MAS6019-200234131-ASSIGNMENT2

Introduction. This report will analyse various machine methods in there capability to accurately model the data set gamma.csv. This data set consists of 10000 entries of data on high energy particles attained from an atmospheric telescope. The aim of the report is to distinguish which method best classifies the particles into gamma or hadrons. The particles have 10 features including variables such as Length, Width, Size, Conc (concentration of brightness), Conc1 (maximum brightness) and Asym(Distance from brightest pixel to the center).

The dataset will be split into training, validation and a final test set. The training and validation set make up 80% of the total data, with a ratio of $7:1$. The 7000 entries in the training set are used to fit the model, while the 1000 validation entries are used to tune parameters for Support Vector Machines and Neural Networks. In order to evaluate the models, each method will be fitted to the same dataset of 1000 entries (test set). This dataset is kept completely separate from all testing and parameter tuning to prevent any information leakage to the test set. Before the data is split, the class data was processed into a binary class rather than 'g' or 'h'. The methods will be evaluated using the accuracy of this model on this test set, while specificity and sensitivity will not be used as this task is asking for the general classification accuracy of gamma and hadron particles. All accuracy metrics will be given to 3 decimal places.

Logistic Regression (LR). For LR there was only one parameters to tune, this was the cost function. The cost function is used to specify the type of penalization of the model. However, after altering the cost term, the function which gave the highest accuracy score on the test set was the default function with an accuracy score of 0.764.

Discriminant Analysis (DA). Linear DA, similarly to LR had no real parameters. The model was fitted to the training data and then tested on the test data. Predictions of the training and test data were calculated and these gave accuracy scores of 0.761 and 0.767 respectively. A 10-fold cross-validation was executed for the LDA model, which produced a mean score of 0.760. Quadratic DA was also tested on the training data and again on the test data, however this performed worse that LDA and so was not used in the comparisons.

Support Vector Machines (SVMs). The SVM used for the classification was the Supoort Vector Classifier. The model was tested using three kernels, including linear (default), polynomial and the radial basis function. A validation set of size 1000 was split off from the training set at the start of the method and was used to tune the parameters. A grid search was run for each kernel to find the optimal parameters for the cost value C $(1, 2..., 10)$ and the gamma value $(0.1, 0.5, 1, 2)$. The optimum values for the cost C and gamma were 10 and 0.1 respectively. For each kernel, the model was run on the validation set. Processing time was extremely slow even when running the even the linear kernel. The linear kernel gave a training accuracy of 0.754 and a test accuracy of 0.767. Suggesting low amounts of over fitting occurred but this is a weak prediction of the test data.

Neural Networks (NN). In order to start the NN, the input data had to be standardised. This is because the features were measured in different units and standardization allows the data to have a common scale without distorting the ranges of data. For the NN model, the large amount of parameter tuning mean that a validation set of size 1000 was again taken from the training data. This stopped any information leakage which would distort the test accuracy score.

To optimise the networks architecture, a number of different models were attempted, including models with the number of hidden layers $(1, 2, 3)$ and the number of nodes at each layer $(5, 10, 25, 45, 75, 100,\ 150)$. Each model was fitted to the training set and then testing on the validation set to allow for parameter tuning. Once the parameters had been tuned on the validation set, the model with the highest accuracy score on the validation set was tested on the final test set. The architecture with the highest accuracy on the test set was that of 2 hidden layers, the first hidden layer has 100 nodes and the second layer has 45 nodes. The final model is extremely wide and short with the activation function of the first 2 layers being 'relu' and with the output layer using the 'sigmoid' activation function.The optimizer for the model was 'Adam' and it was compiled using the 'binary crossentropy' loss function. The architecture also used a dropout function after the first hidden layer. This function drops units randomly which means that it did not allow the model to rely too heavily on variables which may have lead to overfitting. The dropout function increased the models performance when compared to when it was not used in the architecture. A grid search was used to find the best value for some of the parameters. This included the optimum

size of the batch size. The grid search found the optimum number for batch size to be 20. The training accuracy was plotted against the validation score, along the number of epochs. Figure 1 shows that both the validation and training score reached 0.8 after only $10-15$ epochs. The validation accuracy began to level off and reached it maximum at around 90 to 100.

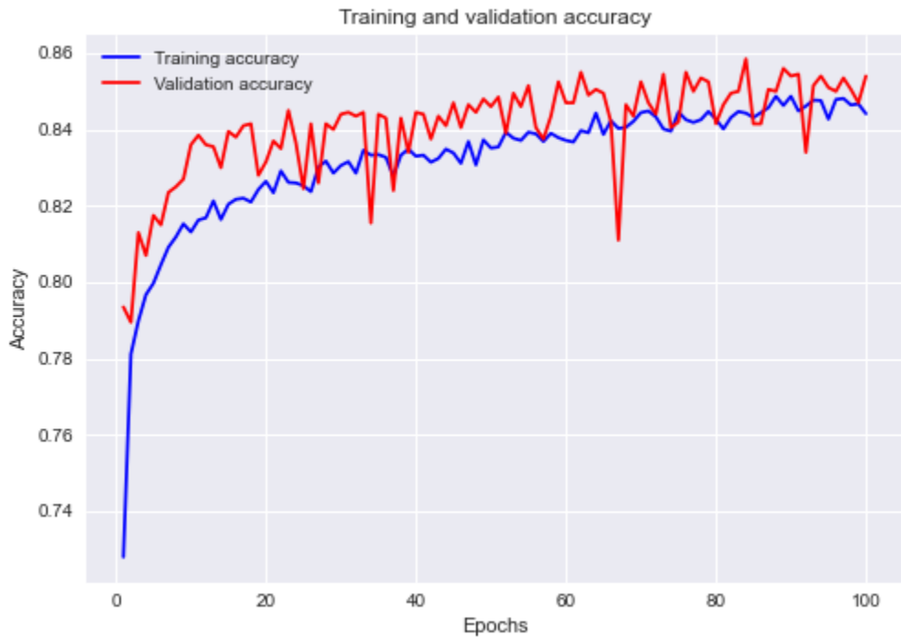

FIGURE 1. The training and validation accuracy after each epoch

The final model was tested on the test data and produced an accuracy score of 0.806. The neural networks took $2-3$ minutes to run and were initially over fitting the data.

Decision Trees. For the decision tree method, the package DecisionTreeClassifier was used for the model. This model was fitted to the training values while the parameter for maximum depth was tested on different values, including $(3, 6, 9, 10, 12, 15)$. The optimum max depth value was 10 which attained an accuracy score of 0.804 on the test set. A 10-fold cross-validation on the this model gave an accuracy score of 0.806.

Random Forests (RF). The RF model uses a ravelled training set. The ravel function is used to flatten the array of data. The parameter of the model, the max depth, was added to a grid search and the optimal max depth was 27. The number of estimators was set to 50 as this gave the highest accuracy against values of 100 and 200. This model gave an accuracy score of 0.857 on the test data. A 10-fold cross-validation gave an accuracy score of 0.852.

Gradient Boosting (GB). For the gradient boosting machines, both the adaptive gradient boosting and the extreme gradient boosting were used to test the difference in processing time and accuracy score. The maximum depth and minimum samples split parameters were tested for different values including $(5, 7, 9, 11, 13, 15, 17, 19, 20)$ and $(200, 400, 600, 800, 1000)$ respectively. The optimal values were found as being max depth 17 and 400 for the minimum sample splits. The number of estimators was experimented with at multiple values and 50 was concluded at being the optimal value. This gave the shortest processing time and the least overfitting. These parameters were then used in the model which was fitted to the training data. This model achieved a test accuracy of 0.854 and 0.850 after a 10-fold cross-validation.

The Extreme Gradient Boosting classifier was used on the same training data to test if the accuracy could increase further. The XGB model was also used to find the importance of the variables in the data, shown in Figure 2. The accuracy score on the test set was 0.830 and the accuracy score after a 10-fold cross-validation was 0.845. This XGB method scored lower in accuracy and therefore the GB classifier was the preferred method.

Figure 2 plots the score of feature importance. Features 1,2,3 and 9 have the largest importance in the classification of the particles. These features correspond to Length, Width, Size (the log of the total brightness of the ellipse) and Alpha (the angle of the major axis to the axis of the telescope). This suggests that the main difference between gamma and hadron particles is their size, with the dominating feature being the width.

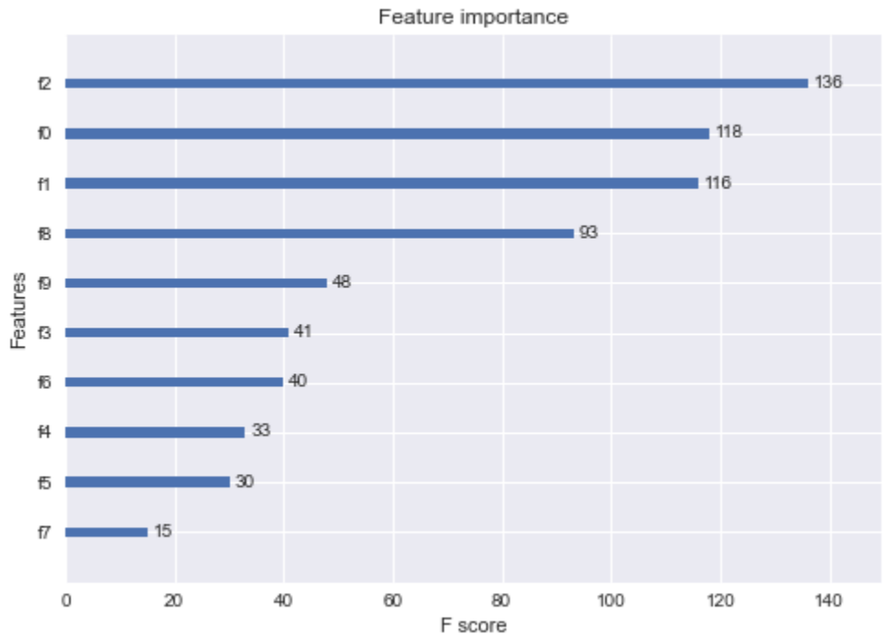

FIGURE 2. Importance of features.

Method	Cross Validation Accuracy
LR	0.764
DA	0.767
SVM	0.768
NN	0.806
DT	0.815
GB	0.850
RF	0.852

TABLE 1. Method Accuracy.

Conclusion. Based on the accuracy score of each method as seen in Table 1, Random Forests with a maximum depth 27 and n estimators 50 had the highest score on the test set as well as a relatively quick processing time in comparison to SVMs and NNs. Therefore we believe this model will perform with the highest accuracy on your test set. This model is able to estimate unknown data once it is trained on the training data and so for the secret test set we make the prediction that this model will attain an accuracy around 0.840 to 0.854, as this was the range the model was scoring after numerous runs.

The RF method is easy to use as it fits lots of different data types and so acts as a good starting point to begin solving a problem. The model performs well with high accuracy, as well as the ability to model and predict missing values. However, this method can be slow at training and is prone to over fitting due to the randomness of sampling and variable selection.