for detecting Hadron events. While its precision for Gamma classification was decent at 73.1%, the recall for Hadron was very low at 37.7%, meaning the model failed to correctly identify a large portion of Hadron events. This imbalance between classes limited the overall effectiveness of Naive Bayes for the task, suggesting that its strong independence assumptions between features may not align well with the characteristics of the dataset.



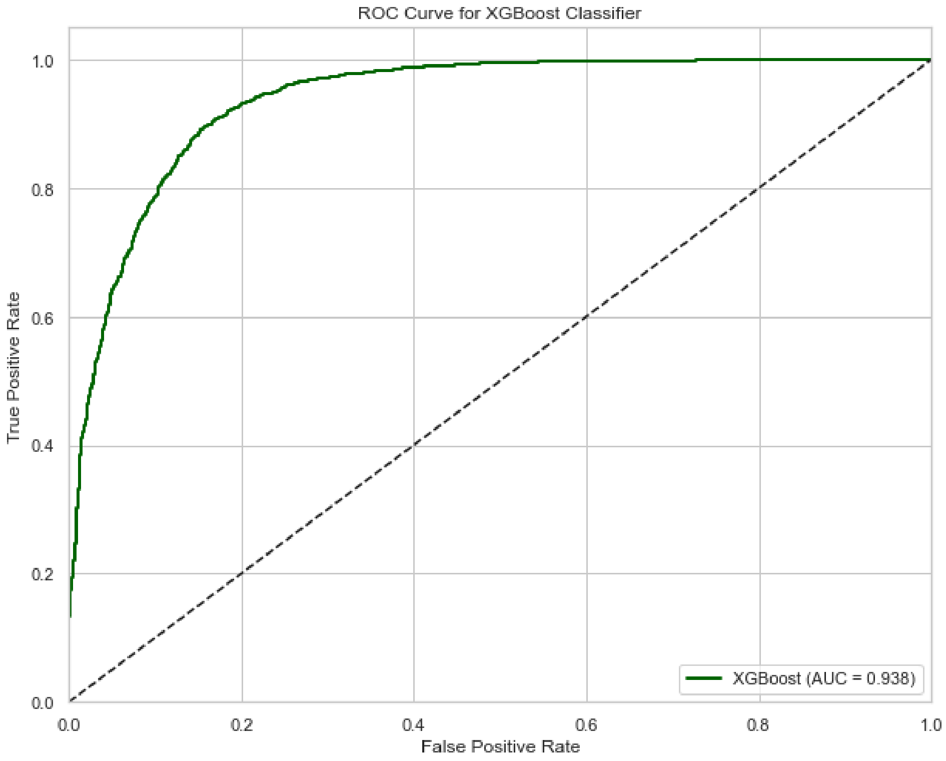
The ***SVM*** model showed strong performance, achieving a test accuracy of 86.65% and a high AUC of 0.9194. Precision and recall for Gamma events were impressive (85.4% precision, 95.8% recall), but recall for Hadron events was lower at 69.7%, leading to 607 false positives and 155 false negatives. The confusion matrix clearly illustrates these challenges.



The ***Random Forest*** model, leveraging an ensemble learning approach, performed robustly with a test accuracy of 88.33% and an AUC of 0.9351. It achieved high precision and recall for both classes, showing consistent and reliable results. Performance was especially strong for Gamma events, where the model reached a precision of 88.6% and a recall of 94.1%, indicating its ability to correctly identify most Gamma signals. The confusion matrix showed relatively few misclassifications. The most influential features in the Random Forest model were fAlpha, fLength, and fWidth, which contributed significantly to the model's predictions and helped improve its classification accuracy.



Finally, the ***XGBoost*** model delivered the highest performance among all the models tested, with a test accuracy of 88.47% and the highest AUC of 0.9381. It excelled in both precision and recall for Gamma events, achieving 89.0% precision and 93.8% recall, which shows its strong ability to correctly identify Gamma signals. The model had relatively few misclassifications, with 429 false positives and 229 false negatives, indicating strong reliability across both classes. The key features that contributed most to its success were fAlpha, fLength, and fWidth. These features played an important role in helping the model distinguish Gamma events from Hadron events effectively.

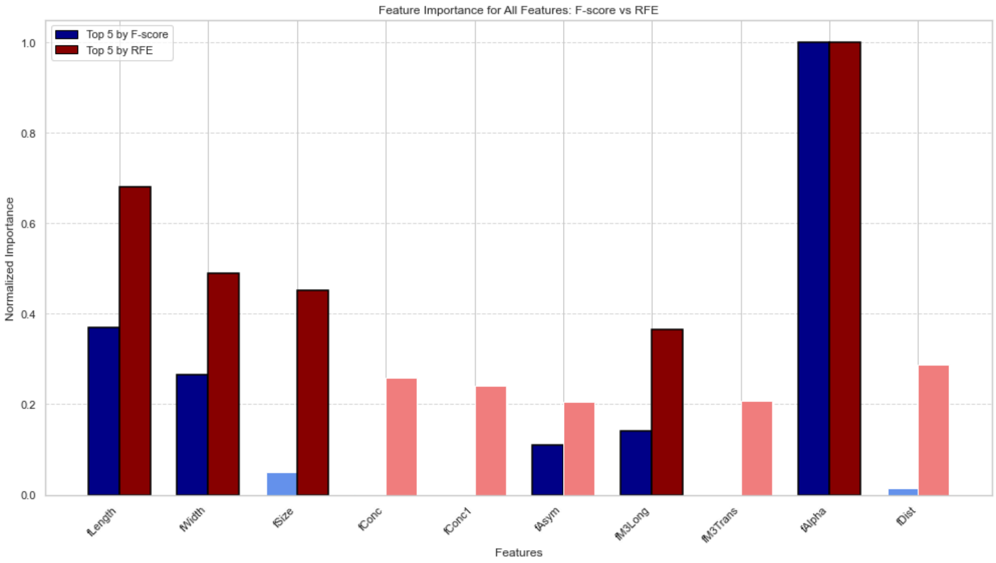


Overall, Random Forest and XGBoost performed the best, showing high accuracy and reliability in distinguishing Gamma-ray signals from Hadronic background events. These models outperformed others like Naive Bayes, and XGBoost gave the strongest results. This shows that more complex models can be more effective for this type of classification task.

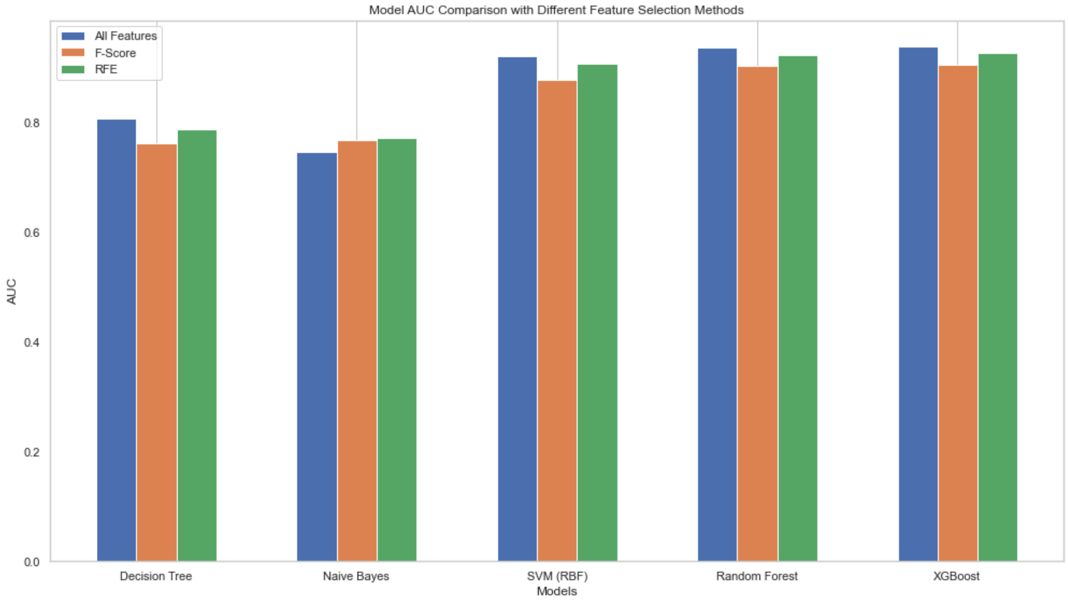
VI. Feature Selection

To make our model easier to understand and more efficient, we used two feature selection methods: SelectKBest with F-score and Recursive Feature Elimination (RFE) with a Random Forest classifier. These methods helped us find the most important features for predicting particle classification.

With the F-score method, we chose the five most important features: fAlpha, fLength, fWidth, fM3Long, and fAsym. These features had the strongest relationships with the target class. For RFE, we used a Random Forest classifier to gradually remove less important features. This method selected a slightly different set of features: fLength, fWidth, fSize, fM3Long, and fAlpha. Four features, fAlpha, fLength, fWidth, and fM3Long, were identified as important by both methods. The main difference was that F-score picked fAsym, while RFE picked fSize.



We then compared how the models performed using all features versus the selected features based on AUC (Area Under the Curve). The comparison shows that the models performed well even with fewer features. For example, Random Forest's AUC dropped slightly from 0.935 (all features) to about 0.903 (F-score) and 0.922 (RFE). Similarly, XGBoost's AUC dropped from 0.938 (all features) to about 0.903 (F-score) and 0.926 (RFE).



Overall, feature selection helped simplify the models by reducing the number of features, while maintaining strong performance. The slight decrease in accuracy and AUC confirms that the selected features were still highly effective for classification. This approach made the models more efficient and easier to interpret without compromising their predictive power.

VII. Model Comparison and Results

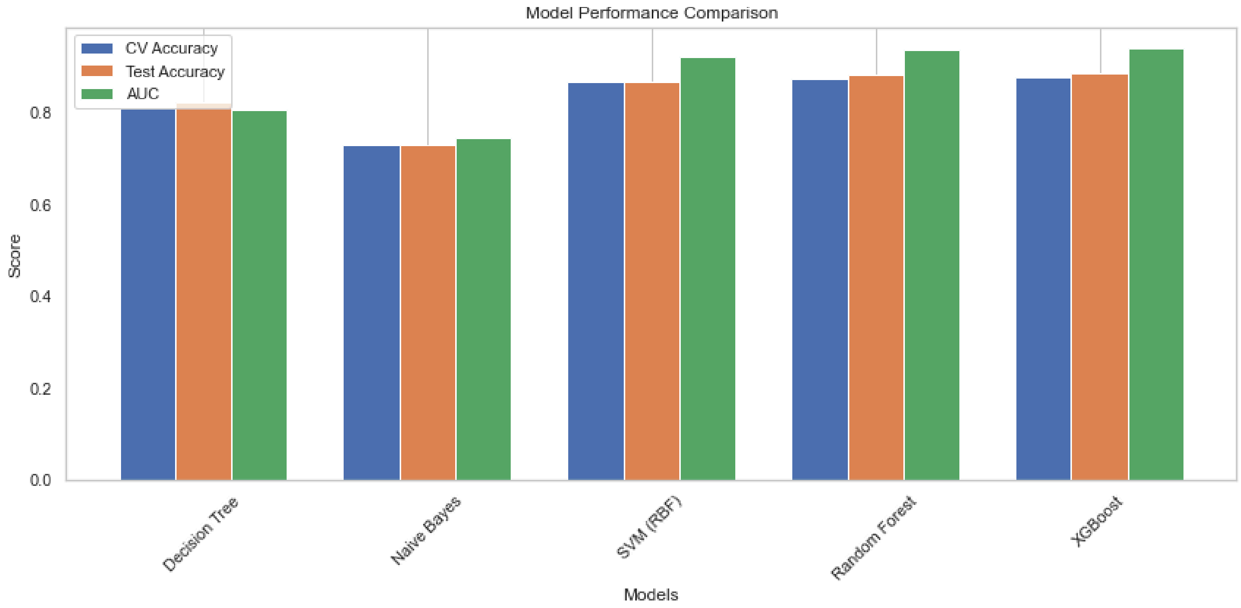
In this section, we compared the five data mining models: Decision Tree, Naive Bayes, SVM (RBF kernel), Random Forest, and XGBoost based on cross-validation accuracy, test accuracy, and the Area Under the Curve (AUC) metrics.

Initially, the Decision Tree achieved reasonable cross-validation accuracy (80.79% ±0.90%) and test accuracy (82.09%), with an AUC of 0.8054. Naive Bayes performed the weakest, showing lower accuracy (72.75%) and an AUC of 0.7446, indicating challenges in accurately identifying

Hadron events. The SVM model showed stronger results, achieving a test accuracy of 86.65% and an AUC of 0.9194, demonstrating good predictive ability.

The ensemble-based methods clearly performed best. The Random Forest model achieved strong performance, with a high test accuracy (88.33%) and an excellent AUC (0.9351). XGBoost slightly outperformed Random Forest, providing the highest test accuracy (88.47%) and the best AUC (0.9381), highlighting its exceptional ability to distinguish Gamma from Hadron events.

A comparison of all metrics, cross-validation accuracy, test accuracy, and AUC is provided in the summary graph below. This visual clearly emphasizes the superior performance and reliability of the ensemble methods, particularly XGBoost, compared to simpler methods like Decision Trees and Naive Bayes.



Overall, our analysis highlights the superior performance of ensemble methods, particularly XGBoost, for this classification task. The results show that XGBoost outperforms simpler models like Decision Trees and Naive Bayes, making it the most effective choice for accurate particle classification.

VIII. Discussion and Challenges

While working on this project, we encountered several challenges that helped us better understand the data and improve our modeling approach. One key issue was class imbalance. Since the dataset contained more Gamma events than Hadron events, some models, especially Naive Bayes, tended to favor the majority class. This resulted in many false positives for Hadron predictions. Ensemble models like Random Forest and XGBoost handled this better, producing more balanced and accurate results.

Another limitation we noticed with Naive Bayes was its assumption that all features are independent. In reality, many of the features in our dataset are correlated. This mismatch between assumption and data likely reduced the model's accuracy and made it less effective compared to more flexible classifiers [3].

We also found negative values in features like fAsym and fM3Long confusing at first. After further research, we understood these values reflect directionality and asymmetry in the particle showers, not data errors. This insight improved how we interpreted the input features.

Lastly, we realized that accuracy alone wasn't enough to evaluate our models. We focused on metrics like precision, recall, and AUC to better understand performance, especially when minimizing false positives was critical for Hadron classification.

IX. Conclusion

In this project, we applied various data mining techniques to classify high-energy particle events using the MAGIC Gamma Telescope dataset. The main objective was to differentiate between

Gamma-ray signals and Hadronic background events using real-valued features extracted from simulated particle showers, advancing our understanding of astrophysical phenomena.

We started with exploratory data analysis (EDA) to understand the feature distribution and relationships. Some features had negative values, which initially raised concerns. Upon closer inspection, these negative values provided insights into the directionality and asymmetry of particle showers, guiding our preprocessing and model development.

We trained and evaluated five data mining models: Decision Tree, Naive Bayes, SVM, Random Forest, and XGBoost. Random Forest and XGBoost consistently outperformed the others, with XGBoost delivering the highest test accuracy (88.47%) and AUC (0.9381), establishing itself as the top model.

Feature selection techniques like F-score and RFE helped simplify the models by identifying key features such as fAlpha, fLength, and fWidth. Even with only selected features, the best models maintained strong performance with minimal reductions in accuracy and AUC. These results suggest that ensemble models, particularly XGBoost, are effective for classifying Gamma-ray events and can support future astrophysics research.

X- References

[1] Barbuzano , J. (2024, September 5). *Magic in the Air*. Sky & Telescope. https://skyandtelescope.org/sky-and-telescope-magazine/magic-in-the-air/

[2] Bock, R. (2007, April 30). *Magic Gamma Telescope*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope

[3] *Naive Bayes algorithm in ML: Simplifying classification problems*. Turing. https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners#disadvantages-of-a-naive-bayes-classifier