C S 487/519 Applied Machine Learning - Machine Learning Project Report

Vensan Cabardo, Likhitha Devineni, Camika Leiva

Motivation

Manually classifying dry bean seeds based on visual observations has proved to be difficult due to certain varieties of dry beans looking very similar to the human eye. This difficulty also increases at high production volumes, as a predicted 4.5% Compound Annual Growth Rate (CAGR) over the course of the next five years (Mordor Intelligence, pg. 1) means producing more beans to keep them affordable and profitable. Despite the high demand, farmers are producing less amounts of bean harvest -reporting last year's harvest with approximately 300,000 less acres of beans. (Held, 2021) This is caused by a multitude of reasons: extreme weather conditions affecting the potential of harvestable crops, the commodity effect, the amount of consumed product per capita around the globe, and the effects of the pandemic of 2019-2021. (Held, 2021)

Classifying dry bean types holds many benefits, as the type of bean is an indicator of its quality and price. Different types of beans also have a range of care requirements, so being able to automatically classify them allows farmers to implement the appropriate agricultural practices and save any expenses associated with overwatering, tilling too much land, etc. This would be especially useful since it would allow farming businesses to navigate around extreme weather and conditions, effectively lowering the risk of a poor harvest. They could also determine which beans would be the most beneficial to grow overall.

Problem Statement

To that end, we will propose and implement a design for classifying these dry bean instances to accurately determine their respective species.

Dataset

The Dry Bean dataset, obtained from the UCI Machine Learning Repository, contains information about 7 different types of dry beans. Features such as the length, area, perimeter, and more were extracted from 13,000+ images of dry beans in order to make up this dataset.

Proposed Solution

We propose implementing classification by utilizing multiple classification algorithms (Kernel SVM, Linear SVM, Decision Trees, and Random Forest Classifiers) on the data. The performance of each model will be obtained and compared in order to determine which is the best for the problem.

Methodology

A Python program was written that uses the SciKit Learn methods for Linear and Kernel SVM classification, Decision Tree Classification, and Random Forest Classification. Each classifier was fitted and used to predict on a set of testing samples. After this point various metrics were collected in order to determine the performance of each classifier on the

dataset. The metrics collected are the following:

- Accuracy
- Precision
- Recall
- F1
- Runtime

Results

The following are the performance metrics for the classifiers being utilized in this experiment. These metrics are not given on a per-class basis, but rather the performance on the dataset as a whole.

```
-------
Accuracy (Linear SVM): 0.9262977473065622
Precision (Linear  SVM): 0.926463360210026
f1 (Linear SVM): 0.9262278151672764
Recall (Linear SVM): 0.9262977473065622
Runtime (Linear SVM): 0.623055253000075
```

Figure 1 : Performance Metrics for Linear SVM

```
Accuracy (Kernel SVM): 0.9076885406464251
Precision (Kernel SVM): 0.9086077227107745
f1 (Kernel SVM): 0.907362173724844
Recall (Kernel SVM): 0.9076885406464251
Runtime (Kernel SVM): 0.623055253000075
-------
```

Figure 2 : Performance Metrics for Kernel SVM

```
-------
Accuracy of Decision Tree Classifier (Score Method): 0.83
Accuracy (Decision Tree): 0.82615083251714
Precision (Decision Tree): 0.9204211100800144
f1 (Decision Tree): 0.8581819774379963
Recall (Decision Tree): 0.82615083251714
Runtime (Decision Tree): 0.623055253000075
```

Figure 3 : Performance Metrics for Decision Tree

```
Accuracy for testing data{Random Forest Classifier}: 0.9272771792360431
Precision score for testing data {Random Forest classifier}: 0.9272771792360431
Recall score for testing data{Random Forest classifier}: 0.9272771792360431
Precision score for testing data{Random Forest classifier}: 0.9272771792360431
Runtime (Random Forest Classifier - testing): 0.623055253000075
```

Figure 4 : Performance Metrics for RandomForestClassifier

In order to more thoroughly assess the performance of each classifier on this particular dataset, per-class metrics were also obtained in order to ascertain to which degree of precision, accuracy, etc. each classifier could identify the labels of each instance.

These metrics are visualized in Figures 1-3 of the appendix.

Discussion

According to the metrics recorded in the previous section and the bar graphs provided in the appendix, it is possible to determine which classifier works best for the Dry Bean Dataset, and thus for solving our problem.

Note on the metrics in Figures 1-4 in the results section. In terms of precision, the Random Forest classifier has the best results, with Linear SVM, Kernel SVM, and Decision Tree following afterwards. The same can be said about the F1 and recall scores. In addition, Decision Tree holds strongest in accuracy for each of the available classes. It is, however, worth noting that the differences in the different metrics are fairly minimal and that when we speak of the differences between the performances of these classifiers that we are speaking in differences to the order of 0.1, 0.2, 0.3, etc.. The reports for the metrics for each dry bean species have also been graphed in the Appendix, Figures 1-3. At first glance, it would seem that the Random Forest classifier is best suited for the dataset, and that the Decision Tree classifier has the worst suitability.

However, the same cannot be said when comparing the results of the bar graphs from the Appendix, Figure 4. These results, as you may view in the project code, are based on the overall performance of the classifiers instead of for the individual classes measured. These metrics provide a base, or foundation, for what to expect for the overall performance after training and testing the dataset, regardless of the class being predicted. In this case, the accuracy and runtime bar graphs both present the Decision Tree classifier as the most effective and efficient approach to differentiating dry bean species. The overall accuracy for the Decision Tree classifier is approximately 1.0, with Random Forest as the second best and Kernel SVM being the least accurate. As for the overall runtime performances, Decision Tree is the most efficient by a landslide, with the Random Forest Tree having the least efficient runtime.

We see that while an ensemble method such as RandomForestClassifier can help increase per-class metrics for precision, recall, and f1, it suffers from lessened runtime performance due to the fact that it is running multiple base classifiers, each of which take additional time to train.

Conclusion

Manually deciding which type of dry bean farmers plan to grow and profit from has proven difficult and extensive, given the growing demand and harsh conditions bringing poor quality harvests.

Classification of beans based on a variety of features assures its quality, which allows agricultural businesses to gain more profit by planting the most beneficial bean for the current conditions they may be planted in.

We have implemented four classification models ‘Linear SVM’, ‘Kernel SVM’ , ‘Decision Tree Classifier’ and ‘Random Forest Classifier’ to determine which would ensure accurate assignments in the smallest amount of time. The performance scores(F1 score, Precision, Recall) for each class label are all good for the classification models implemented. All four classification models result in proficient accuracy scores and precision, but varying runtime scores. Decision Tree will be taken as the best classifier of four models, as it will classify the beans with a high accuracy score with less running time in comparison to the other three classifiers.

Citations

Held, L. (2021, November 15). *Comments on: Beans may be the 'food of the future,' but U ...* Beans May Be the 'Food of the Future,' but U.S. Farmers Aren't Planting Enough. Retrieved April 21, 2022, from https://civileats.com/2021/11/15/beans-may-be-the-food-of-the-future-but-u-s-farmers-arent-planting-enough/feed/

Mordor Intelligence. (n.d.). *Dry beans market: 2022 - 27: Industry share, size, growth - mordor intelligence*. Dry Beans Market | 2022 - 27 | Industry Share, Size, Growth - Mordor Intelligence. Retrieved April 21, 2022, from https://www.mordorintelligence.com/industry-reports/dry-beans-market#:~:text=The%20global%20dry%20beans%20market,of%20their%20long%20shelf%2Dlife.

Appendix

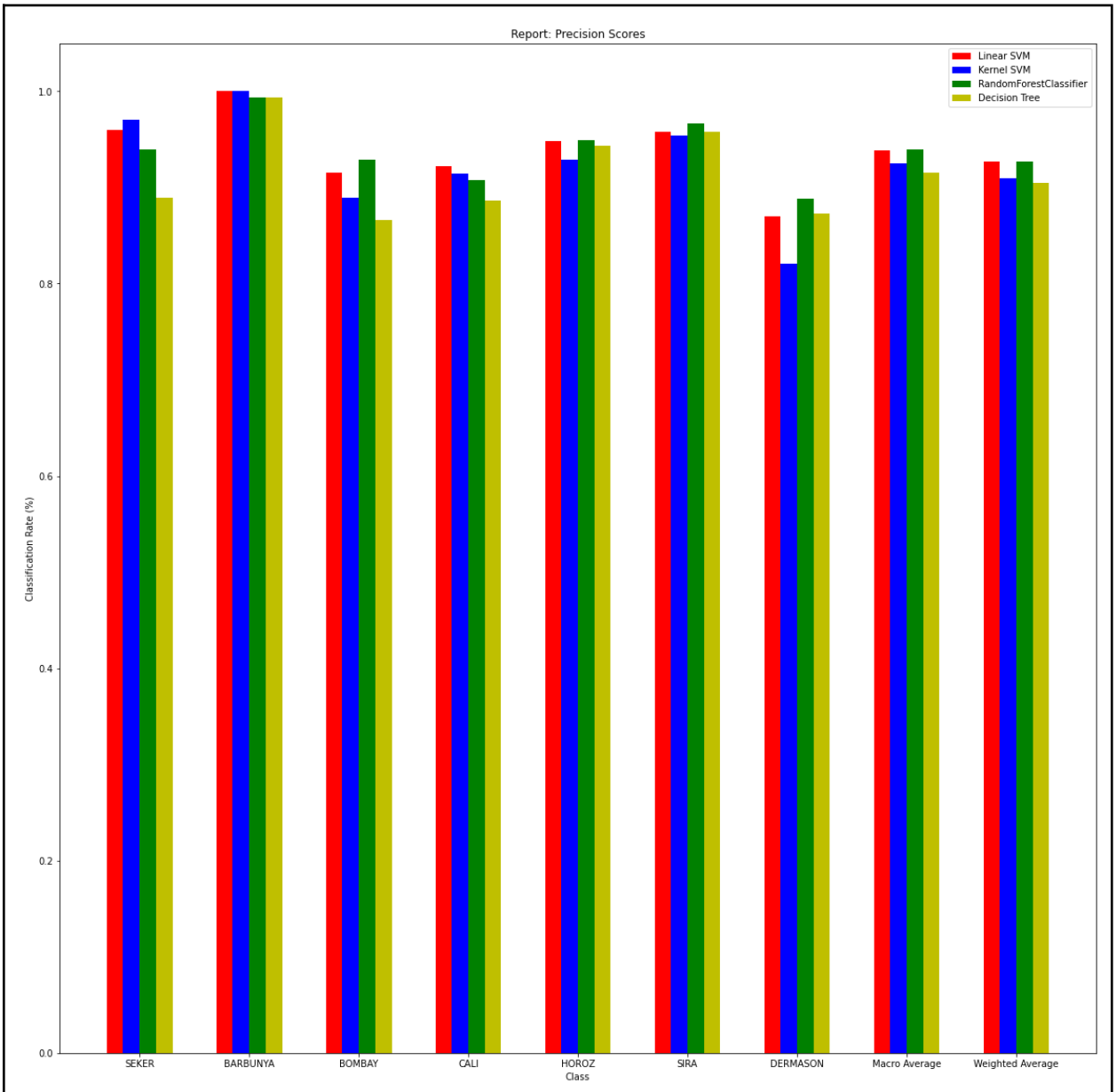

Figure 1 : Per-class precision scores for Linear SVM, Kernel SVM, RandomForestClassifier, and DecisionTreeClassifier

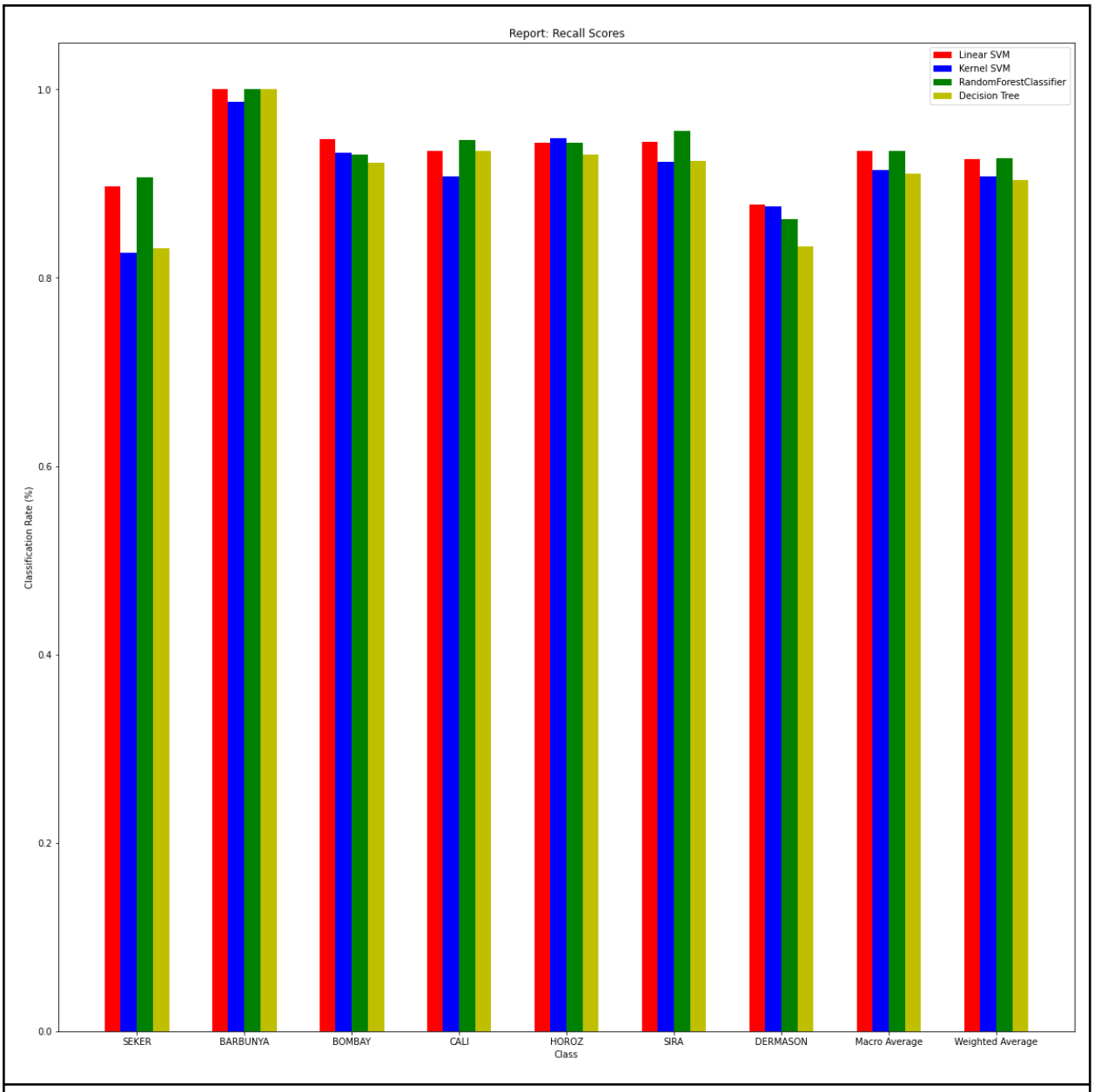

Figure 2 : Per-class recall scores for Linear SVM, Kernel SVM, RandomForestClassifier, and DecisionTreeClassifier

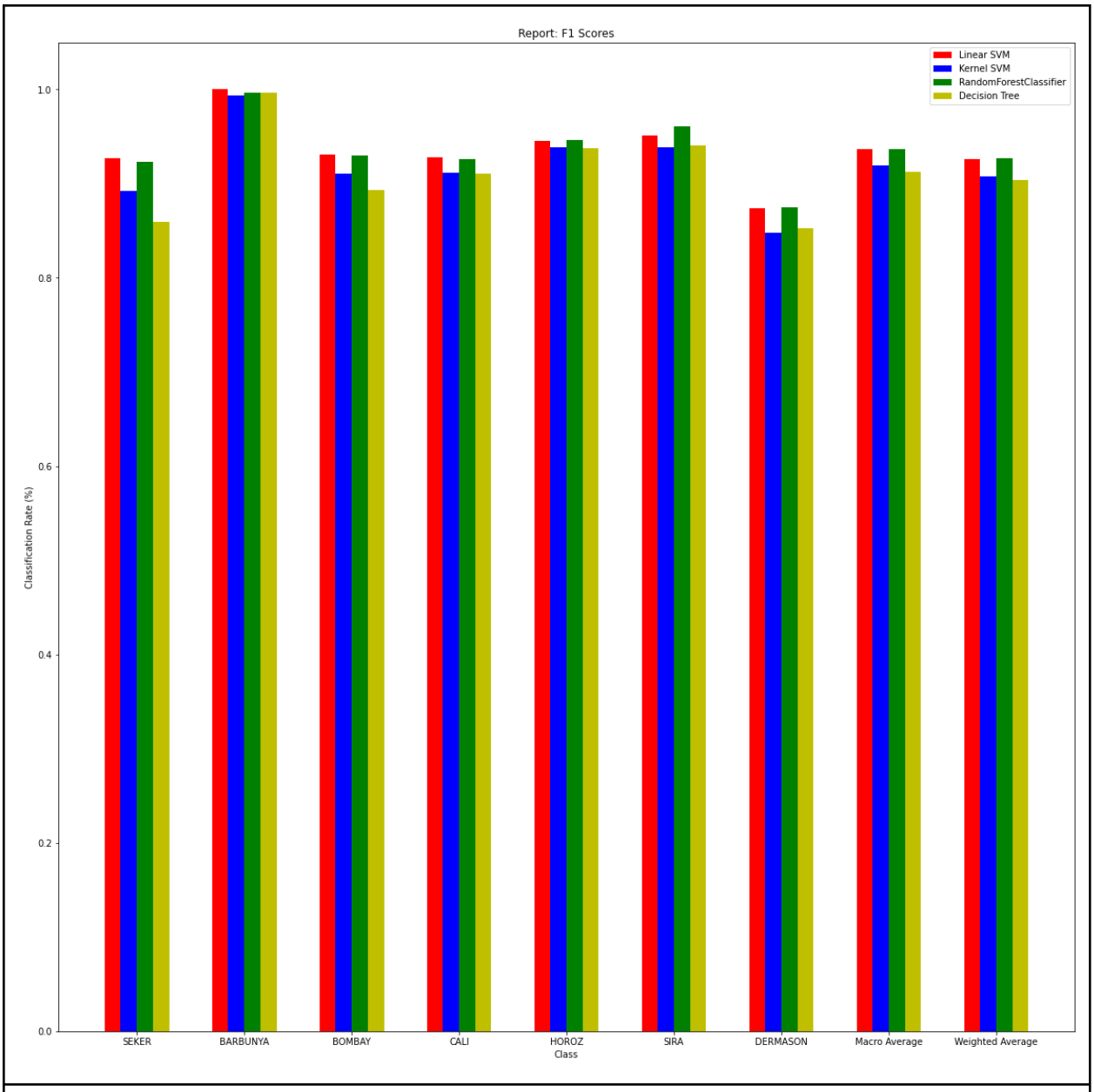

Figure 3 : Per-class f1 scores for Linear SVM, Kernel SVM, RandomForestClassifier, and DecisionTreeClassifier

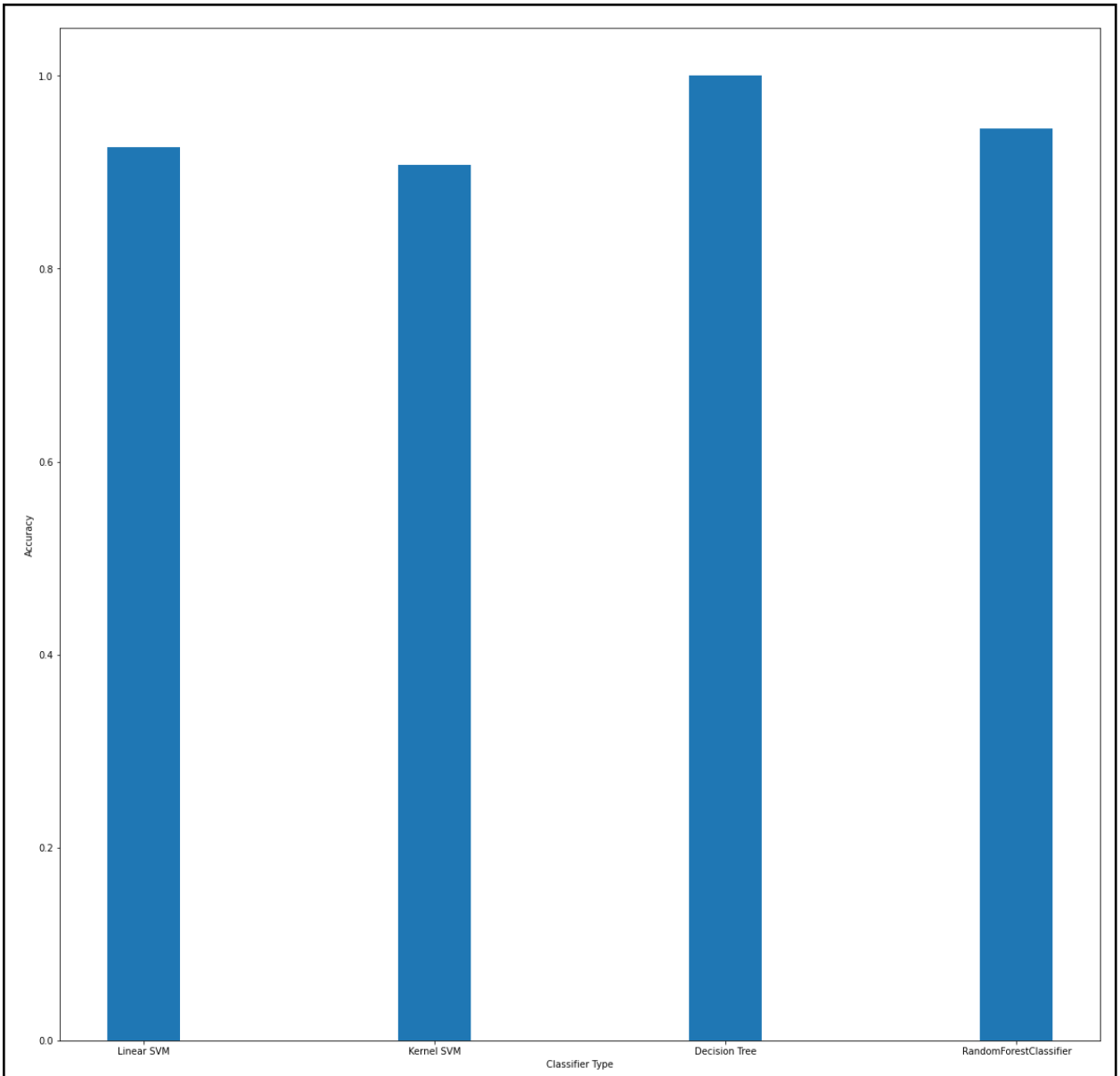

Figure 4 : Accuracy scores for Linear SVM, Kernel SVM, RandomForestClassifier, and DecisionTreeClassifier

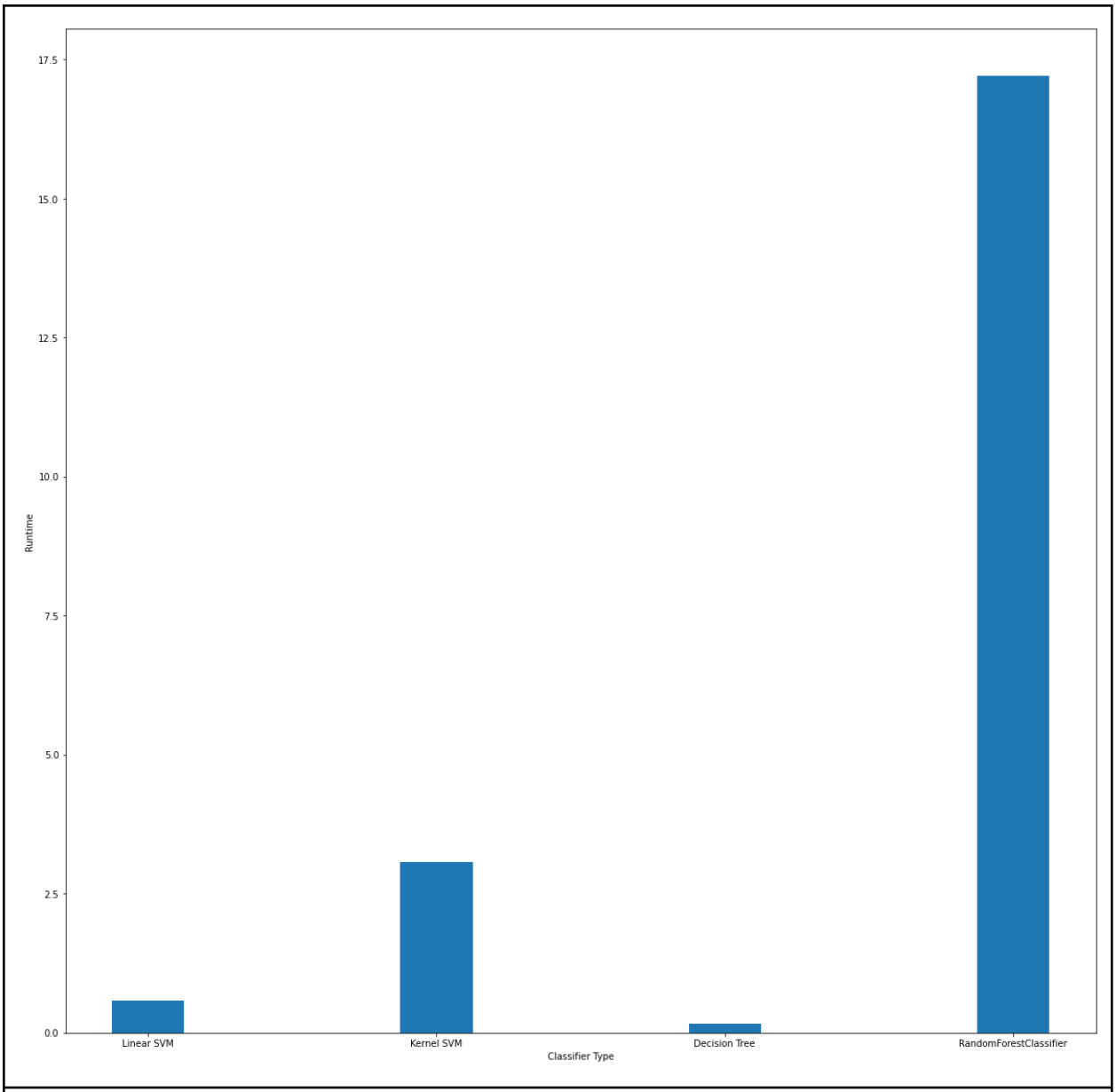

Figure 4 : Runtimes for Linear SVM, Kernel SVM, RandomForestClassifier, and DecisionTreeClassifier