Machine Learning: Classification Models

SEP 787 Course Project

Nazanin Moshtagh, 400046822 Aditya Goel, 400414715

January 2022

1 Objective

The goal of this work is to compare three different classification algorithms on one dataset. In this work, we investigated the models below:

1. K Nearest Neighbors (KNN)
2. Support Vector Machine (SVM)
3. Naive Bayes (NB)

The Receiver Operating Characteristic (ROC) curve and the confusion matrix are generated in order to have a comparison between these models.Moreover, the computational time for both trainig and testing are measured and reported.

2 Dataset

The dataset used here is called "Wine Quality Dataset" provided by UCI Machine Learning Repository and can be found here. We have picked the white 'Vinho Verde' wine samples dataset. The goal is to model wine quality based on physicochemical tests. It consists of 4898 instances of white wine. There are 11 attributes and all are continuous variables having a normal (Gaussian) probability distribution. The target variable is the quality that ranges from 0-10. In order to make the labels into binary form, we have combined the wines with quality of 0-5 and labeled them class 0. The remaining labels (quality: 6-10) are combined to form class 1. As stated in the project requirements, 75% of the data have been used for training and the remaining 25% is reserved for testing. We have preprocessed all features using Standard Normalisation.

3 Classification Models

3.1 K Nearest Neighbors Classifier

The optimum value of 'n' in n-fold cross validation is obtained by hit-and-trial with n ranging from 2 to 10. While no result was ideal, the best graph obtained was when n=4. Figure 1 shows the error vs. K value for a 4-fold cross validation (CV).

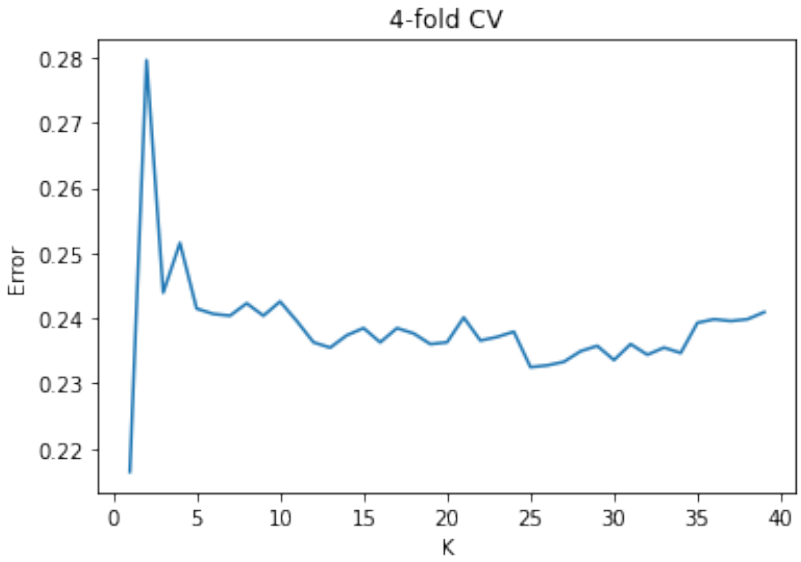

Figure 1: Error vs. K value in KNN model performing 4-fold cross validation.

Value of 'K' in KNN is determined by observing the above plot and using the elbow method. The error stabilizes between 0.235 and 0.24 beginning from K=14. Hence, the value chosen for K is 14.

Therefore the parameter values for KNN are as below:

```
K = 14

n for cross validation = 4
```

3.1.1 Training Phase:

First, the training data is fit in the classifier and its corresponding metric are obtained:

- Computational Time = 65.37 s
- Accuracy = 0.794

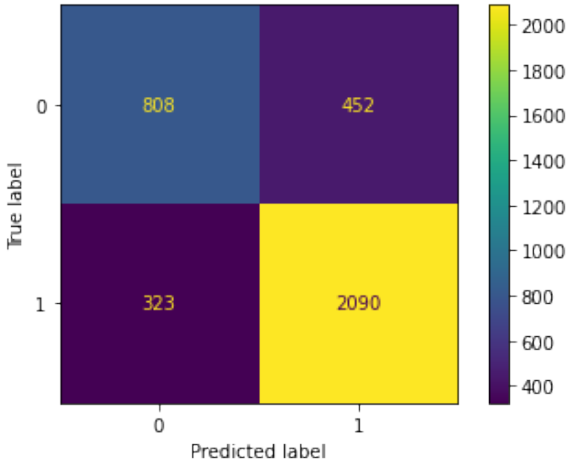

Figure 2: Confusion matrix for KNN training phase.

The ROC curve is shown in Figure 3 is created by the training data.

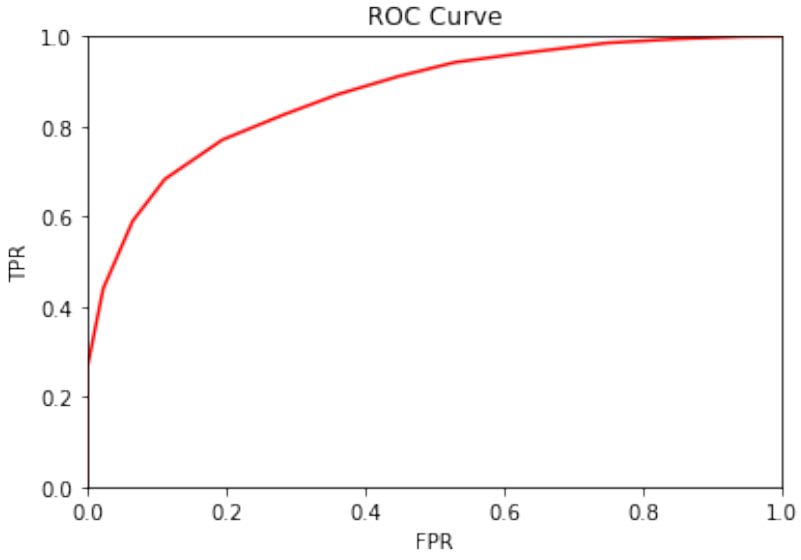

Figure 3: ROC curve, KNN performed on training dataset.

3.1.2 Testing Phase:

- Computational Time = 0.581 s
- Accuracy = 0.754

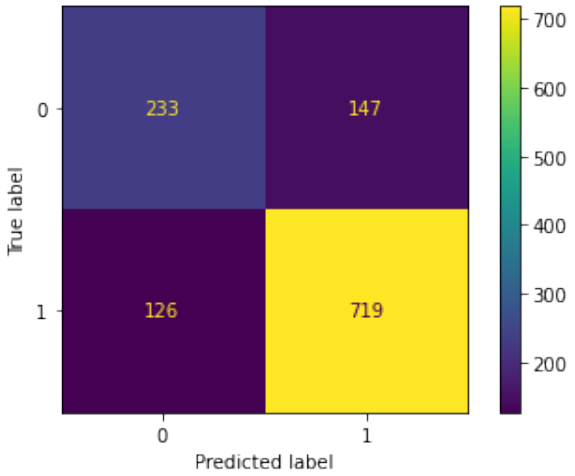

Figure 4: Confusion matrix for KNN testing phase.

The ROC curve is shown in Figure 5 is created by evaluating the trained model on the unseen or testing data.

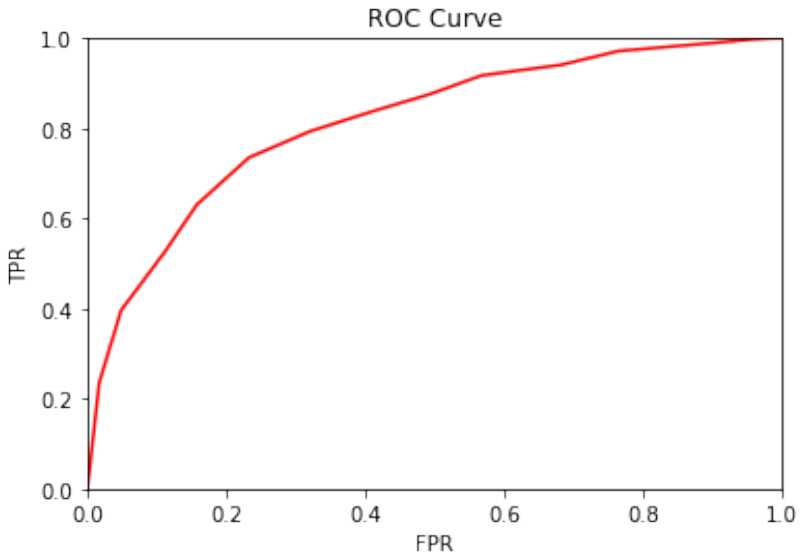

Figure 5: ROC curve, KNN performed on testing dataset.

3.2 Support Vector Machine Classifier

SVM is another model we picked to classify the white wine quality dataset. We performed Grid Search Cross Validation for tuning the hyperparameters. The best parameters found are:

```
C = 10

gamma = 0.01

kernel = rbf

n for cross validation = 4
```

The numbers reported below are the result of 4-fold cross validation.

3.2.1 Training Phase:

SVM model is trained on the training data and the confusion matrix and ROC curves are plotted below.

- Computational Time = 63.98 s
- Accuracy = 0.794

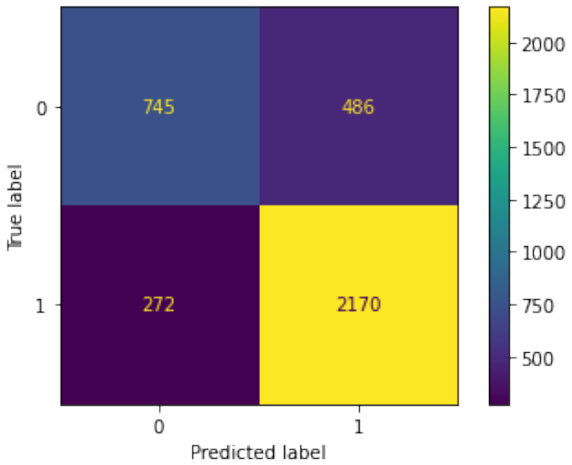

Figure 6: Confusion matrix for SVM training phase.

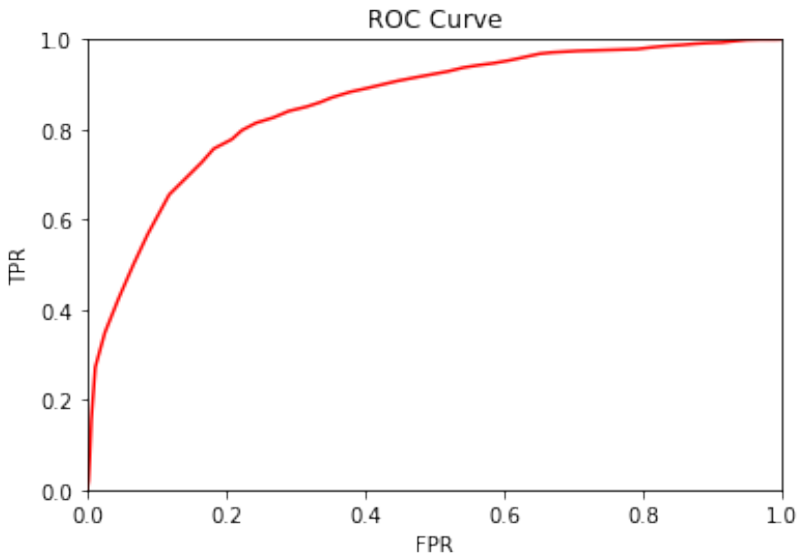

Figure 7: ROC curve, SVM performed on training dataset.

3.2.2 Testing Phase:

The remaining portion of dataset is used to evaluate the performance of the SVM model. Confusion matrix and ROC curve are presented below.

- Computational Time = 0.871 s
- Accuracy = 0.772

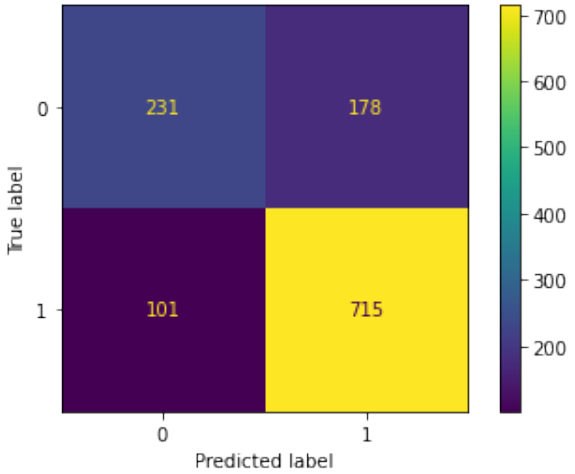

Figure 8: Confusion matrix for SVM testing phase.

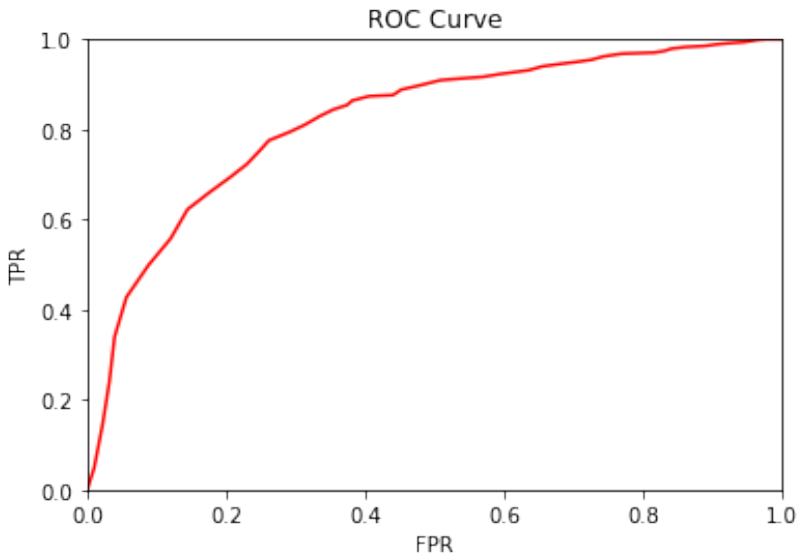

Figure 9: ROC curve, SVM performed on testing dataset.

3.3 Naive Bayes

In this model, n-fold Cross Validation is not required because there are no parameters. Since the features of the dataset are Continuous Variables with a Normal (Gaussian) Probability Distribution, their Prior Probabilities are computed by the in-built sklearn libraries.

3.3.1 Training Phase:

First, the training data is fit in the classifier and its corresponding metric are obtained:

- Computational Time = 0.899 s
- Accuracy = 0.701

The ROC curve is shown in Figure 11 is created by the training data and shows the true positive rate vs. the false positive rate.

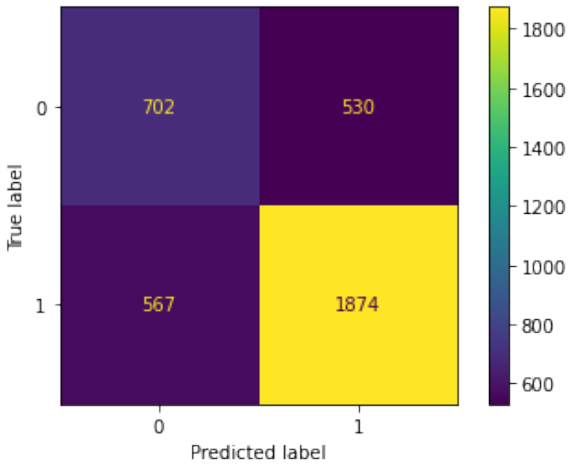

Figure 10: Confusion matrix for NB training phase.

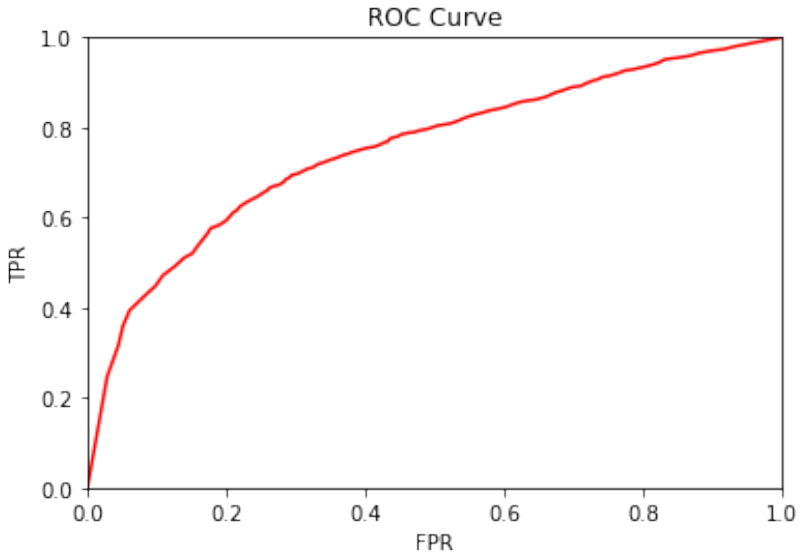

Figure 11: ROC curve, NB performed on training dataset.

3.3.2 Testing Phase:

- Computational Time = 0.589 s
- Accuracy = 0.705

The ROC curve is shown in Figure 13 is created by evaluating the trained model on the unseen or testing data.

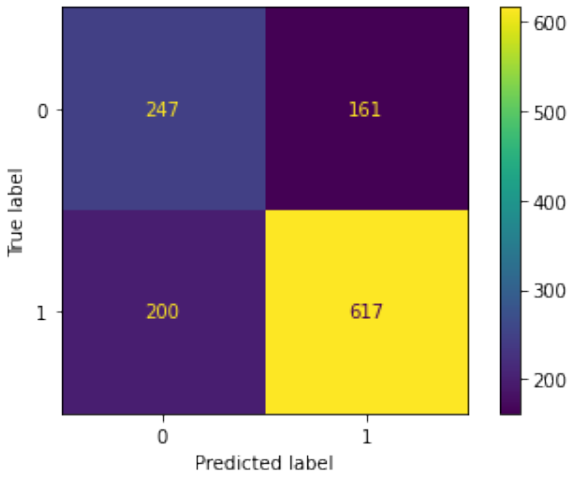

Figure 12: Confusion matrix for NB testing phase.

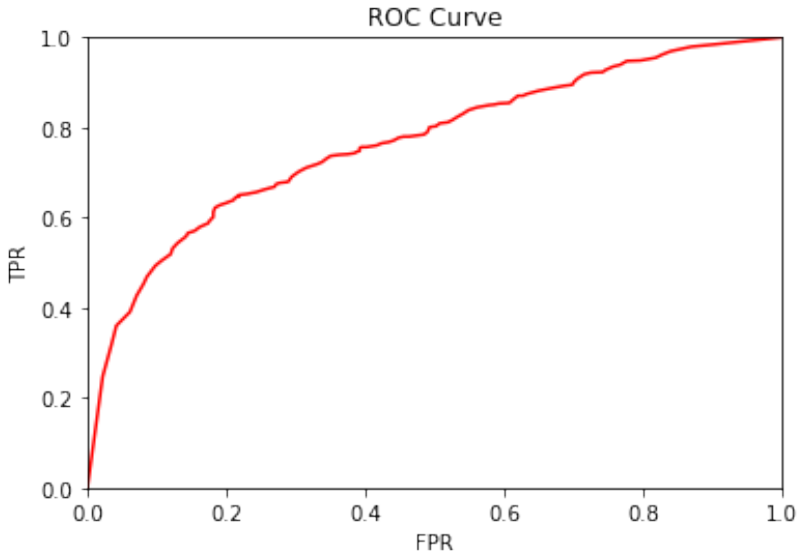

Figure 13: ROC curve, NB performed on testing dataset.

4 Conclusion

Table below shows the result of three different classifiers that we trained and tested in this work. KNN, SVM and NB can be compared using the computational time for training and testing, confusion matrix, and accuracy. Please note that due to data shuffling, if you run the submitted Notebook as you are looking at the results presented below, the numbers will slightly differ.

Model	Comp Time- Training (s)	Comp Time- Testing (s)	Test Accuracy
KNN	65.37	0.581	75.43%
SVM	63.98	0.871	77.22%
NB	0.899	0.589	70.53%

We have trained and tested KNN, SVM, and NB classifiers on the same dataset using a common test bed. We can see that KNN and SVM resulted in a higher accuracy in comparison with NB classifier. Moreover, considering the computational time for SVM and KNN, support vector machine classifier offers a slightly more efficient computation. Additionally, the ROC curve gives us a more accurate image of the classifier performance and as it is shown in figure below as they are compared side by side. The ROC curve for KNN shows a better performance in comparison with SVM.

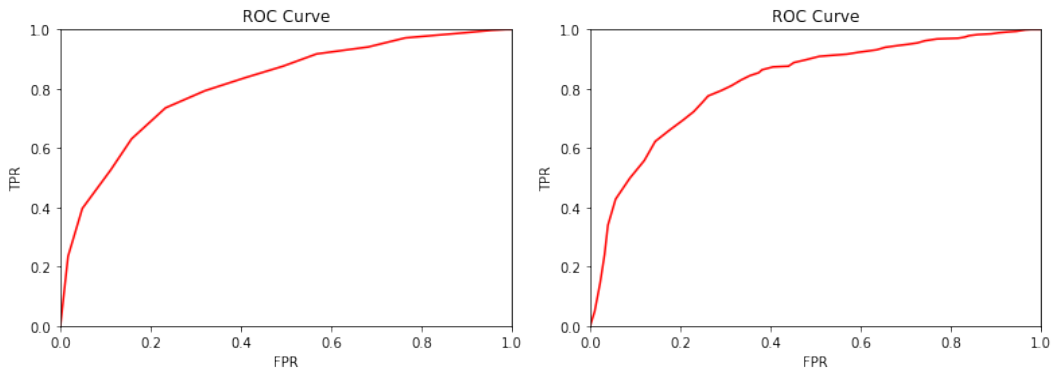

Figure 14: Comparing the ROC curve for KNN (left) and SVM (right) models in testing phase.

5 References

1. Dataset: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

2. Dr. Jeff Fortuna's SEP-787 lecture notes.

3. Scikit Learn, Machine Learning in Python