**MODEL SUMMARY REPORT**

1. Introduction:

- The aim of the problem is to study and analyze very high-energy gamma-ray emissions from celestial sources. The MAGIC telescope is a ground-based Cherenkov telescope designed to detect gamma rays, which are high-energy photons originating from astrophysical objects such as stars, galaxies, quasars, and other cosmic sources.

The primary objective is to identify and classify the sources of gamma rays into gamma (signal) or Hadron (backgrouong) and study their properties. By observing and analyzing the gamma rays, scientists aim to gain insights into the underlying astrophysical processes and phenomena, such as:

* Identifying and characterizing gamma-ray sources: The dataset helps in identifying and understanding the nature of celestial objects that emit gamma rays.
* Studying high-energy astrophysics: By analyzing the gamma-ray emissions, researchers can study processes involving high-energy particles and extreme environments in space.
* Probing gamma-ray bursts: The dataset may include observations of gamma-ray bursts, intense and brief emissions of gamma rays often associated with supernovae or other cataclysmic events.
* Investigating dark matter and cosmic rays: Gamma rays can provide clues about the distribution of dark matter and the origin of cosmic rays in the universe.

- Original owner of the database:

R. K. Bock

Major Atmospheric Gamma Imaging Cherenkov Telescope project (MAGIC)

http://wwwmagic.mppmu.mpg.de

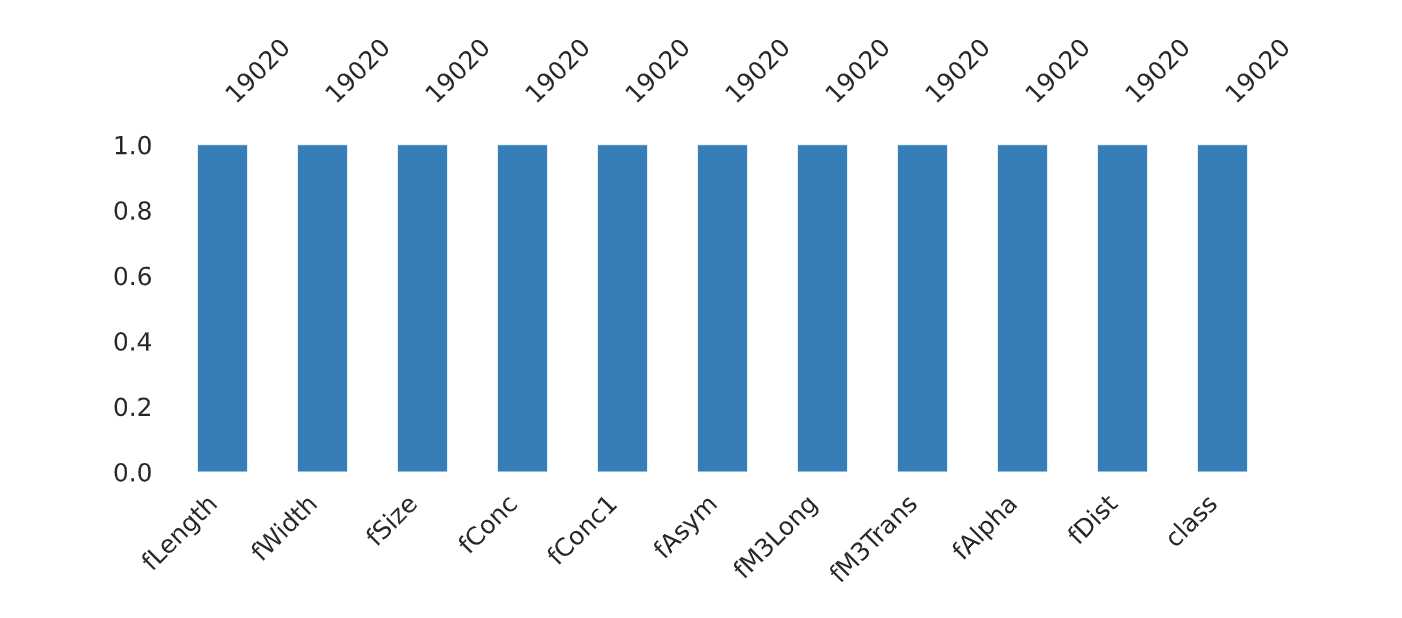
- For this task this dataset is being taken from UCl machine learning repository.

- We are basically going to perform Classification and Clustering algorithms to understand the pattern in the dataset and classify the instances into the correct category using machine learning techniques.

2. Data Preprocessing:

- Data cleaning:

* Handling missing values: we have no missing or null values in our data.



* Duplicates: Only 2 Duplicate rows are present in the data.
* Outliers: Outliers are also not present, and it is backed by a very good explanation and the dataset provides the information about various features of an astronomical ray and rays can having varying features value with some reasonable difference in values.

- Feature engineering:

Originally, we have 10 features and one target column in the dataset namely:

1. fLength: major axis of ellipse [mm], continuous

2. fWidth: minor axis of ellipse [mm], continuous

3. fSize: 10-log of sum of content of all pixels [in #phot], continuous

4. fConc: ratio of sum of two highest pixels over fSize [ratio], continuous

5. fConc1: ratio of highest pixel over fSize [ratio], continuous

6. fAsym: distance from highest pixel to centre, projected onto major axis [mm], continuous

7. fM3Long: 3rd root of third moment along major axis [mm], continuous

8. fM3Trans: 3rd root of third moment along minor axis [mm], continuous

9. fAlpha: angle of major axis with vector to origin [degree], continuous

10. fDist: distance from origin to centre of ellipse [mm], continuous

11. class: g,h gamma (signal), hadron (background)

But after calculating the correlation matrix we have dropped some columns (features) which showed negative or negligible correlation with target feature, those features are 'class','fAsym','fConc1','fM3Long'.

- Data splitting: The dataset after dropping the above features is being divided into training and testing data in a 70:30 ratio.

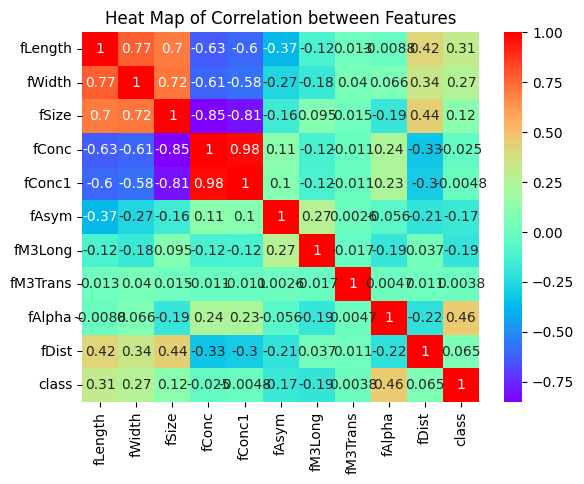
3. Exploratory Data Analysis (EDA):

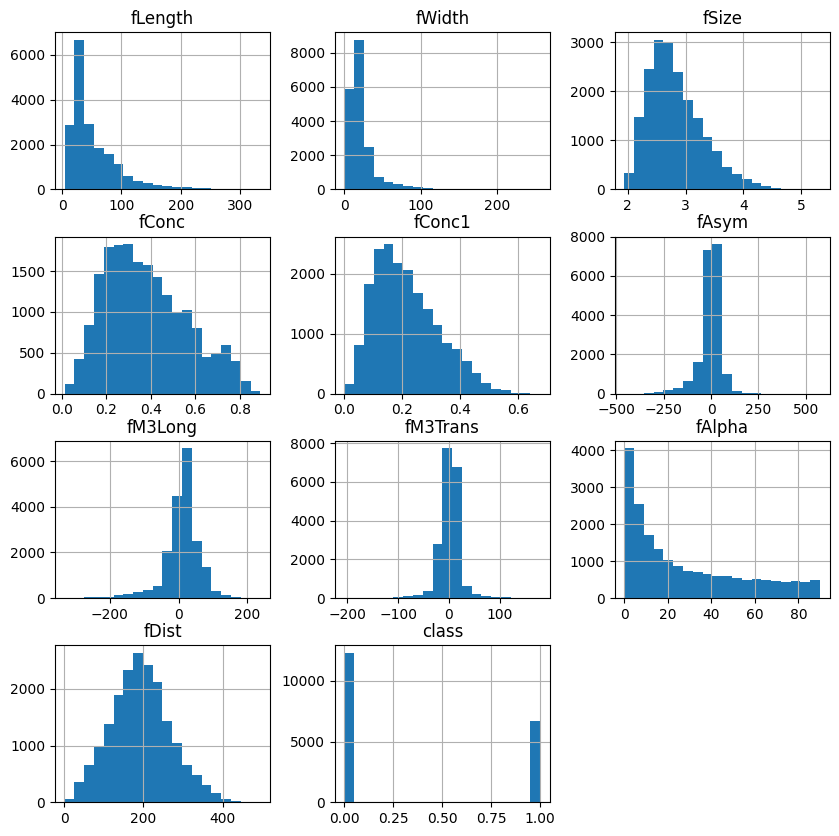
- A very detailed EDA report is being shared in the assignment folder as well as it can be accessed in the Api with ‘/report’ header to the URL.

- Some required Data Visualization are also being shown in the notebook including the correlation heatmap, Scatter plot of complete data, Histogram plots of features with respect to one another, line plots, etc.

- Insights gained from EDA:

* We have a good distribution of feature values which is good for training the to capture more meaningful relationships.
* flength, fWidth, fSize show good corelation between each other, which is obvious.
* fConc and fConc1 also shows similar relationship.
* 'class','fAsym','fConc1','fM3Long' are some features which have no useful relationship with the target features.





4. Model Selection and Building:

- For Classification: Logistic Regression, KNN algorithm, Support Vector Machine, Xgboost, Decision Tree Classifier and Random Forest Classifier.

For Clustering: K-means Clustering, DBSCAN Clustering, Gaussian Mixture Model Clustering, Agglomerative Clustering.

(The use of multiple models is to extract out every possible relationship between the features as well as get the best model having the highest performance).

- Based on the results of the models used for classification, Random Forest Classifier gives the best results. Therefore, I have performed fine- tuning on this model on the following hyperparameters: number of estimators, max depth and minimum sample split. Then using Grid Search CV obtained the best values of this hyperparameters.

5. Model Evaluation:

- The simple classification accuracy is not meaningful for this data, since classifying a background event as signal is worse than classifying a signal event as background. For comparison of different classifiers an ROC curve must be used. The relevant points on this curve are those, where the probability of accepting a background event as signal is below one of the following thresholds: 0.01, 0.02, 0.05, 0.1, 0.2 depending on the required quality of the sample of the accepted events for different experiments.

- In case of Clustering the performance metrics used id Silhouette score. The silhouette score is a numerical value that quantifies how well the data points are clustered. It helps to assess the compactness of clusters and the separation between different clusters.

- Evaluation results:

* For Classification:

Logistic Regression: 0.84

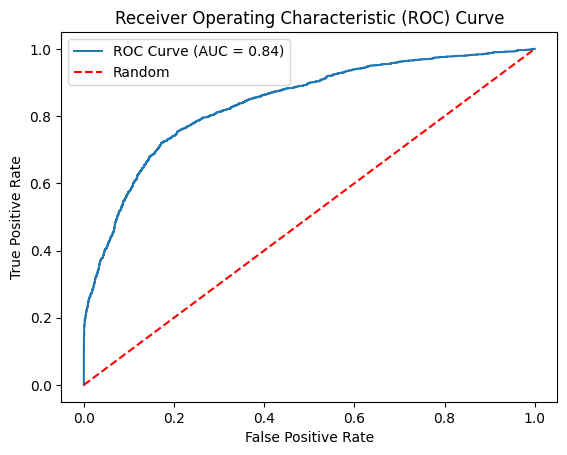
Support Vector Machine: 0.8685612511243068

Decision Tree Classifier: 0.8265905473883963

Random Forest Classifier: 0.9554362974621425

Xgboost: 0.9454338423642469

Ensemble Learning: 0.8430481115717445



* For Clustering:

K-means Clustering: 0.39485505127614073

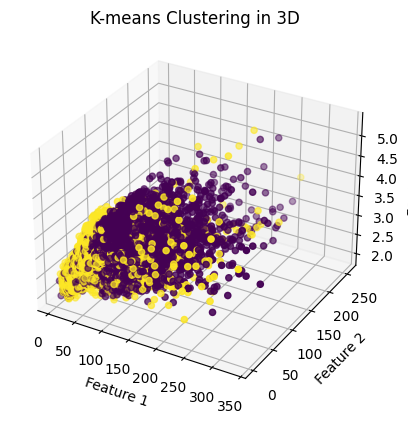
GMM Clustering: 0.39485505127614073

Agglomerative Clustering: 0.33229891425305813

- For this problem the true positive rate should be high and false positive rate should be as low as possible due to the sensitivity of the problem. Gamma classified as Hadron is very dangerous as compared to Hadron classified as Gamma. For all the models we have a good TPR which is a good sign.

6. Discussion and Interpretation:

* **Model Performance**: The model performs good in the classification task with the maximum roc\_auc\_score of 0.95, that means thus model can classify gamma rays for all the inputs given to the model. Clusters formed by performing unsupervised learning are not well separate with the best score of 0.39. The score is good to some extent but not much reliable for a data insight purpose.
* **Feature Importance**: fAplha has the most impact on the target class of the rays likewise in the training of the model, but the model is not solely based only on this feature. Other features are also to be considered will model training to get a more generalized and robust model.
* **Class Imbalance**: There was a big class imbalance with the gamma target count of almost double the count of hadron. But this is for a reason because the model must classify the gamma rays with the highest accuracy as per the delicacy of the model purpose.
* **Overfitting or Underfitting**: No observed overfitting or underfitting.
* **Model Robustness**: The performance of classification model is good on unseen data and results are reliable. The model generalizes well.
* **Limitations**: The model have clustering limitations; the clustering results shows that the clusters are not well-separated.
* **Comparison with Baseline or Other Models**: The base model used is Logistic Regression, along with this I have used KNN, SVM, Decision tree, Random Forest and ensemble learning for classification training. Out of all this Random Forest Classifier gives the best performance. For clustering out of K-means, GMM, DBSCAN and Agglomerative clustering algorithm K-means gives the best results.



7. Conclusion:

- Summary of the results and model's performance: The model gives excellent results on classification task whereas for clustering the model fit the data into clusters with not much confidence I.e., the clusters are not well separated with the score of 0.39.

- The model is fit for classification task and have a great accuracy to classify gamma rays.

8. References:

- List of the resources, research papers, or libraries used in the project.

(a) Bock, R.K., Chilingarian, A., Gaug, M., Hakl, F., Hengstebeck, T., Jirina, M., Klaschka, J., Kotrc, E., Savicky, P., Towers, S., Vaicilius, A., Wittek W. (2004).

**Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope.**

(b) Libraries used: Numpy, Pandas, Matplotlib, Seaborn, sklearn, imblearn, Xgboost, pickle, mpl\_toolkits.

( c ) Dataset: UCL Machine Learning Repository.

9. Appendix:

- The code implementation along with the deployment code is shared in the assignment folder for any reference.