Report: NASA Asteroid Classification using MLPClassifier model

A total of 4687 examples with 40 features in the NASA asteroid dataset were analyzed to classify their hazardousness.

The procedure is as follows: (see detailed coding at:

-Some non-relevant features were excluded from the analysis processing

-Extract 2 groups (X and y) of datasets: X=31 relevant features; y=actual classified results (hazardous or non-hazardous), this data is already binarized.

-Perform normalization to set all features of X in the same range: Scaling process

-Split (X, y) into 2 separate corresponding groups: one for training, one for testing

-Perform experiments by applying MLPClassifier with different setting of hidden layers for each group (training, testing). The number of iterations are the same in all experiments (1000 iterations). In each iteration, the accuracy of classification was calculated, recorded and analyzed.

Evaluation results are summarized as below:

Table 1: SUMMARY OF EXPERIMENT'S RESULTS USING DIFFERENT NEURONAL NETWORK SETTING

Number of hidden layer	Number of 1st hidden layer	Number of 2nd hidden layer	Number of 3rd hidden layer	Overall Training accuracy	Overall Test Accuracy
1	1	0	0	0.956	0.952
	5	0	0	0.995	0.994
	10	0	0	0.998	0.988
	20	0	0	1.0	0.986
	30	0	0	1.0	0.985
2	1	1	0	0.832	0.856
	1	5	0	0.962	0.951
	1	10	0	0.967	0.964
	5	1	0	0.999	0.990
	5	5	0	0.999	0.991
	5	10	0	0.999	0.990
	10	1	0	0.998	0.989
	10	5	0	1.0	0.990
3	1	1	1	0.834	0.848
	5	1	1	0.834	0.848
	5	5	1	0.998	0.987
	5	10	1	0.999	0.990
	5	1	5	0.834	0.848
	5	5	5	0.997	0.989

Note: Yellow label ➔ best cases

Below are some plots in the typical experiment using 2 hidden layers, one has 10 neurons and one with 5 neurons.

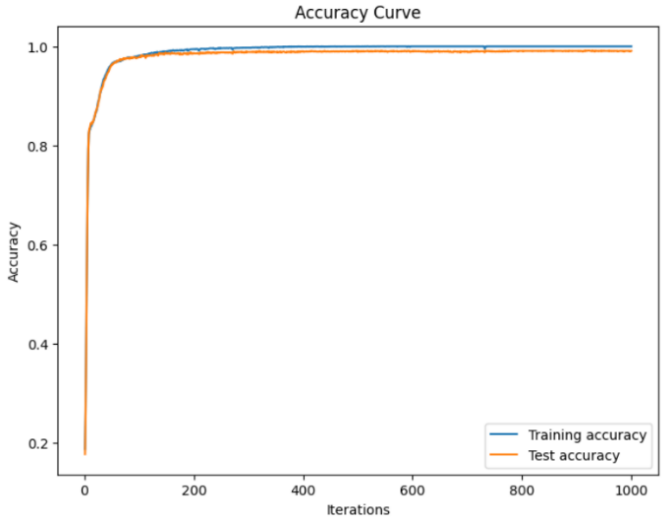


FIGURE 1

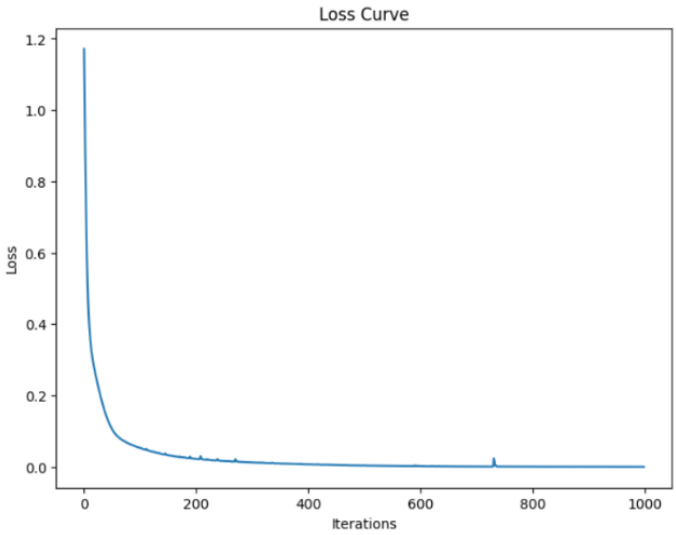
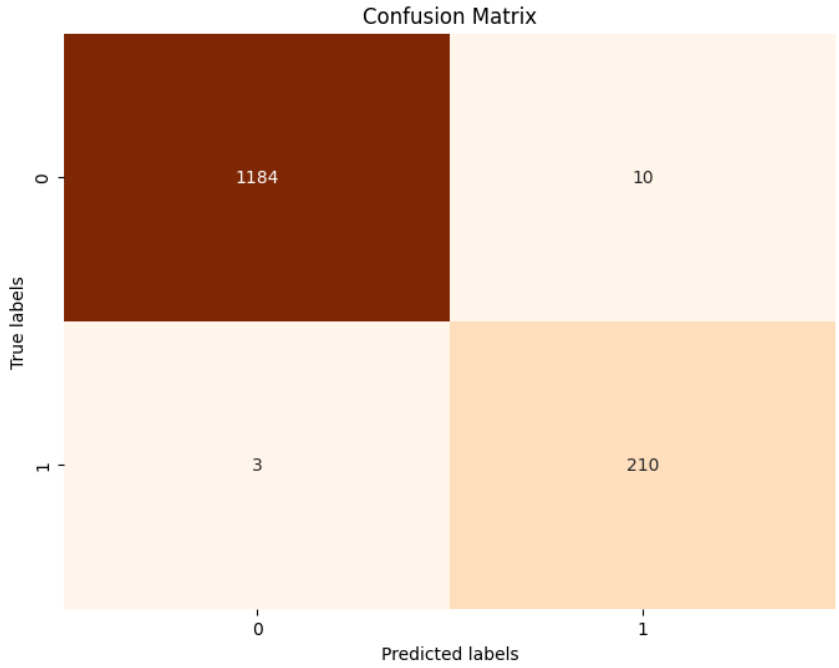


FIGURE 2

From all performed experiments, we can see that MLPClassifier can give a great improvement in accuracy of classification after around 100 iterations (see typical accuracy curve, Fig 1). This is reconfirmed when the difference between the predicted values of the model and the actual values in the training dataset was measured by the loss function (Fig 2).

For both training and testing process, the overall accuracies of 80%-100% were observed in different experimental conditions (Tabe 1). Under the setting of 2 hidden layers, one with 10 neurons and one with 5 neurons, we obtained the best overall accuracy for both training and testing processes (100% and 99%, respectively) (Table 1- yellow labeled).

FIGURE 3



In the best classification process performed on a testing group of 1407 examples (sum of 1184+10+3+210) (see Fig 3), MLPClassifier can predict about 99.08% ((1184+210)/1407) of examples with the same result as actually classified (Fig 3), in which the classification result of True Negative accounts for 84.16% (1184/1407) and True Positive 14.92% (210/1407), respectively. In other words, the model has a great possibility in prediction of asteroid's hazardousness, with 99.08% correct and only 0.9% false predictions.