Analysis of the MAGIC Gamma Telescope Data Set

Final Project for MALIS course

December 14, 2021

Simone PAPICCHIO *simone.papicchio@eurecom.fr*
Daniele FALCETTA *daniele.falcetta@eurecom.fr*
Francesco CAPANO *francesco.capano@eurecom.fr*

December 14, 2021

Contents

1 Dataset Overview

The data present in this dataset are generated to simulate registration of high energy gamma particles in a ground-based atmospheric.
Cherenkov gamma telescope (CTA) observes high energy gamma rays, taking advantage of the radiation emitted by charged particles produced inside the electro-magnetic showers initiated by the gammas, and developing them in the atmosphere.
More precisely, when the gamma rays reach the earth's atmosphere they interact with it, producing cascades of subatomic particles. These cascades are also known as air or particle showers.
Light travels 0.03 percent slower in air, thus these ultra-high energy particles can travel faster than light in air, creating a blue flash of "Cherenkov light". Although the light is spread over a large area (250 m in diameter), the cascade only lasts a few billionths of a second.
CTA's large mirrors and high-speed cameras will detect the flash of light and image the cascade generated by the gamma rays for further study of their cosmic sources allowing reconstruction of the shower parameters.
The available information consists of pulses left by the incoming Cherenkov photons on the photomultiplier tubes.
Depending on the energy of the primary gamma, a total of few hundreds to some 10000 Cherenkov photons get collected, in patterns (called the shower image), allowing to discriminate statistically those caused by primary gammas (signal) from the images of hadronic showers initiated by cosmic rays in the upper atmosphere (background).
The goal of the classification is to correctly classify the gamma signal (g) from hadron (h, background).

2 Dataset Analysis

The first step of the project is the data analysis, to do so are used different representation of data, according to the information that need to be extracted. The results are showed below *Table 1*

Attribute	**Type**	**Description**
fLength	*Real*	major axis of ellipse
fWidth	*Real*	minor axis of ellipse
fSize	*Real*	10-log of sum of content of all pixels
fSize	*Real*	ratio of sum of two highest pixels over fSize
fConc	*Real*	ratio of highest pixel over fSize
fConc1	*Real*	distance from highest pixel to center, projected onto major axis
fAsym	*Real*	3rd root of third moment along major axis
fM3Long	*Real*	3rd root of third moment along minor axis
fM3Trans	*Real*	angle of major axis with vector to origin
fAlpha	*Real*	distance from origin to center of ellipse
fDist	*Categorical*	g gamma (signal), h hadron (background)

The dataset is composed of **19020** distinct records with **11** attributes. The following table shows the labels.

The dataset does not contain missing values.

From the boxplot in Fig. 1 is possible to notice that the range of values for each feature is different. Therefore, in order to not have bias in the model towards higher/lower values, we need to scale the dataset. Moreover, some "distant" measurements are visible in the Figure 1, but we do not know if they are measurement errors or some possible outliers.

Interesting deduction can also been made from the pairplot in Fig. **??**. On the diagonal are plotted the kernel density estimate (KDE) figures. Almost all the features does not have a normal distributed KDE (*fM3Long*, *fM3Trans*, *fAlpha*, etc...). From the scatter plots, it is not possible to see any evident separation of classes. Moreover, in some scatter plots (for example *fSize-fAlpha*) the blue dots (Gamma) are either mixed with the orange dot (Hadron) or completely not visible.

Useful for the analysis is also to understand if there is some correlation between features. The correlation measurement used is the Pearson correlation (value between -1 and 1). it is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations: High ρ correlation means that the two features are linearly correlated, that is, to be related in such a manner that their values form a straight line when plotted on a graph. In Figure 3 is shown, as heatmap, the correlation between features. The features more correlated are:

- **fConc**: [continuous] ratio of sum of two highest pixels over fSize
- **fConc1**: [continuous] ratio of highest pixel over fSize

We could reasonably expect this high ρ by their definitions. We can safely remove *fConc1*.

Crucial is also to understand if the dataset that we are analyzing is balance or imbalance. It is “balanced” when it contains equal or almost equal number of samples from each class, “unbalanced”, as in this case, if the samples from one of the classes outnumbers the other. Figure **??** clearly shows the imbalance of the dataset:

- g = gamma (signal): 12332 => 65%
- h = hadron (background): 6688 => 35%

This kind of skewness towards the label g ma cause problems during the training. Some classifiers, such as Random Forest, fail to address this kind of problem as they are sensitive to the proportions of classes, i.e they tend to favor the class with the largest proportion of observations (majority class).

3 Training Pipeline

1. **Outlier Removal**: Since there are some "distant" measures that may be measurement errors or relevant outliers, we did not manually remove them but we use the *Local Outlier Factor*
2. **Robust Scaler**: typically, this is performed by removing the mean and scaling to unit variance. However, with the presence of the outliers the sample mean / variance can be influenced in a negative way. In such cases, the median and the interquartile range often give better results.
3. **Stratified Training/Test split**: The split must to be stratified otherwise we loose the possibility to replicate the real word with the test dataset.
4. **SMOTE**: We have seen that the dataset is imbalanced towards the class Hadron (65% g and 35%h). There are three solutions: Undersampling the majority class (Gamma), Oversampling the minority class (Hadron) or leave the dataset imbalanced. Undersampling the majority class in this case it is nor recommended because the goal of the classification is not only to correctly classify Gamma but

also minimizing the error on Hadron since classifying a background event (h) as signal (g) is worse than classifying a signal event a background. For this reason, it is better to keep as many records as possible with label (Gamma).For the same reason of above, the third solution has to be excluded since leaving the dataset imbalanced towards g will tend to favor this class respect to h, i.e. leading to more wrong prediction for h.Oversampling the majority class, instead, seems to be the correct choice for this dataset. The selected algorithm for oversampling is the Synthetic Minority Oversampling Technique (SMOTE).

5. **Stratified Cross Validation**: Also in this case it is crucial to select a stratified version of the CV. In addition, only the training set is oversampled (using SMOTE), instead the validation one is untouched in order to have a correct estimation of the true error.

6. **RandomGridSearch**: For our experiments we run a random grid search with several parameters

7. **Random Forest**: It is the first non linear model that we tried. We selected it for its versatility and interpretability.

8. **Ridge for Classification**: We used this simple but powerful linear model with the L2 regularization.

9. **Logistic Regression**: Since this is a binary classification problem, the logistic regression was our starting point

The results are shown in the Appendix

4 What we will do

We want to try the following model: **MLP**, **SVM**, **Gaussian Kernel SVM**, **Polynomial Kernel SVM**
In addition, we want also to create the learning curves to better understand how well our models are generalizing.

5 Contributions

All of us have given a contribution in discussing about the dataset, selecting the best algorithms, and coding.

6 Appendix

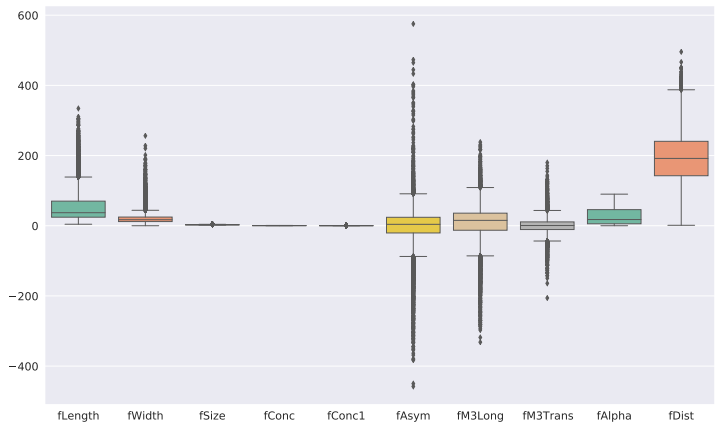


Figure 1: Boxplot of the features compared.

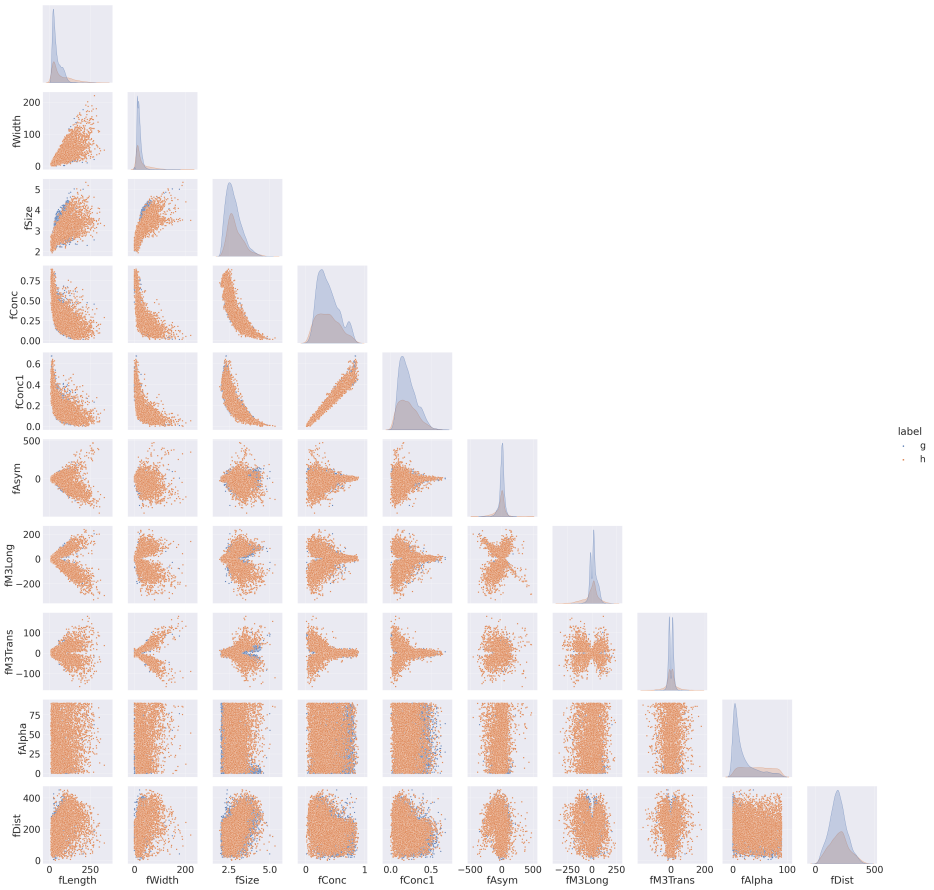


Figure 2: Pairplot between features. On the diagonal there are the KDEs plot. On the lower triangle there are the scatter plots between the selected two features. All the plots use the class label as semantic variable that is mapped to determine the color of plot elements.

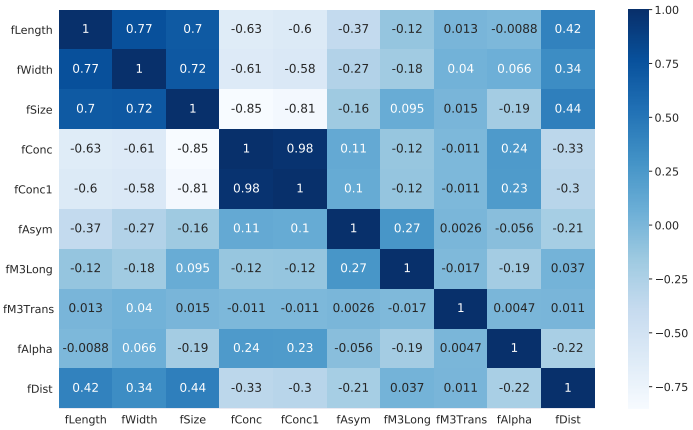
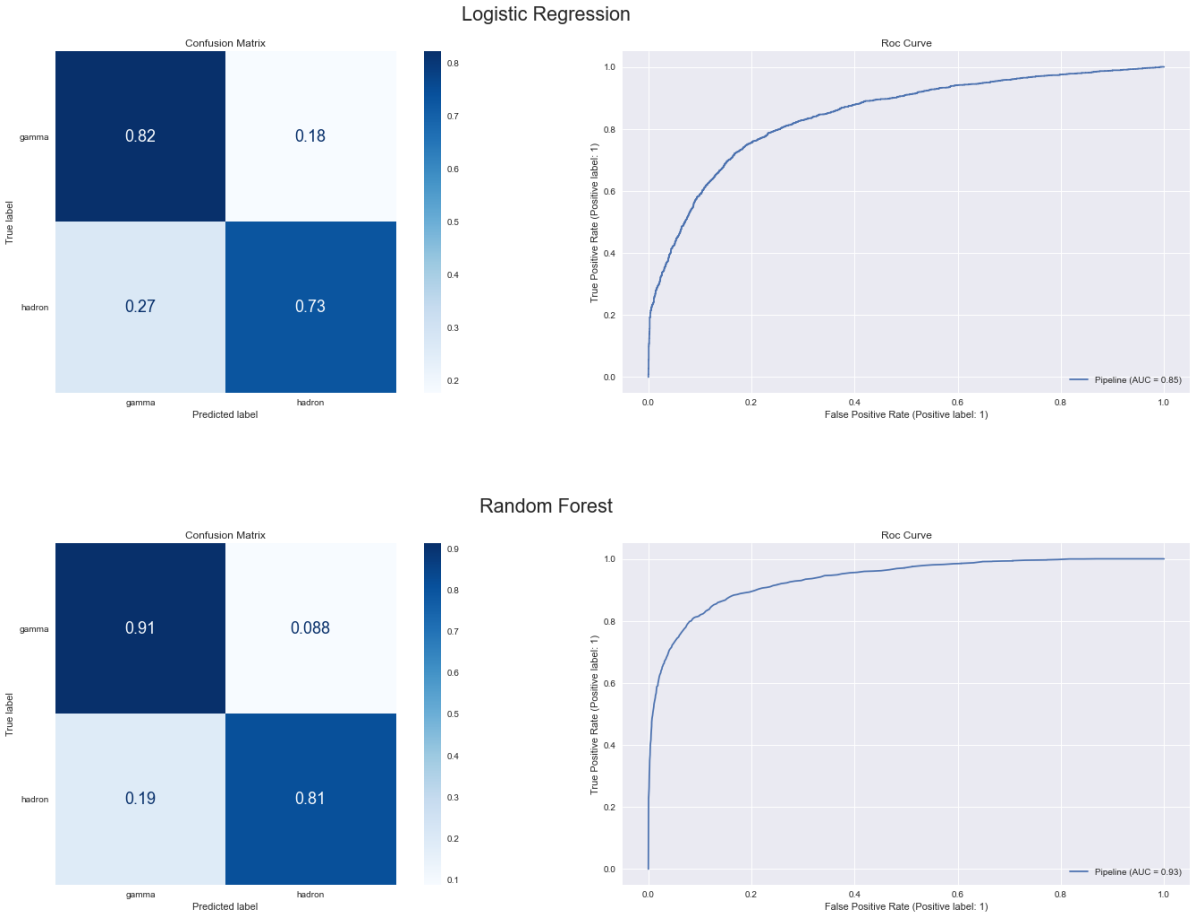
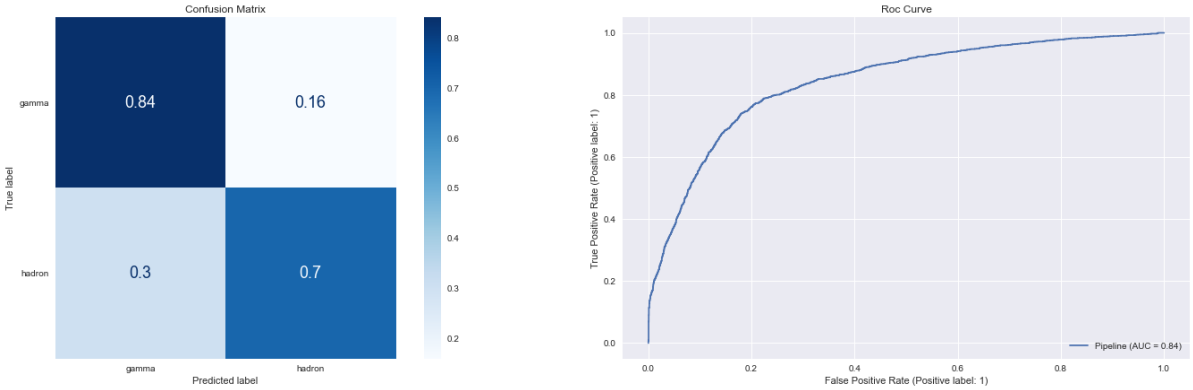


Figure 3: Heatmap of the correlation features using the Pearson correlation.



Ridge


Confusion Matrix
gamma
hadron
0.84
0.16
0.3
0.7
True label
Predicted label
gamma
hadron
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Roc Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Pipeline (AUC = 0.84)
0.0
0.2
0.4
0.6
0.8
1.0