Samantha Albert
Guillermo Blanco
RT Frank
Emily Le
Timothy Tran

Classifying Signals of A Cherenkov Telescope

Introduction:

The Cherenkov Telescope Array (CTA) is a generational and forthcoming ground-based observatory used for exceptionally high energy gamma-ray astronomy.
Despite the fact that Earth's atmosphere works to prevent gamma rays from penetrating the surface, the interactions between these forces produce ultra high energy particles known as showers. Due to their high velocities, the particles create a flash of blue Cherenkov radiation. CTA's high-speed cameras as well as its innovative mirrors will capture and locate these flashes, allowing astronomers to detect the source of these cosmic outbursts.
CTAs entitle astronomers to observe the sky in higher energy resolution for the first time ever, enabling them to scan the atmosphere for dark matter particles and catch the burts as they explode rapidly.
The problem in our project is that we want to distinguish the difference between primary gammas and cosmic rays known as hadrons.All of the cosmic sources of gamma rays, like cosmic rays, black holes, leftovers from supernovas and dark matter are foundational parts of our galaxy and universe. Understanding these events and the high-powered radiation they emit is essential to all parts of astronomy.
As for the motivation for this project, the Cherenkov telescope is an up and coming innovation. It plays a significant role in the study of cosmic particles and the exploration of dark matter. Analyzing its work can help improve scientific knowledge and allow astronomers to learn more about space to further human progression.

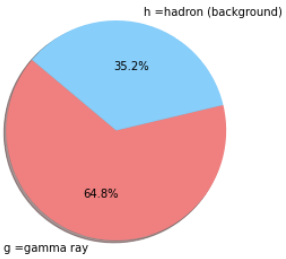
Data:

The dataset has a total of 19,020 rows including 10 features and 1 target which is the class. Since there were no missing values, there were no rows removed. The data was generated by using a Monte Carlo simulation, an algorithm that relies on repeated random sampling to get numerical results, to detect gamma ray particles in a Cherenkov telescope. Since the data set is generated using a simulation then the dataset set used is all synthetic. Within the dataset, there are 12,332 events of gamma ray signals and 6,688 events of cosmic rays or background signals. The dataset was cleaned by renaming the column names and changing the "g" to 0 and "h" to 1 in the class column with g meaning gamma and h meaning hadron.

Figure 1: Data

	fLength	fWidth	fSize	fConc	fConc1	fAsym	fM3Long	fM3Trans	fAlpha	fDist	class	ID
3187	81.4054	25.1619	3.5143	0.1989	0.1014	-51.2494	70.7019	-16.5938	2.3840	288.8320	g	1
15817	109.4600	18.5694	2.6830	0.4461	0.2251	-68.1987	-107.3240	10.6969	75.1910	182.7620	h	2
9560	18.0787	10.2187	2.4082	0.6523	0.3613	20.8493	11.8441	5.8978	24.1656	139.1020	g	3
12702	32.6417	13.4912	2.7948	0.4282	0.2285	32.1499	-17.2488	-4.8722	69.1740	165.9880	h	4
6326	57.1110	15.9127	2.5711	0.4295	0.2161	31.4437	62.9231	-15.8723	39.9411	145.2020	g	5

Figure 2: Pie chart comparing the gamma ray signals and background signals



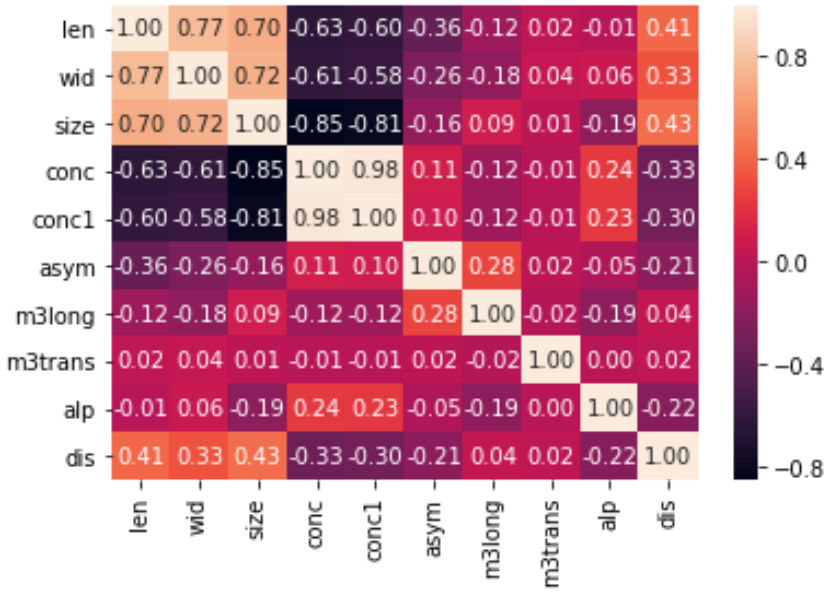
The 10 features of the dataset include length, width, size, concentration, concentration 1, asymmetry, M3long, M3trans, Alpha, and Distribution. The target of the dataset focuses on the class which include gamma ray signals which have bursts that are extremely short lived and hadrons, which for our analysis, are classified as background signals. Cherenkov telescope arrays are ground based and are pointed towards the sky. The data is received as an ellipsoid with the major axis centered on the focal point of the mirror array. The length is the major axis of the ellipse and the width is parallel to the film plane of the receivers. The way the data is interpreted is by a series of three images called moments, with size being the first moment, movement being the second moment, and temperature being the third moment. The size is a log-transformed value of how big large the burst was. Concentration and Concentration 1 are finding a point of the highest intensity within the burst. Asymmetry is the distance from the center of the ellipsoid to the point of the highest intensity. M3Long and M3Trans are the cube root of the transformed data temperature. Alpha is the angular distance from the point of right ascension, which is a fixed point in the night sky. The distance feature is the distance from the event origin. Lastly, the target variable, ‘class’, is denoted with a “g” for gamma signals and “h” for hadrons.

Exploratory Analysis:

Pearson’s Correlation Matrix

Looking at our correlation matrix, we see both strong positive and negative correlations. The strongest positive correlation is between “Conc” and “Conc1”, however, this is to be expected. Both of these values are ratios of the intensity and size of each recorded entry with slightly different calculations. Our hypotheses are most interested in the positive correlations between length and width of the burst.

Figure 3: Correlation Matrix

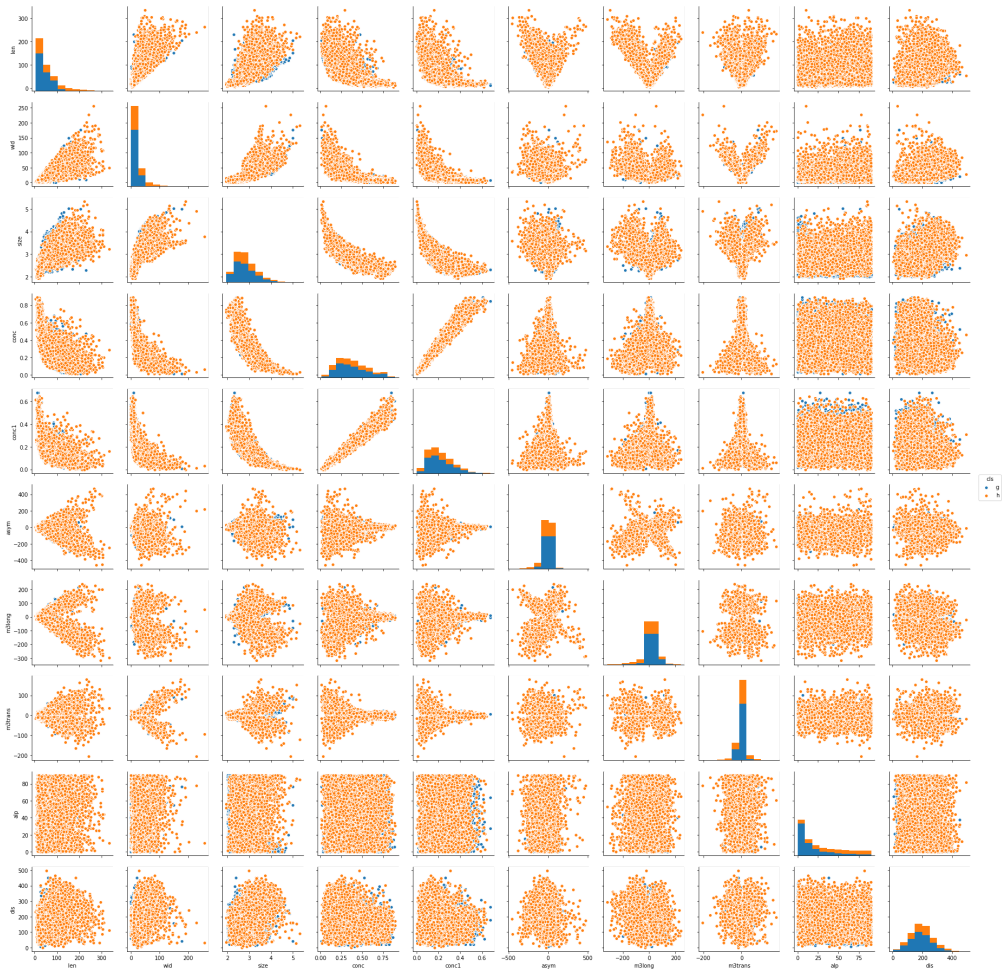


Pairwise plots

For pairwise plots, we observed lots of funneling and non-linear relationships between our variables. The clear linear relationship between the “Conc”, “Conc1” and “Size” are even more apparent, and are very clearly linear. As we mentioned before, the “Conc” variables are both ratios that are calculated based on the highest intensity and the size of the burst, hence their strong relationship.

We see the same strong correlation between the size, length and width of the data points in a mostly linear relation.

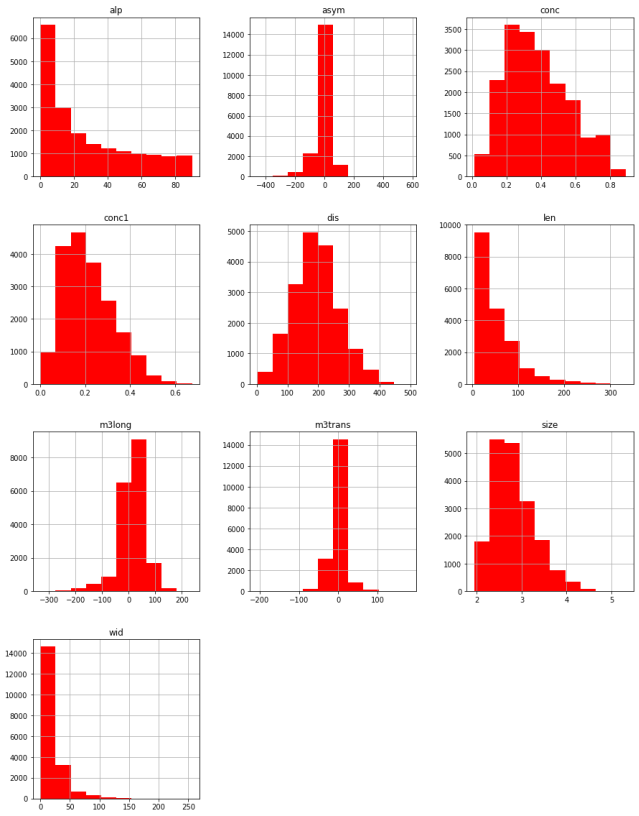
Figure 4: Pairwise Plots



Histograms

In conducting exploratory analysis here, we clearly see why there are so many transformations already applied to the data. There is heavy skewing present in many of the variables. What variables that are normalized are more evenly distributed, albeit some are extremely centered.

Figure 5: Histogram

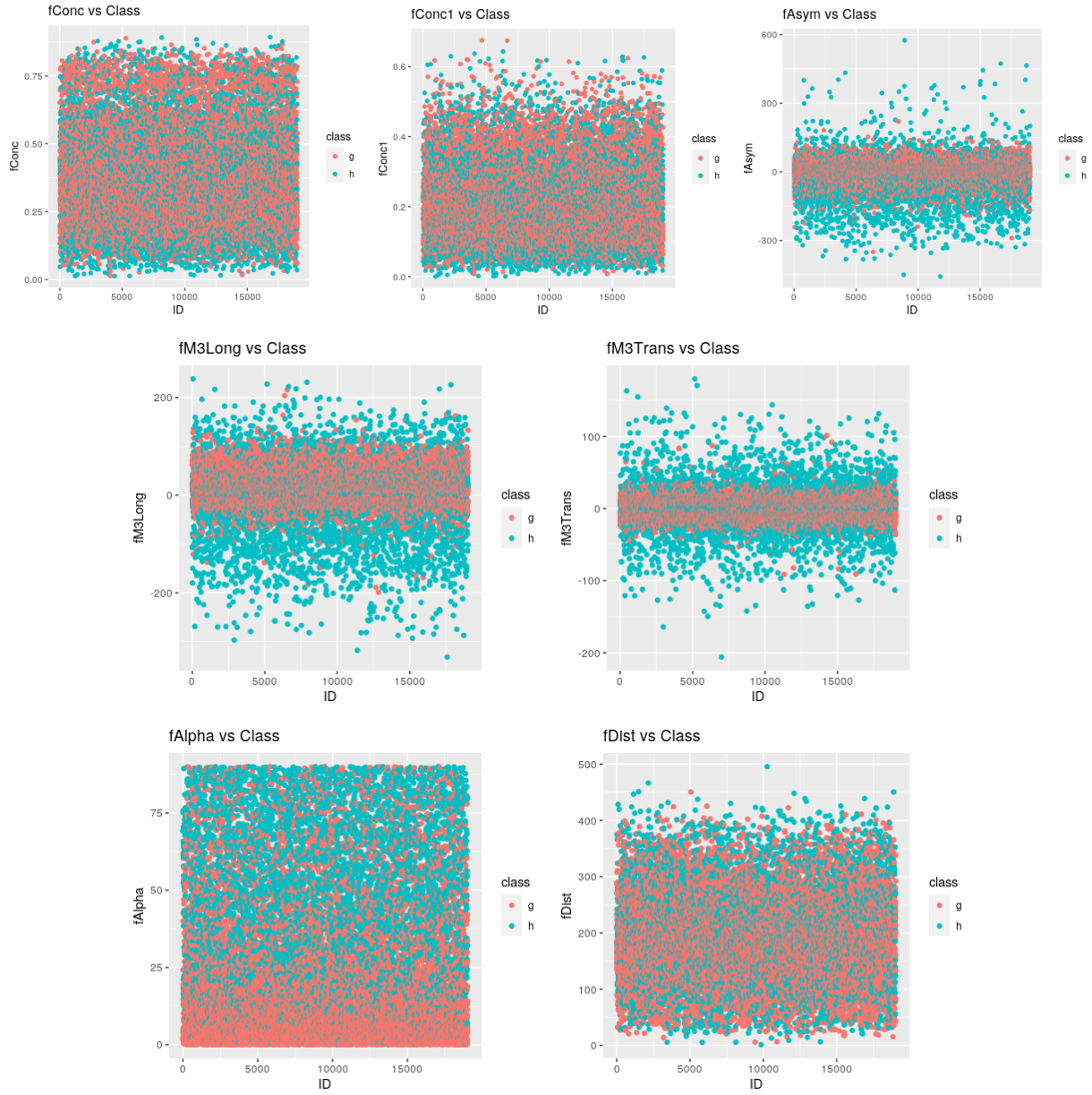


Scatterplots

The dataset was shuffled and assigned an ID, then using the scatterplots, we compared the class to all the features of the data. We can observe that all the variables were not linearly separable, so we were not able to create a line separating the colored points.This can be seen by the overlap of points.

Figure 6: Scatterplots

Classifying Signals of A Cherenkov Telescope



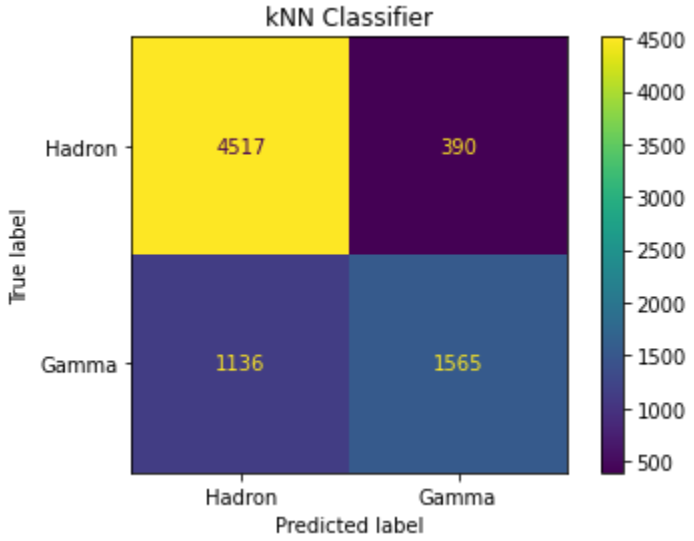
Hypotheses

1. Length, width, and alpha seem to be the most correlated with class.
2. Most gamma points overlap with hadron points, but not necessarily the reverse, so there will likely be confusion in the k-Nearest Neighbors model.
3. Through SHAP analysis, each model should show the same most-weighted factors.

Modeling:

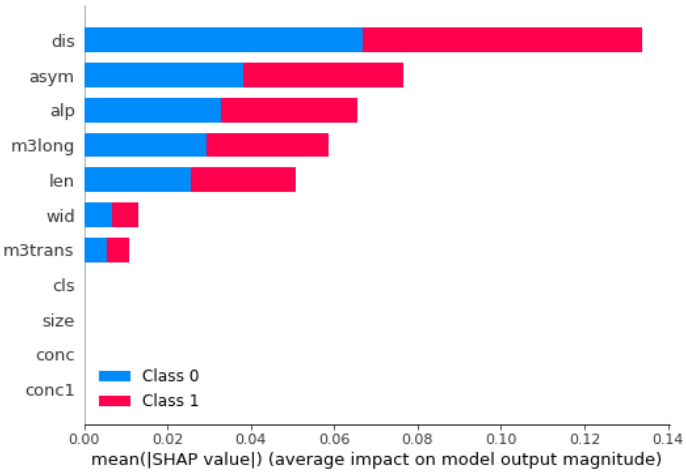
K-Nearest Neighbors

Before deciding the number of groups to use, we ran an exploratory analysis of the model to test the optimal number of groups. We tried lower group numbers, and found that, while more accurate in predictions for the test set, the scoring suffered for the testing data. We settled on 8 total groups, it had high scores in both training and testing, with a minimal distance between the two. The accuracy score overall with 8 groups was 80.415. Below is the confusion matrix for this model.

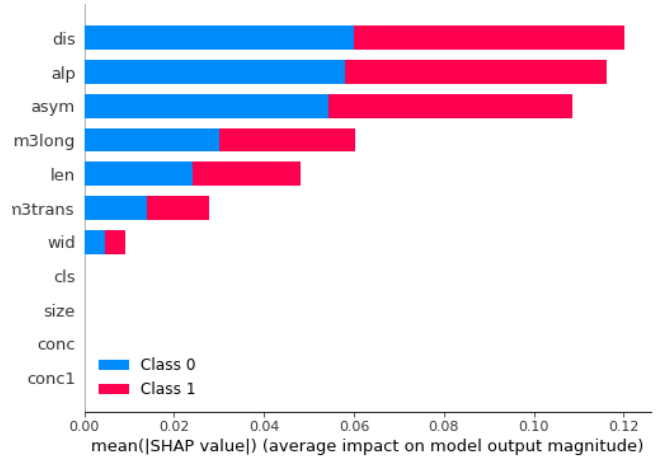


We found that, for the manner in which our data was presented, KNN is a good fit for the style of modeling. We had a large data set, but a small number of features and a binary state of classification. Ultimately, for the simplicity of KNN, it works best for exploratory analysis such as ours. It quickly identifies the target groups, but does not do a great job of providing deep analysis.

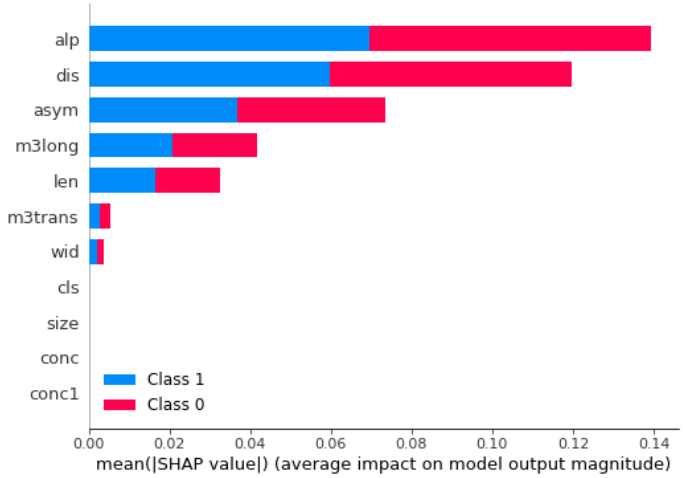
For our Shapley analyses, we realized that the SHAP package is very resource-intensive and would take an extremely long time to run for our full dataset of over 19,000 entries. Thus we decided to run the analyses on just samples of the data for each of the models. First we ran three analyses on samples of n=120 using different random states to see how much the analysis outcome was affected by which random state was used. Below are the results. Class 0 indicates hadron and class 1 indicates gamma. The number in parentheses indicates the random state used for that sample n.



Accuracy score for n(15) = 0.7292.

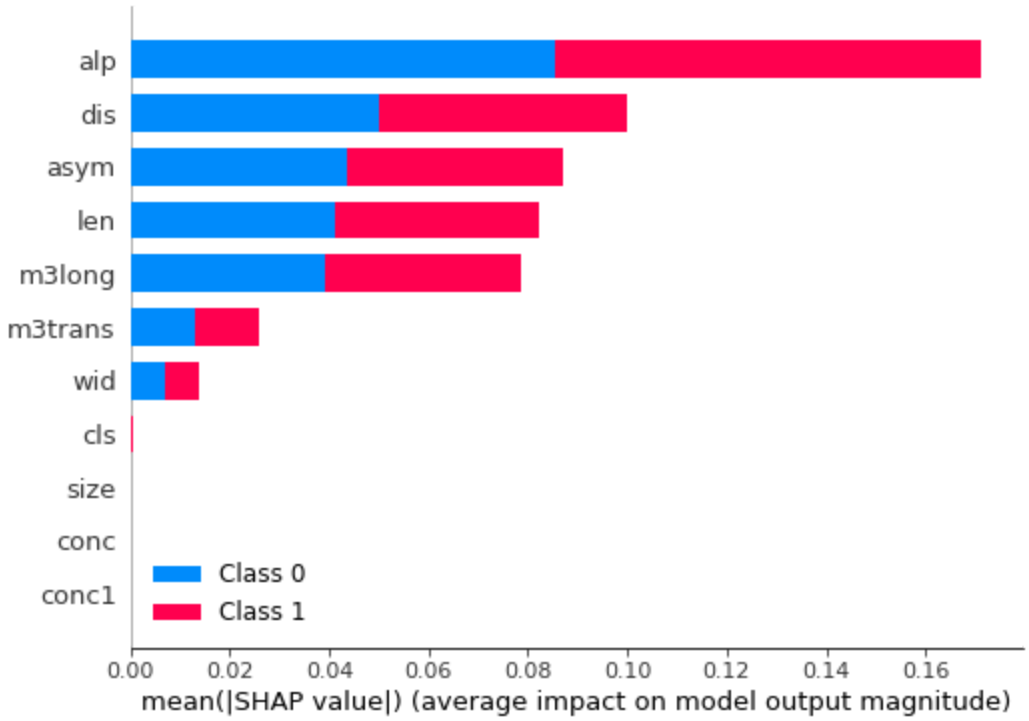


Accuracy score for n(66) = 0.6667.



Accuracy score for n(73) = 0.5417.

When conducting a Shapley analysis on our KNN model for our final model, we used a sample of n(9) = 500.

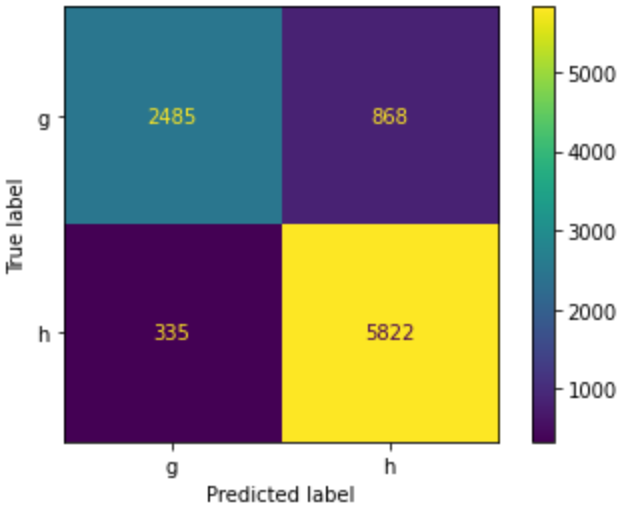


Our sample had an accuracy score of 73: with the alpha and distance values having the highest importance in our model.

Neural Network

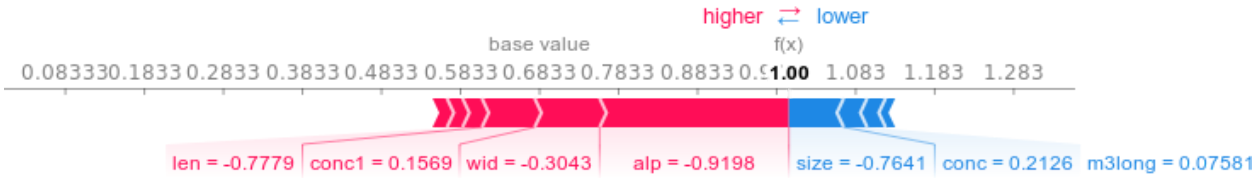
We used a neural network with three hidden layers, of sizes 8, 6 and 5. Half of the data was randomly chosen for the neural network's training, and the other half to evaluate its performance. We capped the number of iterations during the training phase at 5000. Experiments showed that increasing the number of iterations during the training did not increase the neural network's performance significantly.

Overall, our neural network model achieved an accuracy score of 0.87. The confusion matrix is shown below:



The SHAP analysis revealed that the feature with the greatest impact on the model output was the alpha value, which is in accordance with the results obtained with other models. Other features such as size and width also had a significant effect on the model output. The SHAP diagram we obtained is:

And the corresponding force diagram is the following:

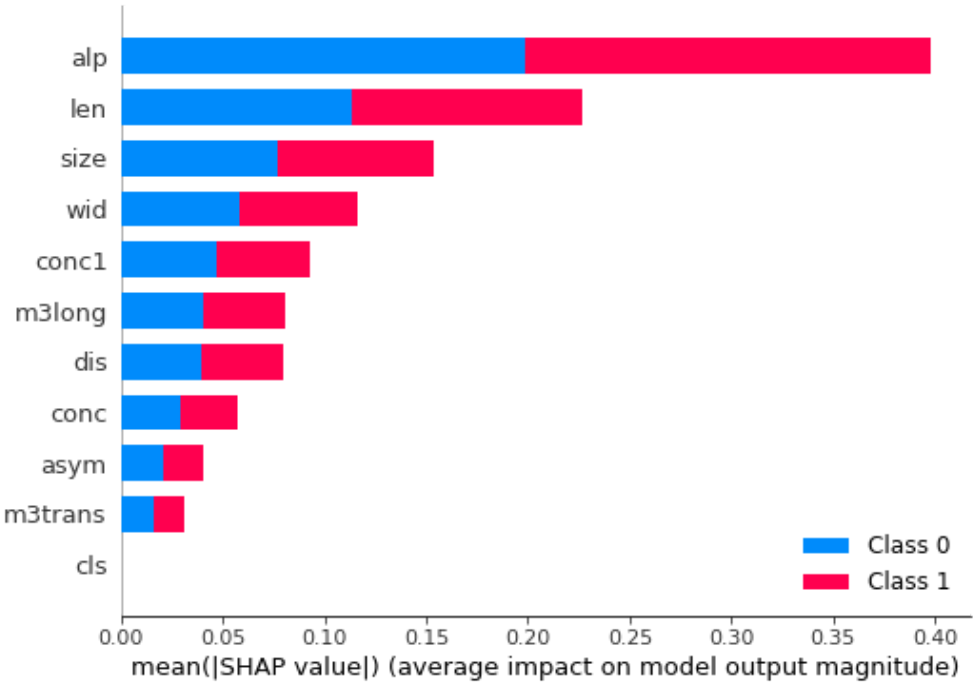


Decision Tree

For our Decision Tree model, we didn't specify a maximum depth so that the model could go until gini=0 (i.e. perfect classification). Using the entire dataset, the Decision Tree returned a model of accuracy score 80.54, very similar to the score from our KNN model. Below is the confusion matrix for the Decision Tree classifier.

As one can see from the confusion matrices, our Decision Tree model was more accurate than our KNN model, but less accurate than our Neural Network.

We then ran a SHAP analysis on the same n=500 sample that we used for our KNN model and Neural Network. The results are shown below. Again, alpha has the highest impact on model output, but was followed by length and size instead of distance and asym (KNN) or size and width (Neural Network). Running our Decision Tree classifier on just the sample returned an accuracy score of 73.2, which is virtually equal to the accuracy score returned by the sample using KNN.



Discussion:

As previously mentioned, our neural network achieved an accuracy score of 0.87. Similarly to our other models, the features that had a greater influence on the neural network's output were the alpha value, size and width.

Previously we had achieved a slightly worse performance of 0.84 by using a network composed of three hidden layers of sizes 20, 40 and 20. However, after following the empirically validated practices described in this link, we found out that using an even smaller number of nodes speeds up the training of the network and improves its performance.

A potential limitation of this model is that the neural network's size and the number of layers might not be optimal. Similar to how we found an improvement on our previous model by following the advice mentioned above, it is possible that choosing another network topology might lead to further performance enhancements. More testing is required to assess whether the selected structure is close enough to optimality for this particular problem.

In addition, a factor that might affect this model's accuracy is that our data is unbalanced: out of 19020 records, 12332 (around 65%) were gamma. This imbalance could potentially introduce some biases into our model.

Furthermore, the dataset was all synthetic, created by the Monte Carlo simulation, therefore it may not be applicable for scientific findings. Due to the dataset being synthetic, the amount of hadron events were underestimated, leading to skewed model testing and analysis. Another limitation is that SHAP analysis on the entire dataset would take too much time. However, as we

saw with our KNN models, the random state that we use for our sample has a big effect on the SHAP outcome.

Conclusion:

We found from our analysis that alpha was actually the most significant factor in determining model output, followed by either length, size, or distance, depending on the model. As we expected, there was some confusion in our K-Nearest Neighbors model regarding false hadron signal predictions. Lastly, we found that SHAP analysis outcomes were heavily impacted by which random state was chosen for the sample, but when the analyses ran on the same sample, there was agreement that alpha had the highest average impact for all three models.

For the future, the model would need to be further trained before using it for real world data. One of the largest hurdles to clear is preparing the models for higher rates of hadron noise. It is important to remember though the purpose of our model: simply sifting through and finding the gamma signals. Each identified instance would have to be examined or passed through and analyzed by another AI.

Acknowledgements:

Samantha did the Decision Tree classifier, along with its confusion matrix and SHAP analysis.
Guillermo worked on the Neural Network model and the website.
Timothy worked on the background research, introduction(problem and motivation), along with running codes for the exploratory analysis for the dataset (Pairwise plots, Histograms, and Scatter Plots).
Emily worked on the data portion of the project as well as running the code for the correlation matrix in the exploratory analysis.
RT helped to tidy data and did KNN model and the accompanying Shapley analysis.

Relevant Links

Repository with the code: https://github.com/spa494/sds_project

Website: https://gjblanco-ut.github.io/ (future work includes choosing a different URL)

Bibliography:

Cherenkov Telescope Array Observatory gGmbH. (2016). Cherenkov Telescope Array CTA — the World's Largest Ground-Based Gamma-Ray Observatory. ESO. Retrieved December 1, 2021, from https://www.eso.org/public/teles-instr/paranal-observatory/cta/.

Wolfgang Wild "Cherenkov Telescope Array (CTA): building the world's largest ground-based gamma-ray observatory", Proc. SPIE 10700, Ground-based and Airborne Telescopes VII, 107000X (6 July 2018); https://doi.org/10.1117/12.2313470

"MAGIC Gamma Telescope Data Set." UCI Machine Learning Repository: Magic Gamma Telescope Data Set, https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope.

jj_ (https://stats.stackexchange.com/users/80688/jj), How to choose the number of hidden layers and nodes in a feedforward neural network?, URL (version: 2018-05-31): https://stats.stackexchange.com/q/180052