Technische Universität Dortmund

Physics Department

International Master Advanced Methods in Particle Physics

Analysis of IceCube Data

Laboratory report

Date: October 27, 2024

Simone Garnero - simone.garnero@tu-dortmund.de

Bastian Schuchardt - bastian.schuchardt@tu-dortmund.de

Dortmund, Germany May, 2024

Contents

Analysis of IceCube data

Advanced Laboratory Course: Particle Physics

Simone Garnero, Bastian Schuchardt

October 27, 2024

Abstract

The goal of this analysis is to perform a neutrino selection on data from the IceCube experiment to differentiate between atmospheric and astrophysical neutrinos. At first, the minimal redundancy maximal relevance (mRMR) method is used to extract the best features. After that, a naive Bayes, random forest, and multi-layer perceptron classifier are trained and their performances are evaluated on different metrics.

1 Introduction

Neutrino astronomy is an emerging field of astroparticle physics that utilizes neutrinos to investigate high-energy processes in the universe. Neutrinos are chargeless and nearly massless particles and only interact rarely by weak interaction with matter. The advantage of neutrinos for astronomy compared to photons is that neutrinos rarely scatter, do not get deflected by magnetic fields, and can pass through large amounts of matter almost without absorption. Moreover, they can be used for multi-messenger astronomy, where the goal is to investigate astrophysical events by multiple signals. One of the most important neutrino observatories is the IceCube neutrino observatory [2] located at the south pole in Antarctica and it mostly detects muons and neutrinos. These muons and neutrinos originate from atmospheric or astrophysical sources and it is of interest to differentiate the astrophysical events from the atmospheric ones to get a better signal to background ratio.

2 Theory

The study of particles arriving at Earth from the universe has been of interest for many centuries but the term "cosmic rays" only emerged in the 20th century and describes rays of highly energetic particles like protons, heavier nuclei or neutrinos. Most incoming cosmic rays interact with the earth's atmosphere and create particle showers that can be measured on the ground. There are many different sources of cosmic rays in the universe, e. g. active

galaxy nuclei or solar eruptions. The energy spectrum of cosmic rays extends up to 10^{20} eV and can be described by a power law

$$\frac{d\Phi}{dE} = \Phi_0 E^{\gamma} \,, \tag{1}$$

where the spectral index *gamma* has an approximate value of -2.7.

2.1 Atmospheric and Astrophysical Leptons

Atmospheric leptons mostly originate from lighter mesons like pions or kaons that are created by the interaction of cosmic rays with the earth's atmosphere. These atmospheric leptons form the background for the astrophysical leptons in the IceCube experiment. These leptons can be classified as conventional because the mesons that decay to a muon and a muon-neutrino lose a lot of their energy due to their long lifetime. This results in a lower energy distribution that is proportional to $E^{-3.7}$. Moreover, some atmospheric muons and neutrinos resemble the energy distribution of astrophysical leptons much more and can be classified as prompt. They originate from heavier hadrons like D mesons or Λ_C baryons created by high-energy interactions with the atmosphere. They have a short lifetime that leaves them less time to lose their energy. The neutrinos from these prompt processes are considered the signal for this analysis.

2.2 IceCube Neutrino Observatory

The IceCube experiment is located at a depth of 1450 m-2450 m at the south pole and its purpose is to measure the origin and energy of astrophysical neutrinos [1]. It detects the Cherenkov light produced by incoming charged particles with an array of 5160 photomultipliers. It is split into the DeepCore, in-ice array, and IceTop, which have different densities of photomultipliers for different energy ranges. The DeepCore as an example has the highest photomultiplier density and has an energy threshold of 10 GeV and the in-ice array one of 100 GeV. In addition, the IceTop is located at the top of the experiment and is used as a shower detector and veto for discarding atmospheric muons.

Neutrinos only interact with matter through the weak interaction with the water molecules mediated by charged currents (CC) and neutral currents (NC) through processes like

$$\begin{aligned} \nu_l \bar{\nu}_l + A &= l^{\mp} + X \,, \text{and} \\ \nu_l + A &= \nu_l + X \,. \end{aligned}$$

The leptons from charged currents can be differentiated by the signature that they leave in the detector. Electrons leave spherical showers, and muons leave a long trace of light because they do not lose that much energy due to their longer lifetime. On the other hand, tau signatures have signatures that are comparable to the signatures of electrons because they decay much faster. Additionally, the hadronic particles of neutral current interactions induce cascades that are similar to the ones caused by electrons. Moreover, the cascade events have a good energy resolution and a poor angular resolution. The reverse is true for the muons that just pass through the detector.

3 Analysis Strategy

For this analysis, Monte Carlo generated data will be used. The background and signal parts of the dataset get assigned a label to later train machine learning models on them. Moreover, features not present for background and signal are removed and the datasets will be cleaned from missing or infinity values.

3.1 mRMR Selection

The minimum redundancy, maximum relevance (mRMR) selection will be used to select the best features for the training of the models. It utilizes the joint information of two variables, x and y,

$$I(x,y) = \int p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy\,, \tag{2}$$

where $p(x)$, $p(y)$, and $p(x,y)$ are probability densities of the corresponding variables. The mRMR algorithm is iterative and independent of the applied learner.

3.2 Naive Bayes

The naive Bayes classifier is based on the Bayes theorem and treats the probability $P(B|A)$ is only dependent on B, where $A/\bar{A}$ stands for signal/background. It calculates

$$Q = \prod_{i=1}^{n} \frac{P(B_i|A)}{P(B_i|\bar{A})}\,, \tag{3}$$

where a value greater than one classifies it as signal and a value less than one as background.

3.3 Random Forest Classifier

The random forest classifier uses the average result of many binary decision trees. Binary decision trees apply cuts to nodes of the dataset to form branches. Each branch is cut until a certain depth of the decision tree is reached or there are no possible cuts anymore. The confidence or signalness c of the random forest classifier is given by the arithmetic mean of the N decisions P_i trees by

$$c = \frac{1}{N}\sum_{i=1}^{N} P_i\,, \tag{4}$$

where P_i ranges between 0 and 1. To classify an event as signal a threshold τ_c must be chosen.

3.4 kNN Classifier

The k-nearest neighbour (kNN) classifier computes the mean of the k-nearest neighbours to classify the data. The number of nearest neighbours k and the metric used to calculate the distance must be chosen according to the problem at hand to yield the best result.

3.5 Model Evaluation

The precision p and the recall r are defined by

$$p = \frac{TP}{TP + FP}\,, \tag{5}$$

$$r = \frac{TP}{TP + FN}\,, \tag{6}$$

where tp, fp, and fn are the number of true positive, false positive, and false negative classified events. The ideal case would have precision and recall being close to one. This cannot always be achieved but by setting the β parameter of the f_β score

$$f_\beta = (1 + \beta^2)\frac{pr}{\beta^2 p + r} \tag{7}$$

not to one it is possible to give a higher importance to precision or recall. Additionally, the receiver operating characteristic (ROC) curve can be used to evaluate the quality of the prediction. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) given by

$$TPR = \frac{TP}{TP + FN}\,, \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

for a given value of τ_c. The area under the curve (AUC) score is 1 for a perfect model and 0.5 for a random classifier and can be used as a performance metric.

4 Analysis

4.1 Data Preprocessing

The signal and background data for this experiment have been simulated separately. Once they have been loaded, only common attributes of the two datasets can be kept. Subsequently, the two datasets are merged and shuffled. All the "truths" coming from Monte Carlo simulations (namely, the features containing the terms Corsika, MC, Weight and I3EventHeader) are removed and values that are not identified as numbers, i.e. NaN (not a number) or inf, are handled in the following way: features with a number of invalid values greater than 10% are dropped and data with single randomly distributed invalid values are just not considered. Eventually, a label attribute is assigned to the data and then separated. This feature will be the training target of the learning methods.

4.2 Attribute Selection

The dataset contains many features and in order to reduce the computation time during the machine learning processes, the set of attributes must be reduced by only considering the

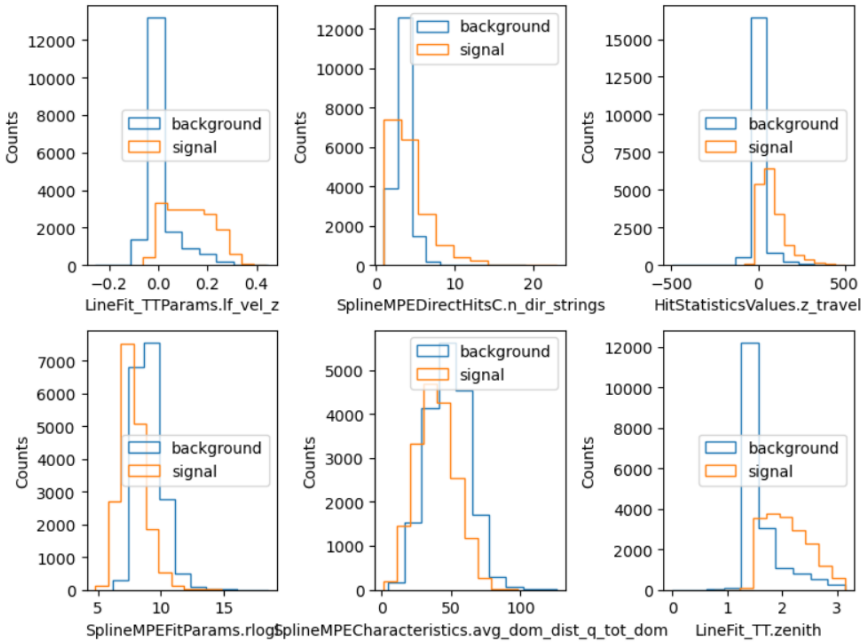

Figure 1: Distribution of the first 6 selected features.

features containing the most relevant information to separate signal from background, which is our final goal. This has been done by the implementation of the mRMR selection method, discussed in subsection 3.1. The chosen number of features is 15. The distribution of the first six is depicted in Figure 1 as an example.

4.3 Multivariate analysis

As previously explained, three different machine learning methods have been tested. At the end, only the best-performing model will be used to predict the labels on the test data set.

4.3.1 Random Forest Classifier

This algorithm, presented in subsection 3.3, has been trained using a $n\_estimators = 100$, which represents thee number of trees in the forest. The resulting confusion matrix from the application of the RF classifier on the testing set (after having trained it on the training set) is shown in Figure 2, while the corresponding ROC curve is plotted in Figure 3. The quality parameters obtained from this method were:

$$\text{Accuracy} = 0.9428$$
$$\text{Precision} = 0.9566$$
$$\text{ROC AUC} = 0.9846$$

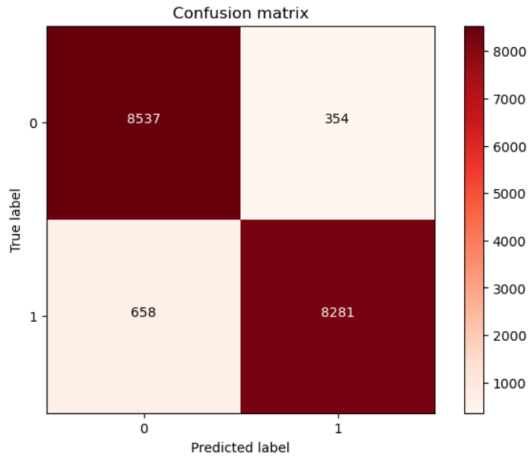

Figure 2: Confusion matrix from the application of the Random Forest Classifier algorithm

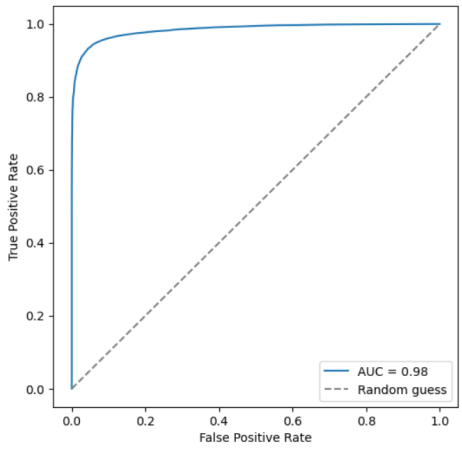

Figure 3: ROC curve with corresponding AUC value for the Random Forest Classifier

4.3.2 kNN classifier

The same thing has been done for the kNN classifier, described in subsection 3.4. The number of neighbours selected for this task was 5. The confusion matrix in this case is depicted in Figure 4, whereas the corresponding ROC curve is shown in Figure 5. This time, the values obtained from the quality parameters were:

$$\begin{aligned} \text{Accuracy} &= 0.9264 \\ \text{Precision} &= 0.9335 \\ \text{ROC AUC} &= 0.9665, \end{aligned}$$

slightly worse than the previous one.

4.3.3 Naive Bayes

Finally, the Naive Bayes algorithm, characterised in subsection 3.2, has been tested. The confusion matrix and ROC curve are shown respectively on Figure 6 and Figure 7. In this last case, the quality values are not great:

$$\begin{aligned} \text{Accuracy} &= 0.8305 \\ \text{Precision} &= 0.8858 \\ \text{ROC AUC} &= 0.9179 \end{aligned}$$

The trial on the testing set, after having trained the algorithms on the training set, revealed that the most performing machine learner is the Random Forest Classifier. Hence, that is the one which will be used for the final task.

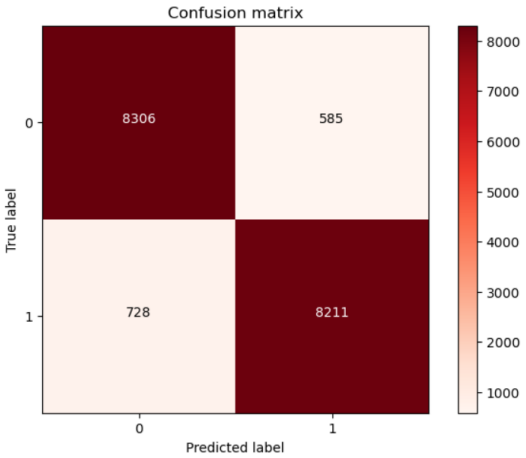

Figure 4: Confusion matrix from the application of the kNN algorithm

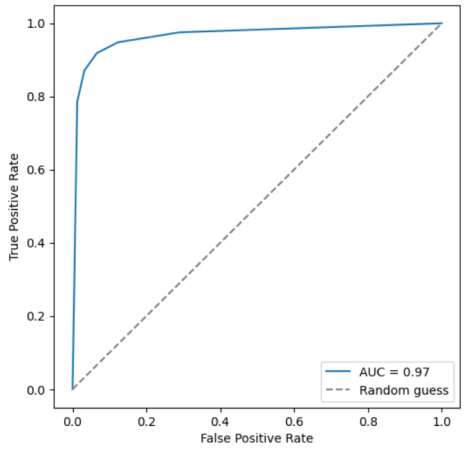

Figure 5: ROC curve with corresponding AUC value for the kNN algorithm

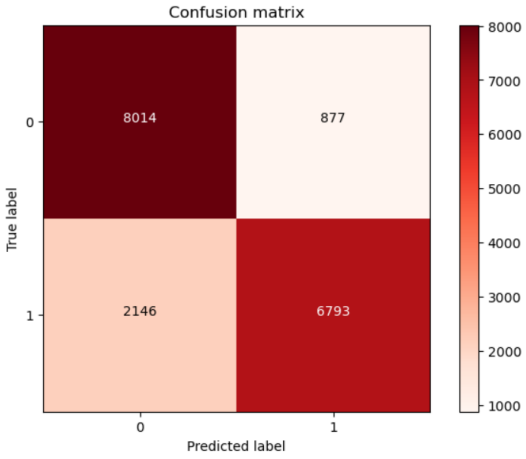

Figure 6: Confusion matrix from the application of the Naive Bayes algorithm

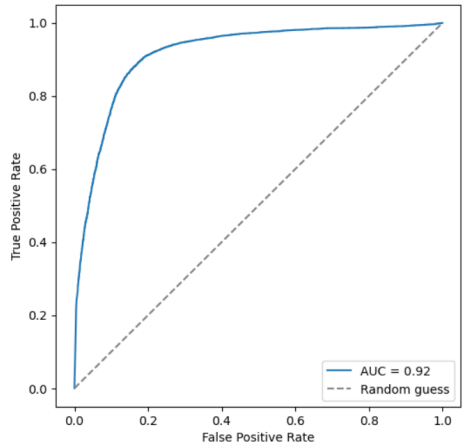

Figure 7: ROC curve with corresponding AUC value for the Naive Bayes algorithm

4.4 Cross-validation

The chosen model is then trained on the training dataset using a k-fold cross-validation method. The dataset is first partitioned into k equally sized folds. Subsequently, k iterations of training and validation are performed such that within each iteration a different fold of the data is held out for validation while the remaining k - 1 folds are used for learning. The final performance given by the k-fold cross-validation is the average of the values computed in the loop. The advantage is that now an uncertainty value can be assigned.

The accuracy after this cross-validation process using the Random Forest Classifier is:

$$\text{Accuracy} = 94.35 \pm 0.17$$

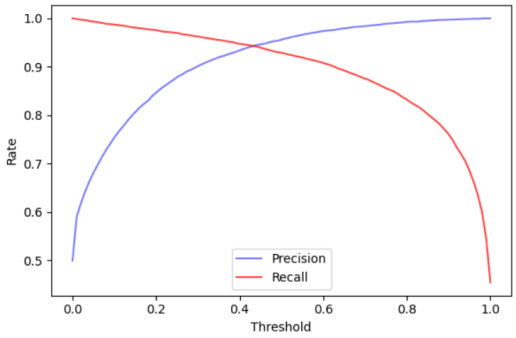

Figure 8: Precision and recall values as functions of the threshold value

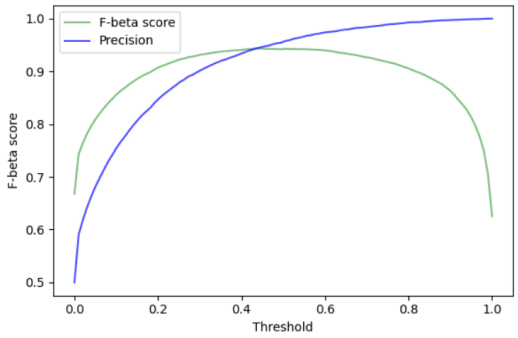

Figure 9: Precision and f_β values as functions of the threshold value

4.5 Model evaluation

The parameters discussed in subsection 3.5 are computed for the chosen model for different threshold values. In Figure 8, the *precision* and *recall* parameters, obtained from Equation 5 and Equation 6, are shown as functions of the threshold value. In order to find the best threshold, the f_β value is computed as well from Equation 7 using $\beta = 1$. Its behaviour is depicted in Figure 9, together with the precision and both as functions of the threshold value. The chosen threshold value is the one that maximizes the f_β score. In this case, the highest f_β value is $f_\beta = 0.945$ and it is reached at a threshold value $\tau = 0.475$. The model with the above-defined threshold is finally applied to the test dataset, obtaining the predictions that are then saved in the file *predictions.csv*.

5 Conclusion

The goal of the experiment was to test the performance of different machine learning algorithms in the task of recognising signal events from background in Monte Carlo simulated data from the IceCube experiment. After a preprocessing phase on the dataset, three different algorithms have been tested. Considering the quality parameters, such as accuracy and area under the ROC curve, the Random Forest Classifier resulted to be the best-performing algorithm. Said machine learner, with the prediction threshold maximizing the f_β score, has then been trained and tested with the *test.csv* file. The results are then saved in the *predictions.csv* file, attached to this report.

References

[1] IceCube Collaboration. Observation of high-energy neutrinos from the galactic plane. *Science*, 380(6652):1338–1343, 2023.

[2] R. Abbasi et al. Icetop: The surface component of icecube. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 700:188–220, 2013.