Photometric Identification of Infrared Sources

Machine Learning Project* Phase 2: **Traditional Techniques**

May 8, 2021

1 Introduction

So far, we have prepared, processed, and analyzed our data. In this phase, we are ready to apply different classification models to it. Our data consists of many infrared sources spectrums that we are going to classify into five classes —YSO, O-rich, C-rich, S-rich, post-AGB.

2 Measures of Evaluation

In this phase we are ready to train different classification models on our data. To measure the performances of models, we first calculated the confusion matrix to find the values of *TP, FP, TN, FN*; and then calculated the metrics to optimize our models: precision, recall, and F_1 score.

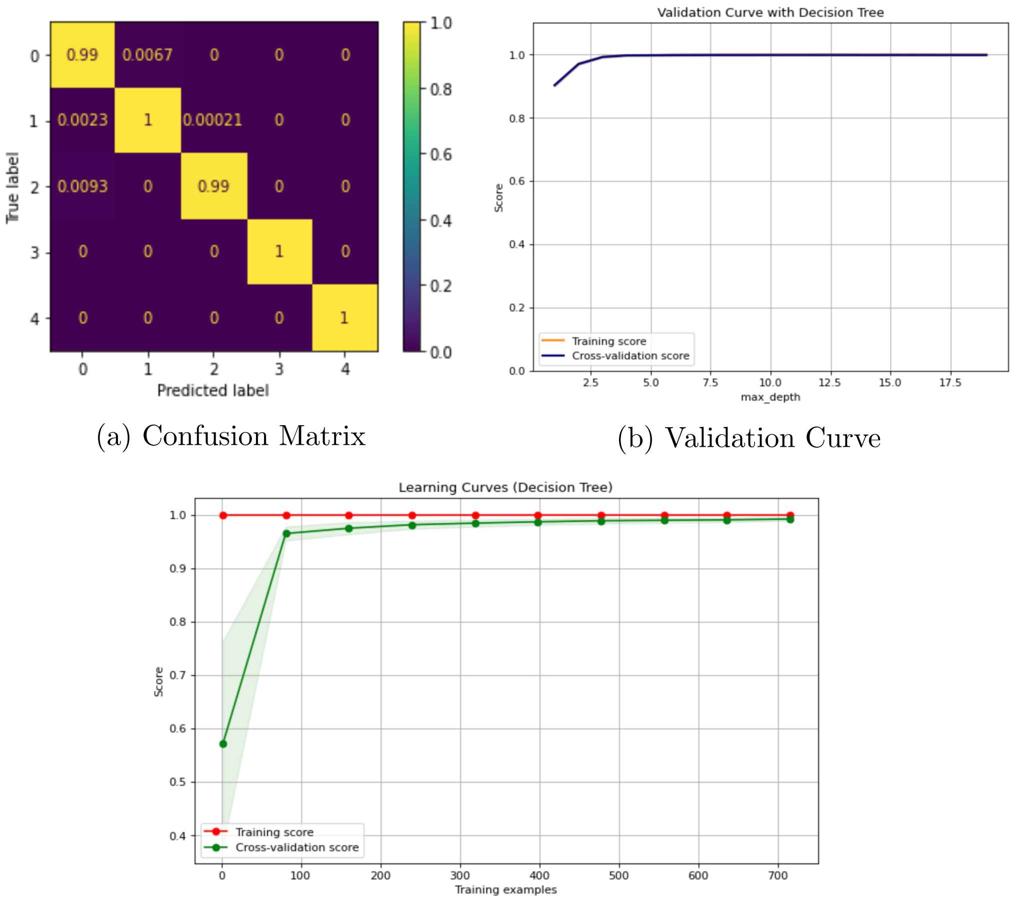
3 Train Different Models

We applied six classification models to our data: decision tree, SGD, SVM, KNN, LDA, and Gaussian NB. These classifiers were applied to our data in three different conditions: raw data, scaled data, scaled and transformed data —using PCA. For each model, we chose the results in the condition with best score.

*Group members: Mobin Saidy, Reyhane Javadi, Amirhosein Masoudnezhad, Negin Khosravaninezhad

3.1 Decision Tree

One of the classifiers we tested on our data is decision tree. We used gini criterion to optimize this model. This model had its best performance on unscaled transformed data. It can be seen in Figure 1 that the learning curve has converged quickly so we can say that our data is adequate and the decision tree model is being trained fairly well.



(a) Confusion Matrix

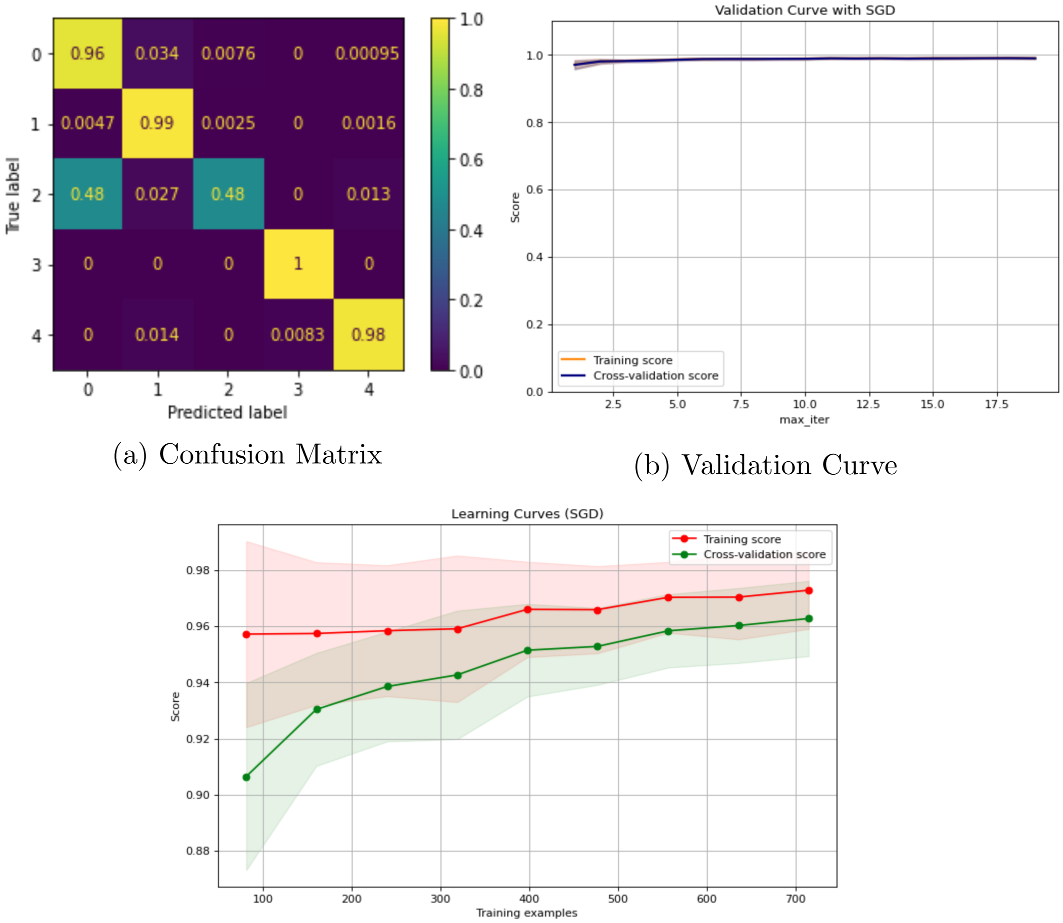
(b) Validation Curve

(c) Learning Curve

Figure 1: Results of the Decision Tree Model on unscaled data with PCA

3.2 Stochastic Gradient Descent (SGD)

Another classifier we tested on our data is SGD. We used hinge loss function to optimize this model. This model had its best performance on standard scaled data without transformation. It can be seen in Figure 2 that the learning curve has converged almost quickly. This was not the best fit though it was a good fit, given that this classification model generally has some fluctuations.



(a) Confusion Matrix

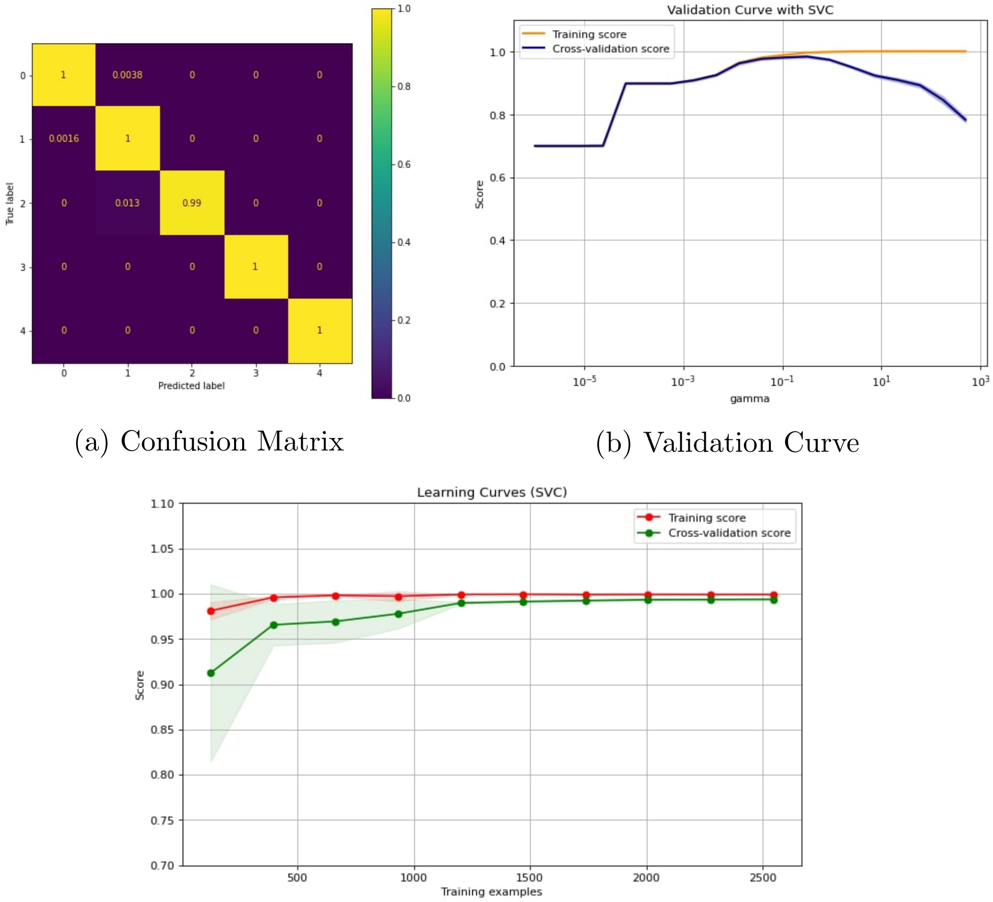
(b) Validation Curve

(c) Learning Curve

Figure 2: Results of the SGD Model on standard scaled data

3.3 Support-Vector Machine (SVM)

Another classifier we tested on our data is SVM. We used squared hinge loss function to optimize this model. This model had its best performance on standard scaled data with PCA. It can be seen in Figure 3 that the learning curve has converged quickly so we can say that our data is adequate and the SVM model is being trained fairly well.



(a) Confusion Matrix

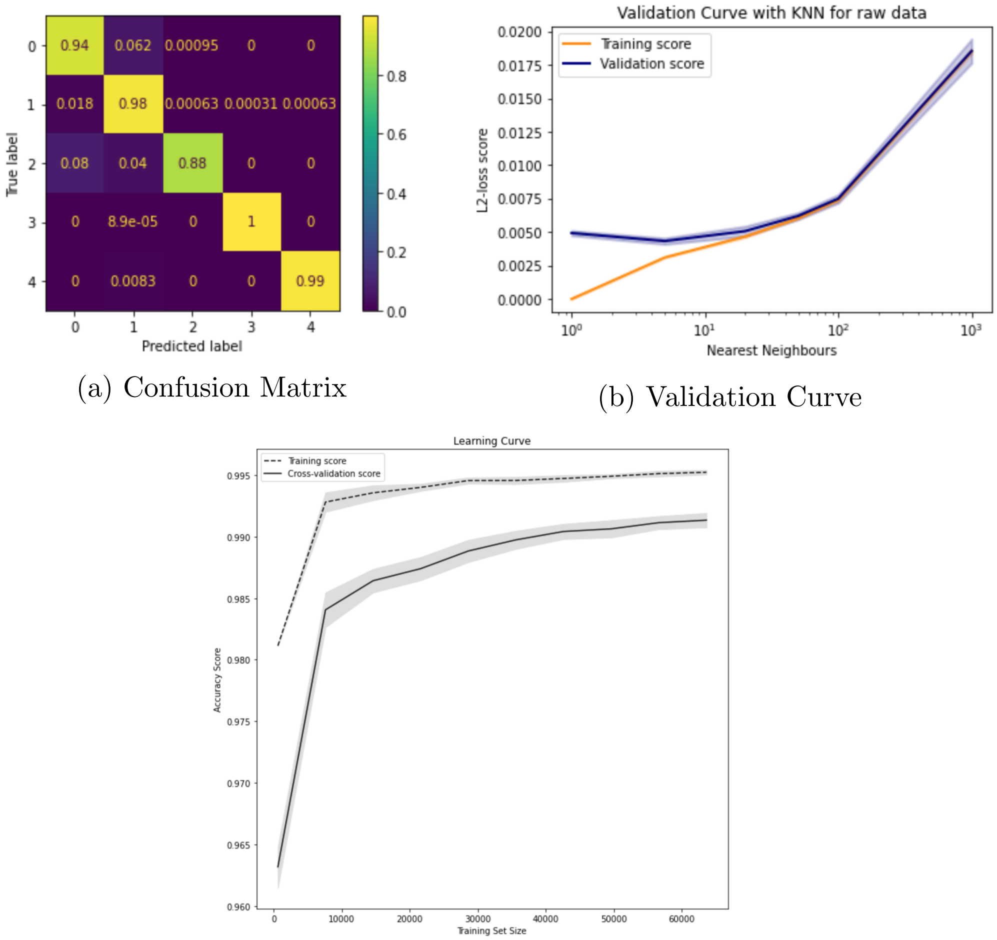
(b) Validation Curve

(c) Learning Curve

Figure 3: Results of the SVM Model on standard scaled data with PCA

3.4 K-Nearest Neighbor (KNN)

Another classifier we tested on our data is KNN. We used mean squared error loss function to optimize this model. This model had its best performance on standard scaled data. It can be seen in Figure 4 that the learning curve has converged quickly so we can say that our data is adequate and the KNN model is being trained fairly well.



(a) Confusion Matrix

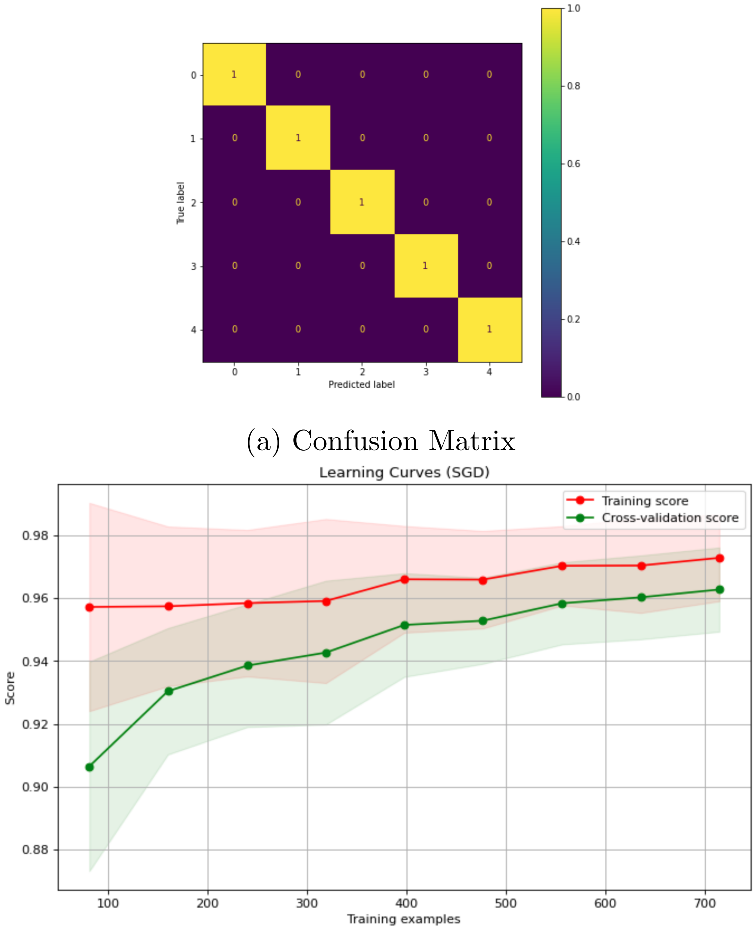
(b) Validation Curve

(c) Learning Curve

Figure 4: Results of the KNN Model on standard scaled data without PCA

3.5 Linear Discriminant Analysis (LDA)

Another classifier we tested on our data is LDA. This model had its best performance on standard scaled data with PCA. It can be seen in Figure 5 that the learning curve has converged quickly so we can say that our data is adequate and the LDA model is being trained fairly well.



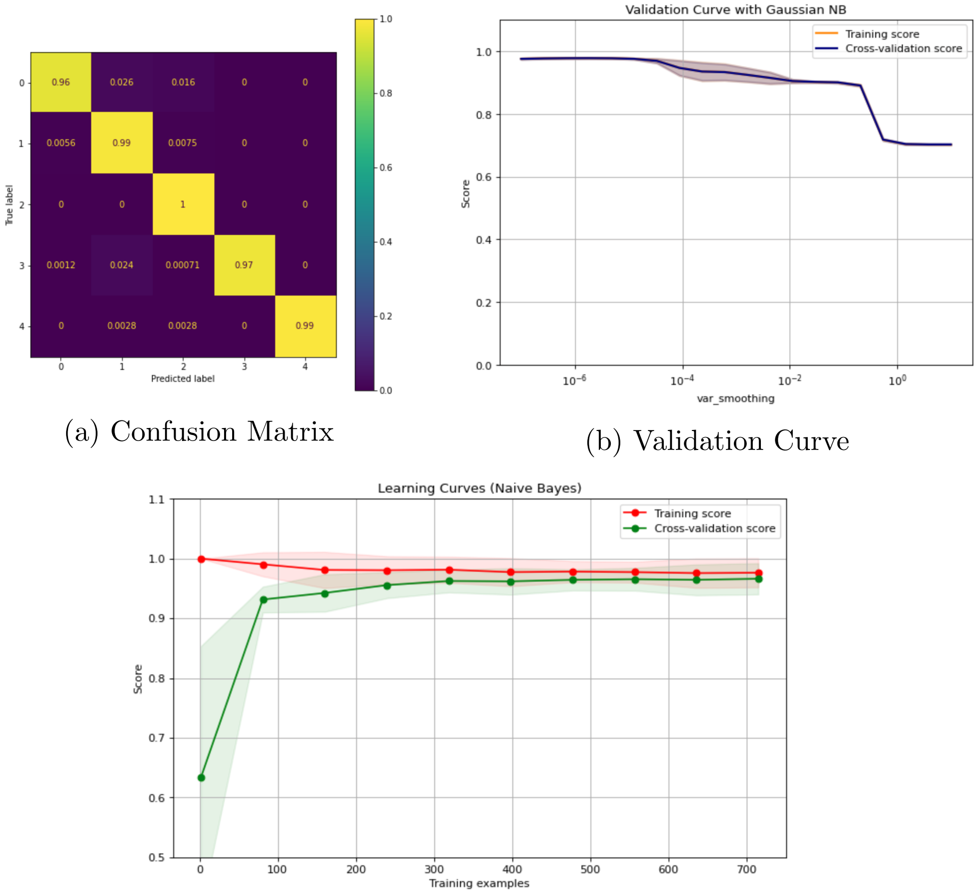
(a) Confusion Matrix

(b) Learning Curve

Figure 5: Results of the LDA Model on standard scaled data with PCA

3.6 Gaussian Naive Bayes

Another classifier we tested on our data is Gaussian NB. This model had its best performance on standard scaled data with PCA. And the reason is that it gives the possibility with Gaussian distributions. It can be seen in Figure 6 that the learning curve has converged quickly so we can say that our data is adequate and the Gaussian NB model is being trained fairly well.



(a) Confusion Matrix

(b) Validation Curve

(c) Learning Curve

Figure 6: Results of the Gaussian NB Model on standard scaled data with PCA

4 Comparison

What we are going to do know is compare the six models we ran on our data.

	Metrics			Times	
Model	Precision	Recall	F_1 Score	Training Time	Prediction Time
Decision Tree	1.00	1.00	1.00	16^s	0.2^s
SGD	0.99	0.99	0.99	23^s	1^s
SVM	1.00	1.00	1.00	2^m	4.3^s
KNN	0.99	0.99	0.99	2^m21^s	27^s
LDA	1.00	1.00	1.00	$46s$	0.3^s
Gaussian NB	0.98	0.98	0.98	20.9^s	0.2^s

5 Discussion

5.1 Model Selection

Considering almost all the learning curves of the models —that converged quickly— we can say that we do have adequate data. Most of the classifiers have small bias and variance as can be seen in their learning curves. This was actually predictable because our initial observational data was very clean. However, it can be said that the decision tree model, which has both the highest possible score and the shortest time, is the best one.

5.2 Metric Selection

In our case, as the scores of models were very high and all were close to each other, we cannot distinguish much between different metrics. But generally, for a classification problem like ours, F_1 score seems to give a more accurate score than the others as the number of data in different categories is very different, precision and recall may be falsely high, when in fact the model is not really good. But F_1 score considers the value of both to be reasonable and gives a more accurate measure of model performance.