York University
PHYS3130 W2025

Classifying Neutrino Flavor Based on Cherenkov Radiation Optical Signal In IceCube Neutrino Observatory Final Data Exploration and Applied Machine Learning Techniques

Joseph Cuzzupoli
`https://github.com/josephcuzz2004/Phys3130_JC`
April 16 2025

Abstract

I present the final exploratory data analysis (EDA) and applied machine learning results for the IceCube dataset, regarding methods of neutrino flavor classification. I utilize different statistical techniques ranging from Gaussian analysis, correlation matrices, random forest regression, k-means clustering, among others to group, classify and predict neutrino flavour from the interaction events detected by IceCube. In particular, I use the parameters of hits per unit of optical measurement to decipher the neutrino flavour based on key differences initially observed by simulations of these events. While results are not conclusive, it would seem that with a more thorough analysis, this method can be a useful means to classify neutrino flavour of events detected at IceCube.

Contents

1 Introduction

Neutrinos are one of the most illusive, yet intriguing particles predicted and discovered by the standard model of particle physics. Their dynamics and origins hold important information regarding the intense astrophysical environments that created them, such as black holes and quasars. Being able to observe these particles has proven to be quite challenging however, detectors such as IceCube have made it possible to detect and measure these neutrinos, as well as their interactions with other particles surrounding them. IceCube is an in-ice astrophysical neutrino observatory located in Antarctica. It is lined with 5160 digital optical modules made of borosilicate glass that detect light created by fundamental particle interactions in the ice (Observatory, 2025). When muon neutrinos, in particular, interact with these atoms, they can result in a decay where relativistic muons are released. Locally within the ice, the muons created move faster than the speed of light. When this occurs, similar to when moving at Mach speeds and a sonic boom is created, the muon creates a boom of light known as Cherenkov radiation (Cass, 2018). Why is all of this information useful? While extensive research has been conducted to actually detect neutrinos themselves, as well as piece together the tracks of their events through the IceCube observatory, one area which has remained relatively unexplored is actually being able to classify neutrino flavour by the optical signal they emit in the detector. By understanding which neutrino flavour is being observed based on the optical parameter, it is possible to gain insights into the proportion of neutrino flux attributable to cosmic events without the need for more complicated methods used presently. Techniques such as classifying neutrino point-sources using methods of coincident observations of neutrino flux with gamma-ray telescopes (IceCube-Neutrino-Observatory*, 2025), deep-learning and maximum likelihood analysis (Hunnefeld, 2021) and the ratio of observed neutrino flux from charged and neutral current interactions ((A. Palladino, 2015); (Steven W. Barwick, 2022)) serve as the current methods to **indirectly** classify neutrino flavour used to understand the cosmic events that created them. So what is the goal? The research question at hand is whether it is possible to classify the neutrino flavour that induces an event detected at IceCube using the optical signal parameter. How will I do this? It was discovered via simulations that the Cherenkov radiation signals that are emitted and detected by IceCube are unique to the neutrino that created them. Aspects of this signal such as the mean optical intensity and spread of the signal around its mean characterize the signal into three classes which can be understood as the following sources; a muon, electron and tau neutrino. Not only is using the optical parameter useful in classification, it is also useful to consider the hits per unit time of a signal detection - these will be discussed extensively in the later sections. The hits per unit time provide insights into the speed and intensity of the products of the neutrino interactions. When there are many hits per unit time, this would indicate a large ensemble of light sources which are detected in short succession. This may help to indicate the Cherenkov radiation signals associated with different flavours as they are traveling through the experiment. The statistical techniques that will be used in this investigation include Gaussian deductive statistics to describe the data in terms of their means and standard deviations, as well as machine learning techniques.

2 Methodology

The research paradigm that I will be using to complete this study is a positivistic paradigm, as the results will be objective and based on empirical observations made at the IceCube neutrino observatory. The analysis makes use of data collected using random sampling methods; hence, the subsequent data analysis is completed purely deductively (Cuzzupoli, 2025). As stated above, the research question being addressed is as follows; it is possible to classify the flavour of detected neutrinos in the IceCube neutrino observatory using the optical signal parameter. The primary objectives of this analysis are to be able to group the signals across all events into three clusters which are represented by each of the neutrino flavours. Due to the optical signal parameter being found to have unique patterns and signal characteristics for each of the neutrino flavours, the goal of differentiating these traits in the data can be realized. Furthermore, once these cluster groups have been identified, I would like to be able to predict the optical signal parameter based on predictors such as cartesian space location, hits per unit time, etc. To do this, I will be making use of a series of deductive statistical techniques and machine learning algorithms to specifically be able to group/cluster and predict the optical signal parameter. This analysis makes use of the libraries including Numpy (Harris et al., 2020), Matplotlib (Hunter, 2007), Scikitlearn (Pedregosa et al., 2011), etc. The machine learning algorithms that are implemented include the following. The first of which is Random forest regression, which is used on the data in several ways to model and predict the signal parameter that corresponds to the different neutrino flavours. Why have I chosen to use a random forest regression model? As will be seen in the proceeding sections, this data exhibits rather non-linear relationships and is complicated experimental data. When observing the different trends, one finds that the relationships do not show much linearity and also include many extreme value outliers. Since this data is physical data, it would **not** be best practice to disregard these extreme points, since it likely contains information regarding high energy activity. Furthermore, random forest regression is well suited for noisy signals as well as outliers. Based on the data, the ability to handle outliers will become paramount to the ability to predict the values of the optical parameter. Random forest regression is also optimized for use on continuous data, which is of course applicable in this instance. Finally, the random forest method reduces overfitting which will also become useful. With these considerations, it is clear the random forest regression method will be well suited for this analysis. How do I plan on using it? I first plan to use a random forest regression to be able to see how well the model can predict the hits/q parameter for all the data across the 10 events. This serves as a baseline for how well the predictors (including cartesian coordinates of hits, among others) can strengthen the model to predict the data. I then plan to use a k-means clustering algorithm to be able to group and characterize local clusters that presumably correspond to the different neutrino flavors. By successfully grouping these clusters into three distinctive groups, I will be able to show that there exists some observable difference in the signal parameter or hits/t that alludes to the idea that perhaps it was caused by different neutrino flavours. In this analysis, the scope of the research extends as far as being able to group the data into three categories, as well as being able to predict the signal parameter from these groups. If I am able to show signs that this is possible, then it could be possible to reconstruct or better understand the astrophysical events that created them, quantify signal to noise ratio of neutrino flux observed, among others. In this case, I am limited by the access to a larger parameter space as well as the number of sampling points. The Particle Physics Playground group (Particle-Physics-Playground, 2025), which created the dataset, only provided 10 events each with 844 rows and 7 columns (2 of which are the number of hits and indices). Thus, a larger parameter space, more events and more rows of sampled points would reduce some of the sample limitations. Furthermore, it would allow me to be able to observe a greater number of high energy points which may further help in identifying astrophysical muon neutrinos. I also do not have access to the advanced algorithms used in the research papers described in the literature, as well as their data to perform triangulation of results.

3 Exploratory Data Analysis (EDA):

As mentioned above, the IceCube neutrino observatory uses the method of tracking Cherenkov radiation tracks due to interactions involving passing astrophysical neutrinos with particles in the ice to construct the dataset. These tracks are monitored and detected by an array of digital optical modules that record the relative intensity of the light created, hits per Cartesian location, frequency of hits per unit time and number of hits of each event. These measurable quantities are stored in the columns of the data provided by the Particle Physics Playground. Using the number of hits in each event, I engineer features including the optical intensity (q), Cartesian coordinates in distance units and the time hits are detected. The combination of all the mentioned feature columns characterize the different detection events and individual hits in the experiment. Fig. 1 shows the 3D reconstruction of event 0 in the dataset. There are a few key features to observe. The first of which is that there is not only one singular source of light inside of the detector, rather there are multiple sources spread out within. This would indicate that a given event contains the detection of several neutrinos, not just one at a time. Thus, when the machine learning algorithms are implemented, it will be useful to not necessarily look at the data event-by-event, but rather examine the signals in totality, and then attribute each to a certain neutrino flavour. Another piece of information that can be deduced from this plot are the intensities and size of the light detected around certain regions. The light emanates spherically outward from the source of the radiation in multiple regions, each with different intensities and widths. These pieces of evidence can uncover information regarding the decay product that is created via the interaction, and thus the neutrino as well. According to the literature, the profile of the light emitted by an event is specific to the neutrino that created the signal. Thus, I begin the exploratory data analysis attempting to find differences in the optical signal across events. If it is true that the optical signal parameter is unique for each of the neutrinos that are observed at IceCube, then there should be some detectable differences in the mean and standard deviation of the light intensity across events. I begin by extracting the data corresponding to each of the 10 events by first importing the mass data from the public Particle Physics Playground area (found under Libraries and Data Load-In section) and then looping over each of the events (Particle-Physics-Playground, 2025). In doing this, I plot the histograms of the hits per optical intensity (optical parameter) for each event to visualize the differences in the signal. If I can observe noticeable qualitative and quantitative differences, then this will support my methodology and research question.

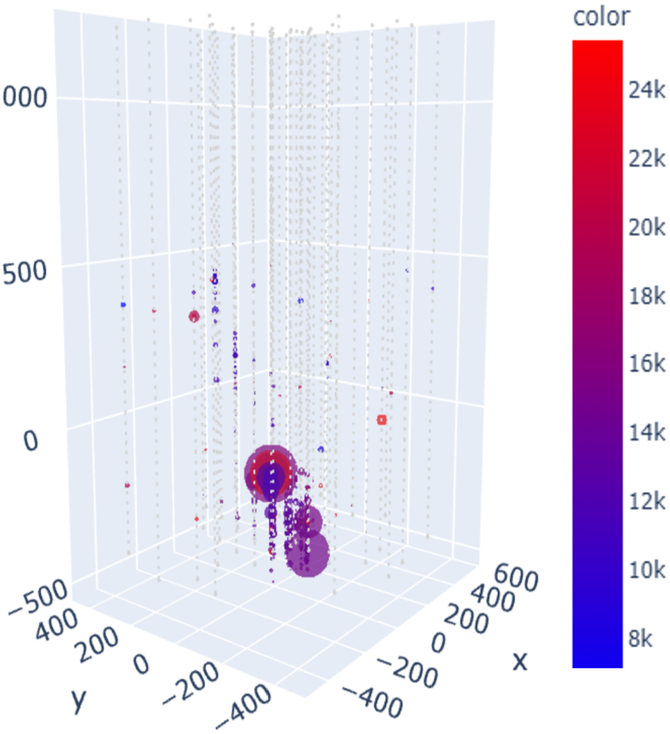

Figure 1: *This figure shows the 3D interactive map of an event detection, specifically event 0. It shows the light detected from the detectors as well as how bright they are. This is done using visualization tools by Particle physics Playground (Particle-Physics-Playground, 2025)*

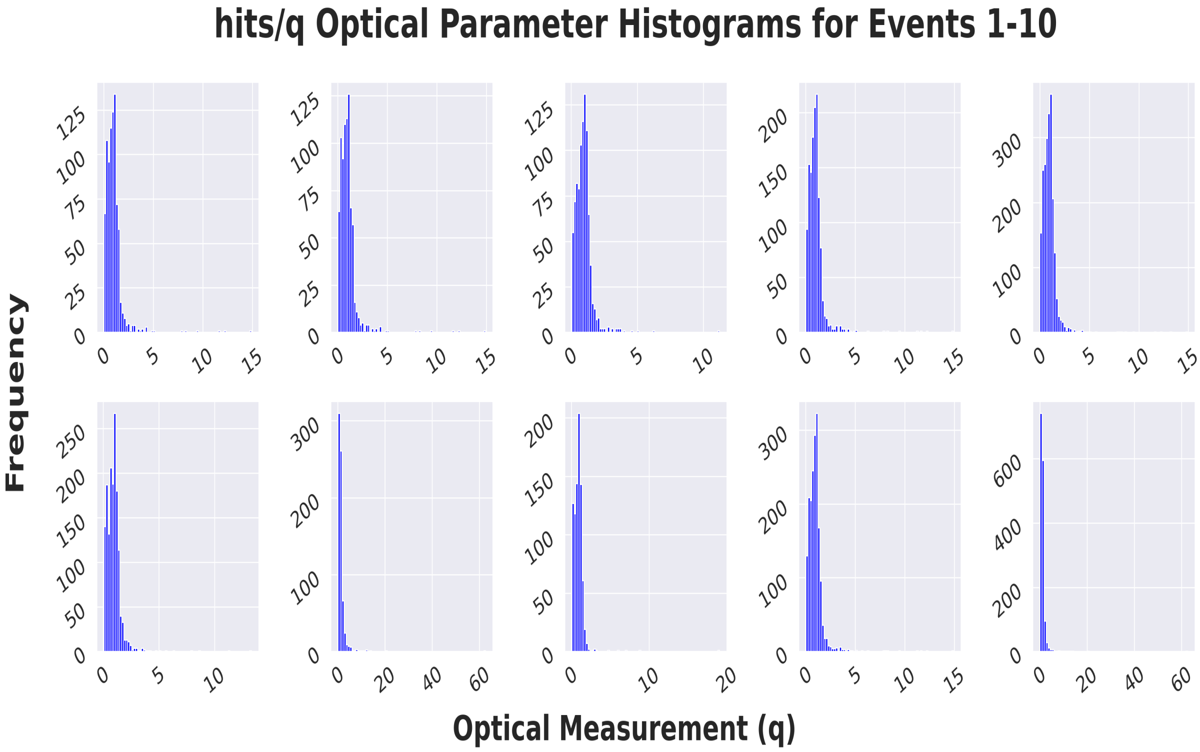

Figure 2: *This figure shows the histogram plot of the optical parameter (hits/q) across all 10 events. This is done with Matplotlib's histogram plotting feature (Hunter, 2007)*

This visualization can be seen in Fig. 2. What this shows is that there are clear differences in the mean and standard deviation of the optical parameter, q, across events 0-10. In particular, qualitatively speaking, one can see that each of the histograms is normally distributed with a slight rightward skew. What distinguishes each of the histograms is at which intensity they are centered around, as well as the spread of the optical values detected during the event. Some of the distributions have a very wide spread of values, while others are more condense. To examine this further, I again loop over the events 0 through 10 however, I calculate both the standard deviation and mean optical intensity in each event. I then present a paired bar plot of this information to visualize the extent of the difference in the optical measurements across the different events. Furthermore, this data is quantitatively displayed in Table 1. This serves as strong evidence to signal the validity of the proposed method above. It shows that there are several large differences in the optical intensity parameter across neutrino detection events. In particular, events 7 and 10 display an above average mean and extremely large optical standard deviation. Thus, it is fitting to suggest that this may be attributable to the high energy astrophysical muons, since the intensity of the optical signal is clearly larger and more dispersed. Another validating piece of information from the literature is that the electron and tau neutrino signals are very similar and sometimes indistinguishable. This observation seems consistent with what is being observed, since there are smaller differences in the mean and standard deviation across the other events.

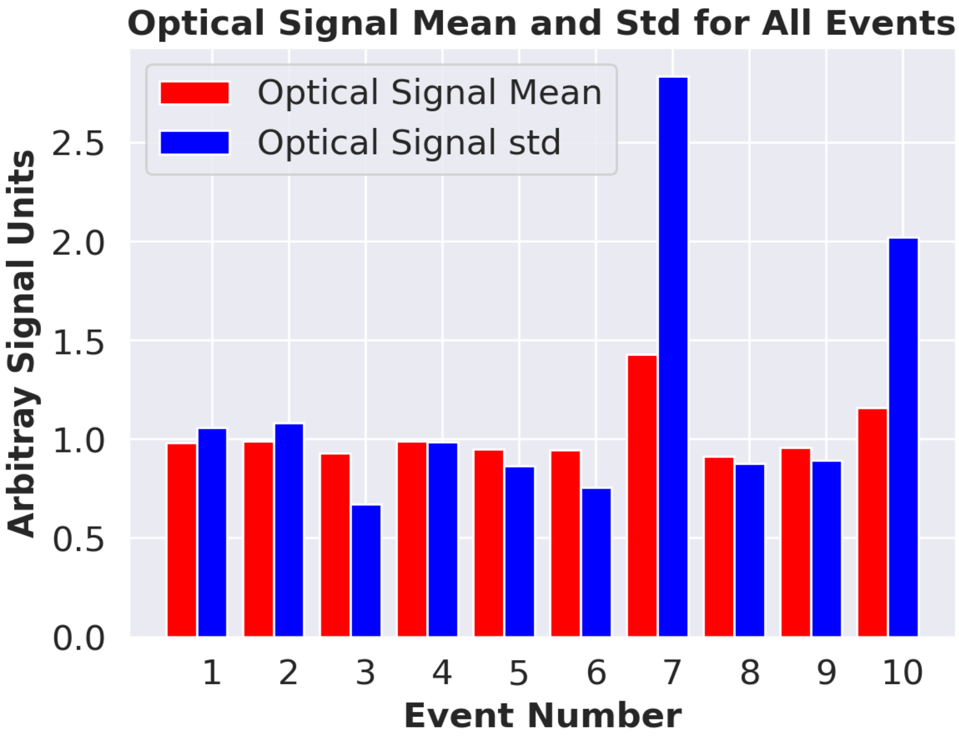

Figure 3: *This figure shows the paired barplots of the mean and standard deviation of the optical intensity parameter for events 0-10. This is done with Matplotlibs's bar plot feature (Hunter, 2007)*

Thus far, it has been observed that there are significant differences in the optical signal parameter across events which may be linked to the interaction of different flavours of neutrino. The next task is to visualize the verity of optical signal values detected in each event and visualize the variance. In other words, how many extreme values of optical intensity are there and how often do these events occur. Presumably, for the high energy muon neutrinos,

Table 1: **Events Data (1-10): Mean and Standard Deviation of the Optical Parameter 'q'**

Event	Mean (arb. opt. units ['q'])	Std. Dev. (arb. opt. units ['q'])
1	0.983	1.058
2	0.990	1.082
3	0.929	0.670
4	0.989	0.984
5	0.949	0.862
6	0.944	0.757
7	1.426	2.835
8	0.912	0.877
9	0.958	0.894
10	1.157	2.022

Table 2: Shows the change in the mean and standard deviation of the optical parameter between events, highlighting key differences in the optical fingerprint which may be caused by the interactions of different neutrino flavors. Distinguishing between them using this method can help determine which are high-energy astrophysical neutrinos and which are background from random sources.

their interactions will generate a significant amount of light intensity. Thus, if one would like to differentiate between background electron and tau neutrino flux from the muon neutrino flux, isolating and understanding the extreme values is important. To do this, I include a box plot, located in Fig. . The box plot shows several features of the data. The central rectangular box displayed represents the inter-quartile range, or in other words, the central 50 percent of the data. The red line represents the median value of the optical parameter for the particular event. Finally, the circles represent the outliers that lie beyond 1.5 times the max values of the inter quartile range (i.e. where the whiskers extend out to). What does this suggest about the nature of the data? This shows that each of the events in the data have a relatively similar median (not to be mistaken with the mean), as well as similar inter quartile ranges. The data points are densely located around the 1-2 optical unit range, showing that many of the detected values are small for all the events. However what is really telling are the abundance of outliers in each of the events, of which events 7 and 10 are most prominent. These high energy outliers extend as far as upward of 60 optical intensity

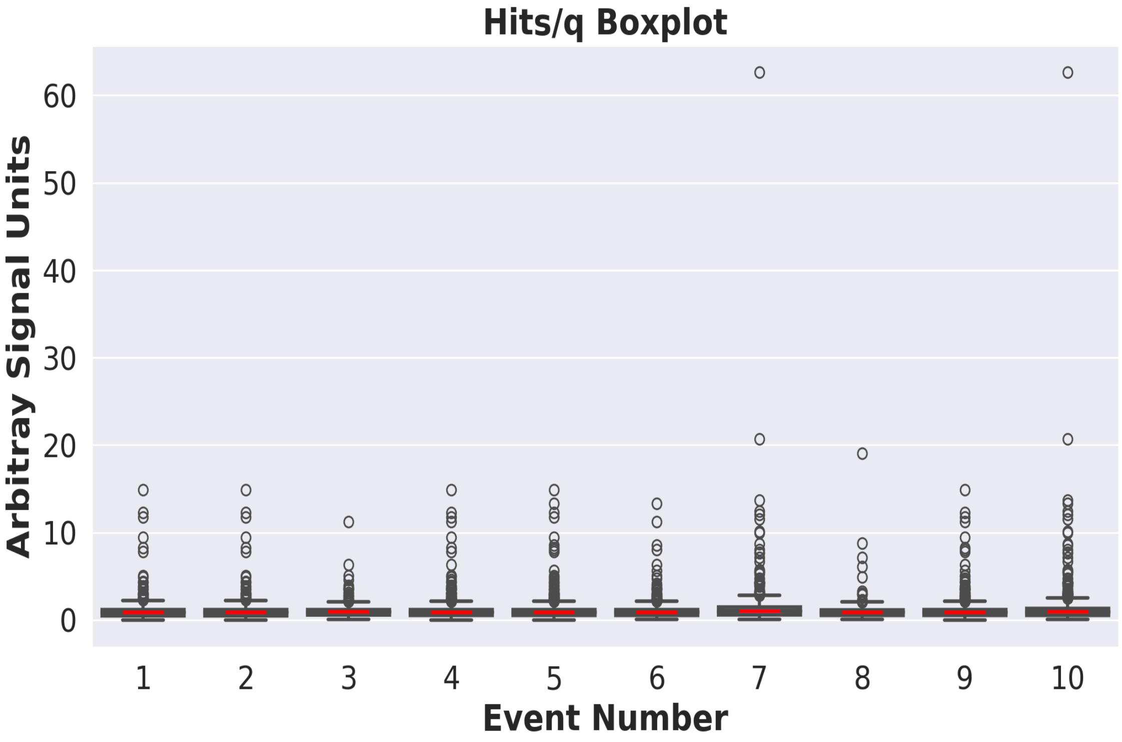

Figure 4: *This figure shows a box plot of the hits/q value across the events 0-10, which visualize the abundance of extreme values of optical intensity*

units. This would suggest high energy muon neutrino activity occurring, which again bolsters the validity of the methodology. The final plot and exploratory metric being explored in this section is the Pearson's correlation to test for parameter linearity. This matrix is calculated on all of the data which has been concatenated across all events to see how these variables are related on a mass scale. The Pearson's correlation score takes the following form;

$$\text{Pearson correlation } r = cor(X,Y) = \frac{cov(X,Y)}{std(X)std(Y)} \tag{1}$$

where 'corr' denotes the correlation between the x and y parameter, 'cov' denotes the corresponding covariance and 'std' denotes the standard deviations. This matrix is visualized by the heatmap given in Fig. 5. What this shows is that the data is not strongly linearly correlated. However, there are some interesting variables which show a quantifiable anti-linear correlation. hits/x and hits/z, hits/t and hits/z as well as hits/y and hits/z show an anti-linear correlation. These anti-linear relationships may be the subject of

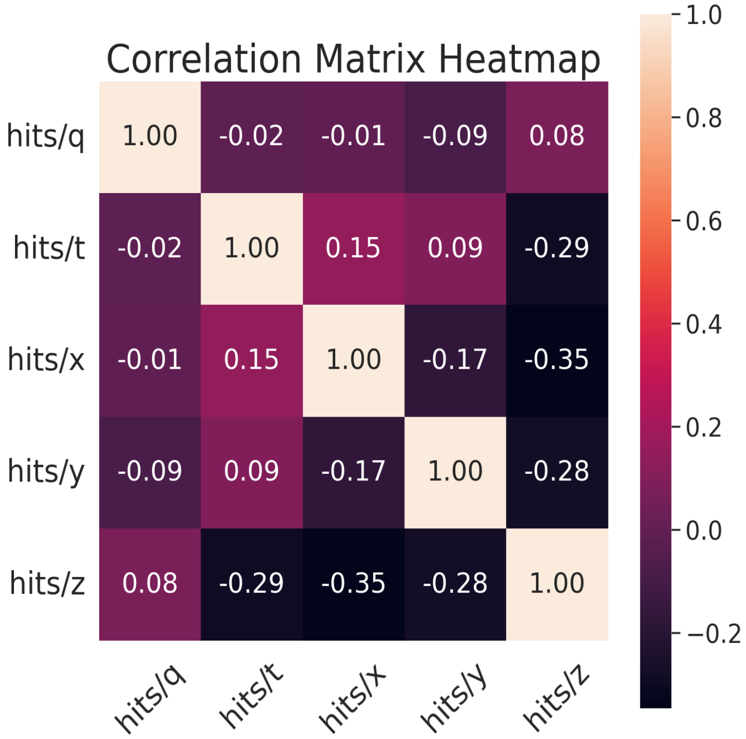

Figure 5: *This figure shows a Pearson correlation matrix heatmap plot that visualizes the linearity score between any two parameters in the data. This is done with Seaborn's heatmap feature (Waskom, 2021)*

a further discussion outside of this. The non-linear relationships (values near 0) will become the foundation for the usage of the random forest regression techniques used. Based on these key pieces of exploratory data analysis, it is clear that the data fits the methodology. In all of the statistics implemented, it was clear that there is an observation of extreme high energy phenomena. These high intensity bursts of light detected by IceCube, in agreement with the literature, can be attributed to muon neutrino events. Not only that, but we were able to show that there were clear signal differences between events (most notably between 7 and 10) which suggest an abundance of muon neutrinos or lack thereof in the rest of the events. These stark differences will cater to the research question of being able to classify the neutrino flavour by the optical parameter, since it is clear that there are quantifiable differences. In particular, these significant quantifiable differences will benefit the results of the subsequent k-means clustering algorithms, due to the stark differences observed. To set up the implementation of the machine learning algorithms, the rest of the analysis uses the mass data that was used to calculate the Pearson's correlation matrix. Since the events are likely detecting several neutrinos during the time an event occurs, we would like to place all of the neutrino detections together. With this, we can then plot all events together to expand the amount of active neutrino detections, and hopefully show how detections pertaining to the specific neutrino flavours will group together. In this way, we can identify the muon, tau and electron neutrino and be able to predict their optical signals later on.

4 Results and Machine Learning

In this section, I discuss the implementation of the machine learning algorithms I used, as well as the interpretations of the results I generated. In the concluding part of this section, I will mention where this research can be taken in the future (with more time and resources), as well as what was tried and did not work, requiring omission. To begin, prior to clustering, I wanted to see how well I can predict the optical parameter for all the events, based on the predictors present in the data. I chose this as my preliminary test since it would tell me whether or not the Cartesian space information or hit frequency would be strong feature variables to accurately predict my optical parameter. However, this will be an incomplete picture and will be further refined later on. Why is it incomplete? Since I have amalgamated the events data together to try and visualize the grouping of each unique neutrino flavour event, this data contains different physical phenomena. This means that the accuracy of the model will not be at peak performance, since it has to try to converge to a solution that is governed by different physics. Again, this section should serve as a proof of concept which will aid in the later, more sophisticated version of this task. The first thing I do is determine which model I want to use to predict my target variable. As per the deductive statistical result observed with the Pearson correlation matrix, we find that the parameters are, for the most part, not strongly linearly correlated. This means that a simple linear regression would not work well here. Thus, I have opted to go with the random forest regression model to predict the optical signal parameter. I have decided to use this model for a few key reasons. The first of which is the non-linear nature of the data. The random forest regression model uses decision trees to predict the likelihood of where a predicted data point is likely to fall. These decision trees make it possible for the random forest regression model to accurately capture complex relationships and patterns in the data, that would otherwise not be modeled well in a linear regression. Secondly, random forest regression models work quite well at handling outliers in the data. Since this data has an abundance of outliers which contain important information about the physics of the neutrinos in the ice, it is paramount that the model must handle these points well. Finally, random forest is also strong at preventing overfitting. Before implementing these algorithms, I need to address two questions;

- What will the target value be
- What are my predictors going to be

Upon examining the results of the initial EDA, there are very important results that should be noted. I am trying to be able to predict the flavour of the neutrino using the optical signal parameter. Hence, my choice of a target variable is either the hits per optical signal (hits/q) or the optical signal (q) itself. I find that hits/q is the correct metric to predict. Suppose there are two events, one of which has a very high optical signal and one that does not. Without understanding the number of hits the detector measured, then the intensity of the burst per each hit will be unknown. Exclusively knowing that the optical intensity is bright does not provide sufficient information to describe the intensity of the source that created the light. It is possible that the q parameter alone would be misleading, since a detector might be observing the light from many hits, but each hit is relatively weak. Furthermore, it is also possible that the detector picks up a weaker signal but it was caused by only one, strong hit. Thus, the most illuminating and applicable variable to predict is hits/q. What are the predictors I will be using? Certainly I'd omit the hits/q and the q parameters from the list of predictors, as this is the quantity we want to target. Furthermore, the index columns and event number will not contribute to predicting the optical signal parameter. This is because there is no expectation of a relationship between the optical signal and event number or index as they are not physical quantities. Thus my predictors will be the location of the hit (x,y,z), the frequency of hits per detector location (hits/x, hits/y, hits/z), the time a hit is detected, the frequency of hits per unit time and the number of hits. Now that the target and feature variables have been determined, a very pertinent issue must be addressed. These columns are each measured on a unique scale. In other words, the distance measurement cannot be directly compared to a time-based measurement since their scales are completely different. Hence, a standard scaler tool, given by Sklearn (Pedregosa et al., 2011), is used to scale the data, such that it can be used to predict the optical parameter accurately. If the data were not adjusted with a standard scaler, then certain quantities may hold more weight in the regression and negatively affect the results. With these developments and treatments to the data, we can now begin the random forest regression implementation. To implement this model, my first task is determining the number of trees I want to use, such that the cross validation accuracy (model fit accuracy grade) is a maximum. To do this, I loop over 'n' trees ranging from 10 to 100, in steps of 10, perform a random forest regression, and then grade the fit. At the end, I extract the best tree number based on the cross validation score. When doing this, I find the best cross validation score is 0.6920 at 80 trees. This means that the model correctly predicts the optical parameter around 70 percent of the time. This indicates that the method is quite successful. Since I have now determined the optimal number of trees for this regression task, I can implement it to actually predict the hits/q parameter across all events. Doing this, I will be able to plot the predicted and actual hits/q and observe the closeness of the model to describing the actual data. I first use a 70/30 train test split to sort the data into a training target and feature set of data (to train the random forest regression model), and a test target and feature set of data (to test the subsequent model). This is done using the data which has already been scaled, such that the units being compared are all the same. Using 80 trees which was found to be optimal, the model is trained and scored using a cross validation score, R^2 test, and mean squared error in each iteration. A plot of the actual vs predicted optical parameter is also included in Fig. 6. When doing this, I find that the R^2 score for the random forest regression is 0.42 and a mean squared error of 0.6. This shows that the 42 percent of the variance in the data is explained by the regression algorithm. While this is a good result, the mean squared error shows that the average squared distance between the fit and the actual data is 0.6 which is a relatively high value compared to the mean optical intensity (which is around 1.2). This means that there is a large share of outliers that are effecting the performance of the fit and causing a large residual in some cases. Superficially, this seems like a poor result however, this is actually expected. This makes sense, since we are demanding the regression task to predict three distinct physical phenomena (muon, tau and electron neutrino interactions) so these residuals are not a concern. With this being said, the methods applied ahead will take a more complete form by addressing individual classes pertaining to each flavour.

The next goal, and arguably the most important is the clustering phase. As mentioned above, one of the primary objectives is being able to classify the neutrino flavour by the optical signal parameter by grouping the data into three classes. These classes would correspond to the particular neutrino flavour that produced the light bursts detected by IceCube. To actually

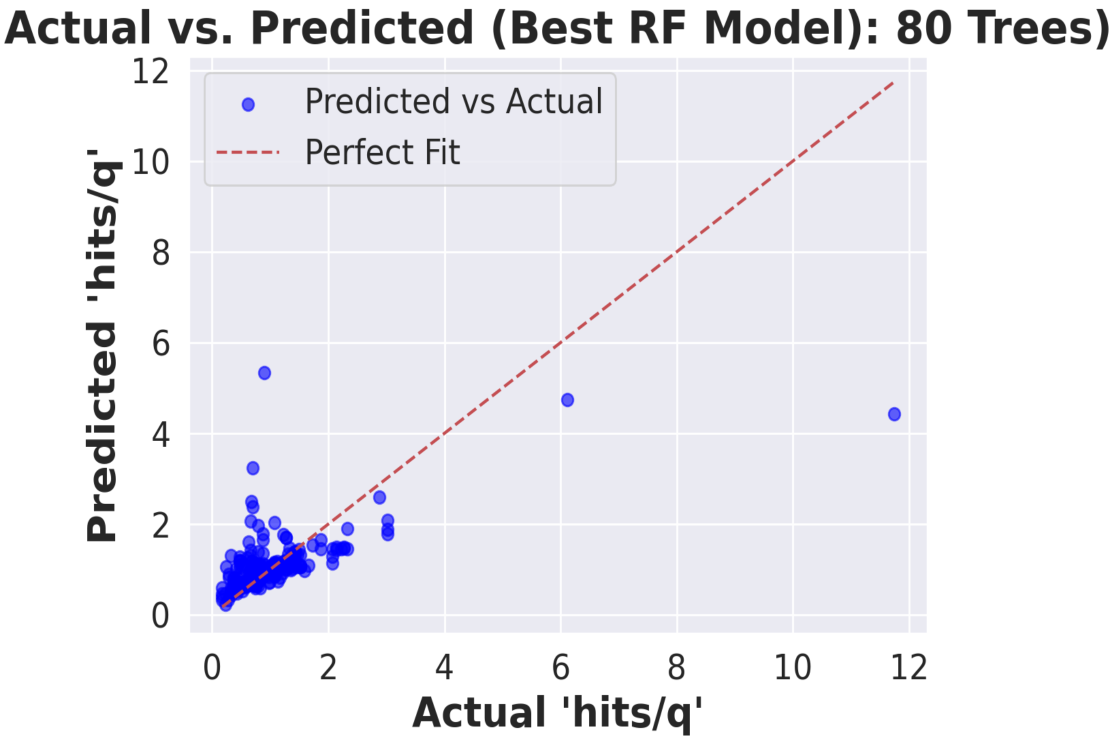

Figure 6: *This figure a plot of the predicted hits/q parameter by the random forest regression on all the data, as a function of the actual hits/q data across all events. This is done with Matplotlibs's scatter feature (Hunter, 2007)*

perform this, I will be using a K-means clustering algorithm to group the data into these three groups based on the patterns produced, by plotting the optical signal parameter as a function of a given correlated independent variable. In particular, I run this K-means clustering twice. Why would I run this twice? I am grouping the data based on the presumption that plotting the optical signal parameter against a particular independent variable will show signs of the data naturally clustering into distinct regions due to the neutrinos that incited the reaction. Based on the exploratory data analysis section, there are clear physical relationships between the optical signal parameter and the frequency of hits per unit time, as well as the frequency of hits as a function of the depth within the detector. Each of these dependent variables holds differential information that can distinguish the neutrinos reaction products (i.e. Cherenkov light signal) from each other. In the case of the frequency of hits per unit time (hits/t), we expect to see the following relationships. High energy events are typically associated with a strong intensity of light per hit, that occurs in a short period of time. This means that there isn't a large flux of light being detected by the DOM's per unit time, but the intensity of the light that is detected is very intense and occurs rapidly. Thus, in this clustering, we expect to see a large share of very intense optical light occurring at a smaller hits/t value, indicating a fast and very intense light burst. We also consider the independent variable of hits/z, that describes the frequency of hits as a function of the depth. By analyzing hits/q as a function of hits/z, we are examining the spatial relationships of the optical light intensity detected as a function of the distance beneath the surface of the ice in the detector. What we'd expect to find, is that the most energetic neutrinos (namely the muon neutrinos) of extremely high energy will penetrate further into the ice and create light at higher values of z. Thus, it follows that when I plot the hits/q as a function of hits/z, I will find at large hits/z values that there will be an abundance of high energy optical measurements which presumably correlate to the muon neutrinos that penetrate deeper into the ice. Since there are two key independent variables that, when plotted against hits/q, associate into discernible groups, I perform a K-means clustering on both plots. I now discuss the implementation of this technique on the hits/q vs hits/t data. Using TopCat, I plot the data and notice a general grouping of the data into clusters around 11,000 hits/t 13,000 hits/t and 20000 hits/t. The means of the optical parameters are 0.99, 1.2 and 1.2 respectively. What this shows is that the data clustering is rather dependent on the frequency of hits per unit time. Interestingly, I find that for lower hits/t, there is an abundance of high intensity optical detections - this is in confirmation with what was initially expected for muon neutrino reactions creating a short-lived high energy light burst that moves quickly past a detector. I initialize the centroids based on the specified clusters observed and use a 3-cluster K-means algorithm supplied by Sklearn (Pedregosa et al., 2011). The results of this are displayed in the notebook under the 'Predictions and Clustering' section. I find that the results show a cluster of high energy, low frequency hits, which are likely attributable to the muon neutrino. The next highest mean intensity cluster is the central frequency group and finally the lowest mean intensity cluster is the high frequency cluster. Since the tau neutrino is generally considered to be more energetic than the electron neutrino, one can assume that this medial frequency cluster is associated with such. Finally, the least energetic cluster likely corresponds with the electron neutrino. Performing the same analysis with the hits/q as a function of the hits/z gives another validating result. To quantify the level of separation between each cluster, I implement the silhouette score. It is a means to grade the K-means clustering algorithm by evaluating how well each cluster is separated from each other, one a scale from 0 to 1. When implementing it on this data, I find that the silhouette score is 0.65, which indicates a well separated set of data. Repeating this process on the hits/z data yields the following results and is visualized in Fig. 7. The K-means clustering finds a high intensity group at large hits/z values, which contains the majority of the highest energy optical hits detected. This result is also in agreement with the theory, such that these high hits/z values correspond to locations deeper in the ice and thus correspond to high energy muon neutrinos that generate very intense light signals in these deep regions. The next highest cluster is again the medial hits/z values which can similarly be attributed to the tau neutrino, and subsequently the lowest energy (being

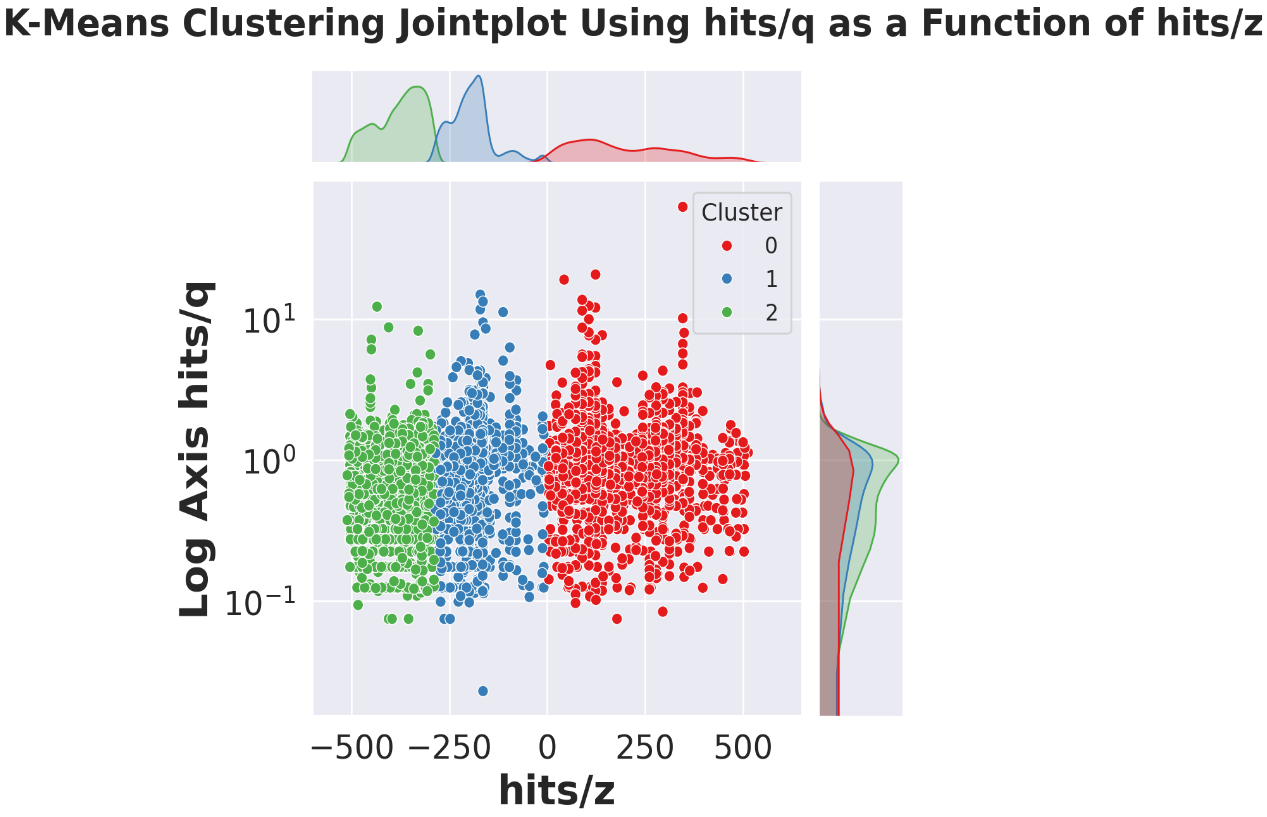

Figure 7: *This figure shows the k-means squared clustering of the events data into three groups. Each group represents a specific neutrino flavour with group 0 presumably representing muon neutrino data, group 1 corresponding with tau neutrino data and group 2 representing electron neutrino data. This was clustered by plotting the hits/q as a function of hits/z. This is done with Seaborn's jointplot feature (Waskom, 2021)*

the lowest hits/z) assumed to be correlated to the electron neutrino. I again apply the silhouette score to this clustering tasks and find that is is 0.34. This indicates that the data is less separated in this domain, but still shows some signs of cluster separation. In the figure and in the notebook, cluster 0 is representative of the muon neutrino population, group 1 represents the tau neutrino group and group 2 is the electron neutrino group. Now that I have created these clusters and identified groups corresponding to the particular neutrino type, I can now revisit the random forest regression model. As per the remarks made above regarding the preliminary regression model, it was observed that there were pitfalls in the method used to predict the hits/q parameter, since we tried forcing the model to make predictions of different physical phenomena (i.e. different flavours all at once). Now that we have presumably separated the data into clusters associated with the different neutrino flavours, we can make three subsets of the data that correspond to each of the clusters. With these subsets, we can run individual regression tasks on each of the neutrino flavour groups and test to see if the predictions get any better. We first test with the hits/t independent variable clusters, again using 80 trees and score the model using cross validation, R^2 score and mean squared error. For group zero, I use a 40/60 train test split to optimize the output of the regression task while not sacrificing the reliability with overfitting. I find that the R^2 score is relatively low with 0.28, with a mean squared error of 2.7. This means that there is a lot of variance and extreme values in the data which the regression has a hard time predicting the optical intensity. This would indicate that the cluster corresponding to the muon neutrino has too many data points corresponding to lower energy events, not triggered by muon neutrinos themselves. This can also be due to low sampling in this cluster that causes issues as well. This is a point of improvement for future iterations of this method - to try and more accurately cluster the groups. However, what this group does well, is show that the mean intensity of the optical signal is the highest among all groups, being 1.1 which is consistent with the expected observation. Groups 1 and 2 model quite well, as the data is more well behaved here, thus I can use a more generous 70/30 train test split. For the tau neutrino group, I find that the R^2 score is 0.44 and the mean squared error is 0.39 and for the electron neutrino group, I find a R^2 score of 0.41 and a mean squared error of 0.47. Performing the same tests of accuracy on the cluster subsets created from the hits/z independent variable plot I find the following results. For the group 0 cluster, using the same parameters for creating the model (specifically referring to the particular train test splits, etc.), I find that the R^2 score for this fit is 0.2 and has a mean squared error of 4.6. This cluster seems to contain a greater proportion of non-muon neutrino events which is causing the model to not accurately predict the physics of the high energy outlier events. Thus, this explains the low R^2 score for a given fold and the large variance, since the high energy outliers stick out so much more than the lower energy data. Once again, the lower energy groups seem to be much more well behaved. For the tau neutrino group, I find an R^2 sore of 0.52 and a mean squared error of 0.4, and for the electron neutrino I find a R^2 score of 0.38 and a mean squared error of 0.31. This shows that the model does a good job predicting the more well behaved and abundant phenomena (being the low energy cases) but struggles at separating out the high energy events, as well as predicting them. As mentioned, this is a point of improvement for future models. The results from this section are visualized in the notebook and summarized in Table 2.

Neutrino Group	**Feature Set**	**R^2 Score**	**Mean Squared Error (MSE)**
Muon (Group 0)	hits/t	0.28	2.70
Tau (Group 1)	hits/t	0.44	0.39
Electron (Group 2)	hits/t	0.41	0.47
Muon (Group 0)	hits/z	0.20	4.60
Tau (Group 1)	hits/z	0.52	0.40
Electron (Group 2)	hits/z	0.38	0.31

Table 3: R^2 scores and mean squared errors for each neutrino group using hits/t and hits/z feature sets.

Summarizing what I've done thus far, I have created three clusters based on the hits/q as a function of hits/t and three clusters based on the hits/q as a function of hits/z. These clusters are presumably describing the same physics, in that they correspond to data that undergoes the same phenomena. This implies that in the ideal case where the k-means clustering can predict all the points in the muon neutrino cluster (in both the hits/t and hits/z feature sets) with 100 percent accuracy, then all the same data points clustered by the in group 0 with the hits/t independent variable, will be the exact same data points clustered in group 0 in the hits/z independent variable. This idea also applies to the electron and tau neutrino groups, as we've suggested that the groups are coincident with one another. Thus, we can use the groups as pseudo labels and test the accuracy of the clustering predictions made by the previous k-means clustering tasks. In other words, we can introduce a confusion matrix to determine the amount of times the k-means clustering models can accurately predict data points being in group 0, 1 and 2 in both cluster cases. This is done using Sklearn's confusion matrix library and is plotted in Fig. 8. The way to interpret the 3x3 matrix is as

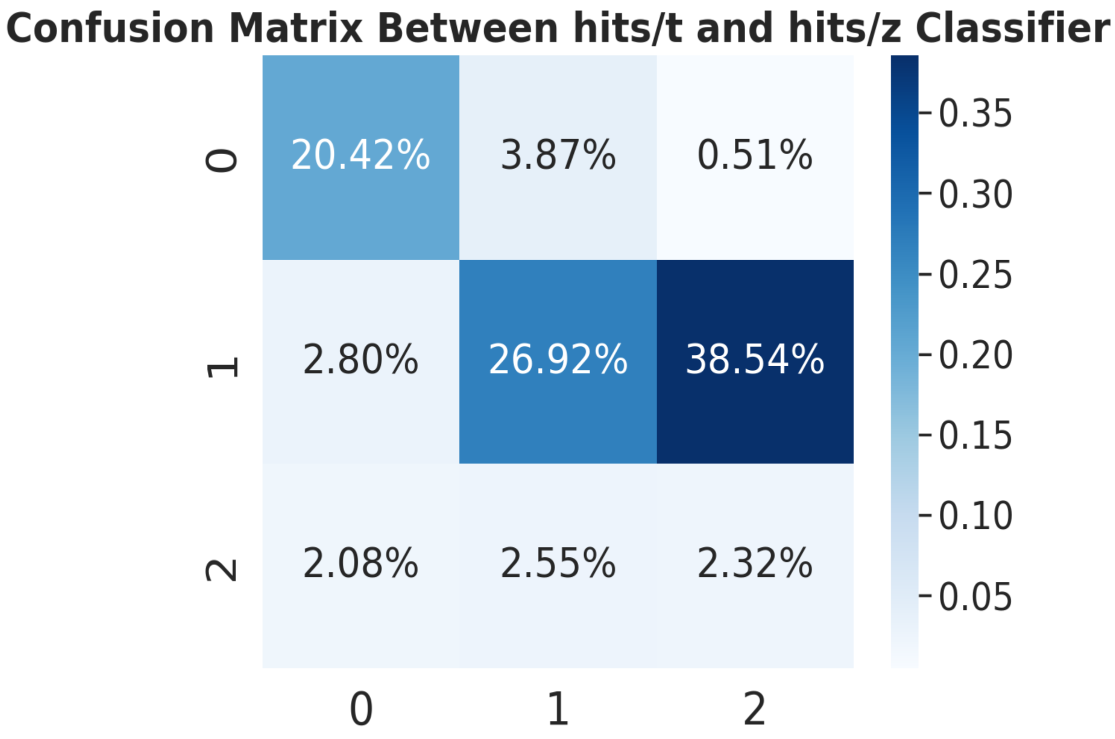

Figure 8: *This figure shows the confusion matrix comparing the amount of consistent predictions of group labels between two independent k-means clustering tasks. This is done with Seaborn's heatmap feature (Waskom, 2021)*

following. The diagonal entries represent the true positives, which are the number of instances when the model placed the given data point in the same group in both cases. The off diagonal entries represent the false negatives. For example, the first column second row corresponds to when the model misinterpreted a group 0 data point in either group 1 or 2. The results of this matrix are quite promising. From this figure, we see that the model groups the same data point in cluster 0 around 20.42 percent of the time, which far exceeds the false negatives for the group 0 data. The same applies for group 2, which is around 26.92 percent acuate. While the results are promising, group two is certainly the weakest, with very little predictability accuracy.

What does this analysis say about the validity of the results? Can or should we trust the result we gathered? The R^2 values, cross validation, as well as the mean squared error and confusion-matrix determinations are all good indicators of the performance of the machine learning algorithms applied on the data. In the first regression task, I find a mean squared error of 0.47 and a cross-validation score of 0.6920. This means that the average R^2 value over 10 folds (specified in the notebook), is around 0.7 with the average squared residual value of 0.47. This would indicate that the model did a good job predicting the hits/q value from the set of specified feature variables. In other words, we can trust the results superficially however, since we know apriori that this data includes the physics of three different phenomena (electron, tau and muon neutrino interactions), we can say that the regression predicted the hits/q value, but not the individual neutrino optical intensities. Then moving to the clustering, we perform a k-means clustering using plots of the optical signal parameter against 2 different independent variables; hits/t and hits/z. Qualitatively, the algorithm seems to have clustered the data quite well pertaining to the different flavours. We see a group of high energy neutrinos, which presumably describe the muon neutrinos, a medial intensity group of tau neutrinos and finally a low energy group of electron neutrinos. We perform a silhouette score to grade the level of separation between the groups. For the hits/t independent variable feature, we find this to be 0.65 and for the hits/z independent variable feature this is 0.34. This tells us that for the hits/t data, we can quite reliably trust that the data is split into 3 clusters quite well, since this scale maxes out at 1. For the hits/z clusters, we can semi-reliably show that the clusters are well defined. However, it borders on the low side, meaning that some data may not be allocated correctly in the correct group. To further validate this, a confusion matrix is implemented to measure the consistency of labeled groups in the hits/z and hits/t feature clusters. We find that there is a 20 percent agreement between group 0 in both cases and a 26.9 percent agreement between group 1. This would indicate that the model can quite accurately predict the groups containing high energy muon neutrinos and medial energy tau neutrinos. However, the model fails when addressing the low energy electron neutrinos, where it incorrectly assigns group 1 or 0 to a group 0 value 38.5 percent of the time. With this, it seems we cannot trust the group 2 data, but group 0 and 1 data is semi-reliable. Finally, based on the predictions using random forest regression of the optical parameter for individual neutrinos using the k-means clustering groups, we find that group 0 is not as well behaved as one would hope. The model seems to have a difficult time predicting these values since the clusters did not separate out the lower energy tau neutrino events from the high energy muon neutrino events enough. This applies to both the hits/z clusters and the hits/t clusters. Therefore, one should take the results presented with a grain of salt. There are certainly strong signs to be suggest that it is possible to classify the flavour of neutrino detected by IceCube by the optical parameter. However, the evidence gathered here does not definitive prove that. A more thorough analysis would need to be used and perhaps a larger set of data.

5 Conclusions

In conclusion, the data and machine learning analysis for the IceCube neutrino observatory data was presented. To summarize the results of this analysis, we first found the optimal number of decision trees to generate a maximum accuracy of the random forest regression, which was found to be 80 trees. We then used this information to perform regression on the whole dataset to try to predict the optical signal parameter. We then carried on to implement a k-means clustering algorithm to be able to group the data into three distinct groups, corresponding to the different neutrino flavours. We did this twice, once with the feature variable being hits/t and then hits/z. We then created subsets of the entire data set, characterized by what was found in the clustering algorithm. We then applied a random forest regression to each of these groups to be able to accurately predict the physical phenomena related to each neutrino flavour. Finally, we compared the data points found in group 0 between both clustering tasks, as they are expected to be governed by the same laws of physics. We test the accuracy of the clustering method by comparing the number of points included in the same group label between the clusters using the hits/t and hits/z using a confusion matrix. We also test the accuracy of the random forest regression models using a verity of accuracy statistics including cross validation, R^2 and mean squared error. This analysis showed signs that there is a potential future involving the ability of classifying the flavour of neutrinos based on the optical signal parameter at IceCube. The methodology, while not 100 percent supported by the results, suggests many hints regarding the ability to classify and predict the hits/q value for each neutrino flavour. I've been able to show that is is possible to generate three distinct groups presumably corresponding to the different flavours of neutrino in the IceCube dataset. While not 100 percent accurate, there are definite signs that the expected theory agreed with what we saw. The confusion matrix confirmed that this method worked relatively well with predicting the muon neutrino and tau neutrino groups. We also saw relatively high success predicting the hits/q value for the electron and tau neutrinos, since they are more well behaved and similar. Showing that these predictions are possible can help train future models to be able to classify the different flavours using the optical signal parameter. Some of the limitations

I encountered included the size of the data. The Particle Physics Playground dataset for this experiment was not as expansive as I would have liked and did not contain a vast number of feature columns. This made it difficult to find relationships between the optical parameter and other variables, especially in clustering and predicting. For future research, I would suggest using a larger dataset with more feature columns. There are no immediate implications of this research on contemporary practices or theory however, if one could refine this method, it would allow accurate prediction of the neutrino flavour observed in IceCube. This would hold valuable information regarding the astrophysical events that created the detected neutrinos, as well as background noise level abundance. Both of these considerations would certainly advance the scope of the classification methods used presently and increase the understanding of these events as a whole.

References

A. Palladino, e. a. (2015). Which is the flavor of cosmic neutrinos seen by icecube? *arXiv preprint*.

Cass, S. (2018). The icecube neutrino detector at the south pole hits paydirt. Accessed: 2025-02-19.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hunnefeld, M. (2021). Combining maximum-likelihood with deep learning for event reconstruction in icecube. *Proceedings of Science*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

IceCube-Neutrino-Observatory* (2025). Frequently asked questions. Accessed: 2025-02-19.

Observatory, I. N. (2025). Icecube neutrino observatory. `https://icecube.wisc.edu/science/icecube/`. Accessed: 2025-12-29.

Particle-Physics-Playground (2025). Ice cube introduction. Accessed: 2025-02-22.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Steven W. Barwick, C. G. (2022). Radio detection of high energy neutrinos in ice. *Encyclopedia of Cosmology II*.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.