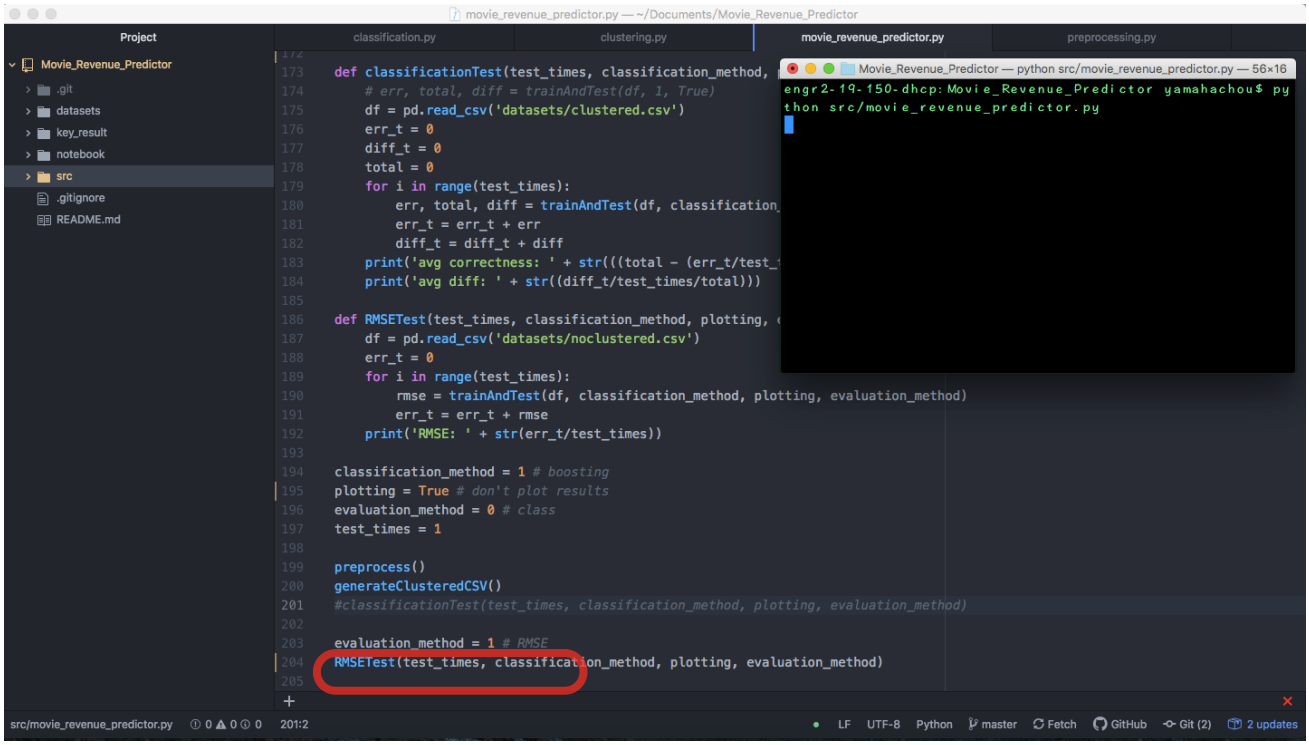
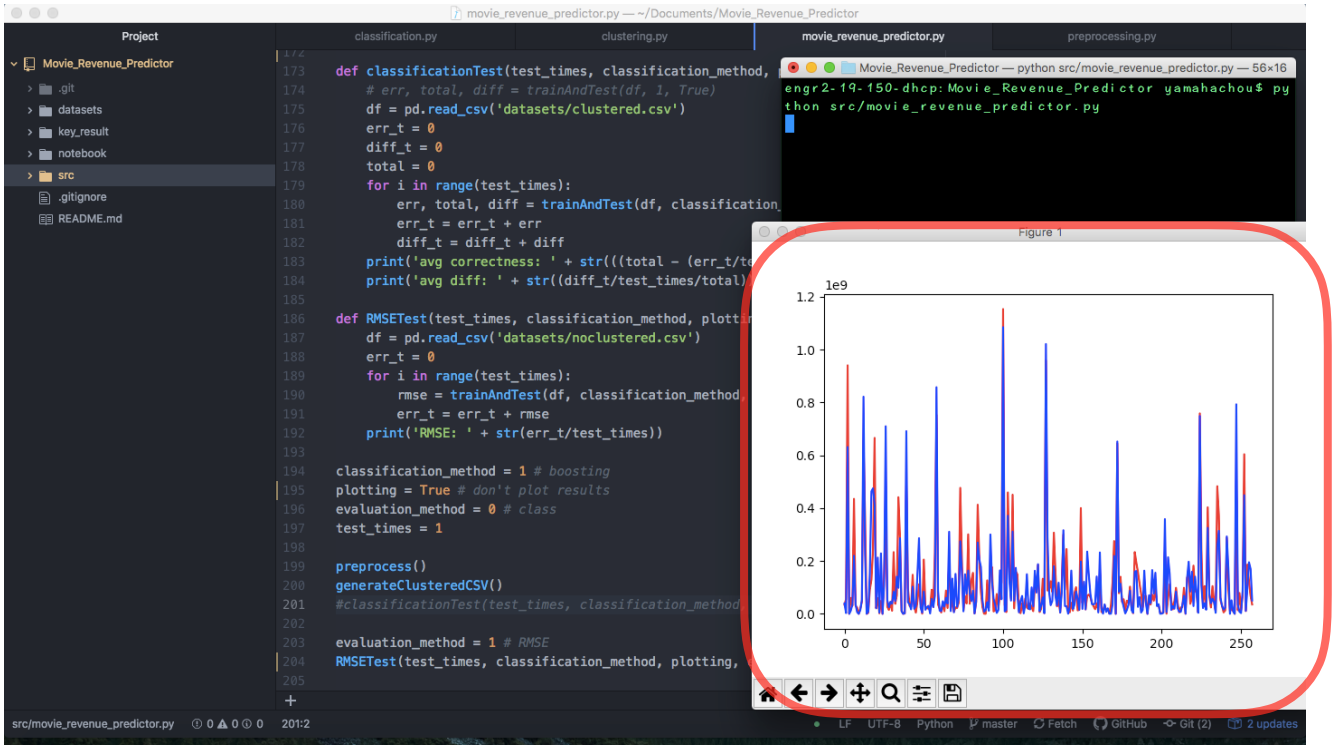
Result

I. Process of running code

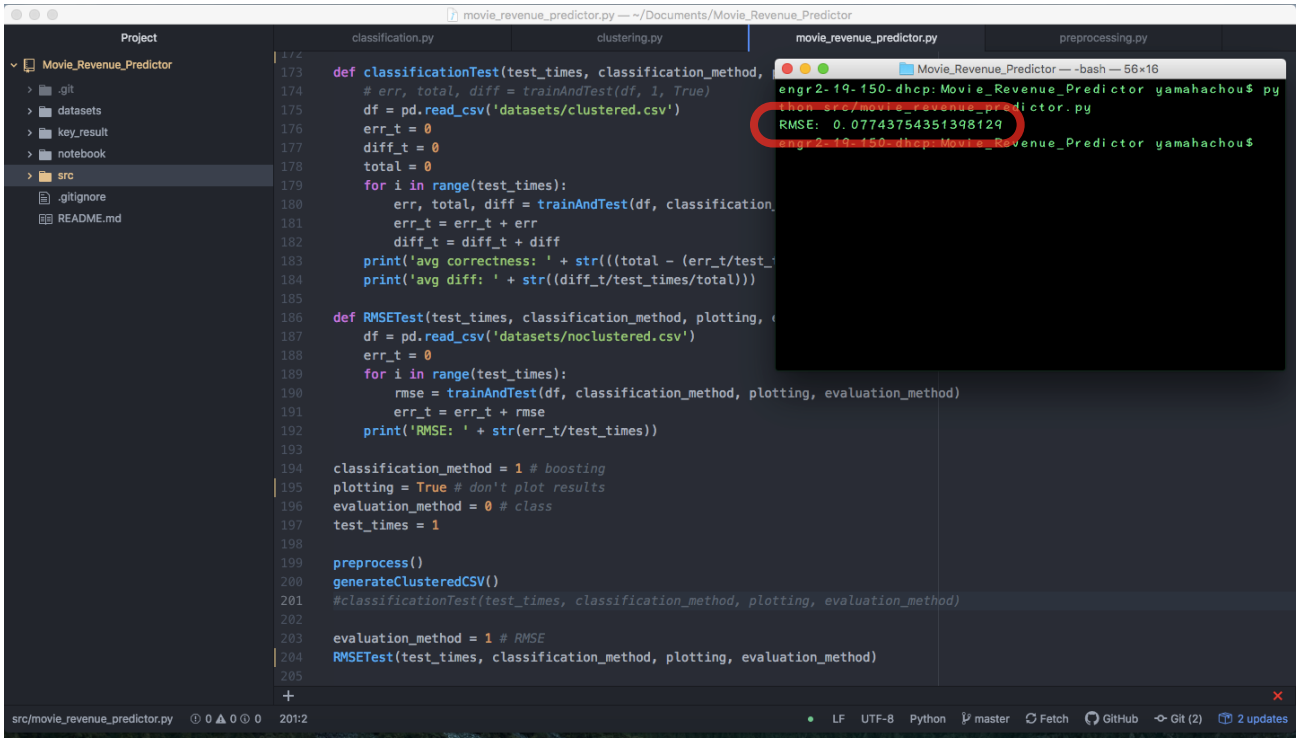
1. Run RMSE test



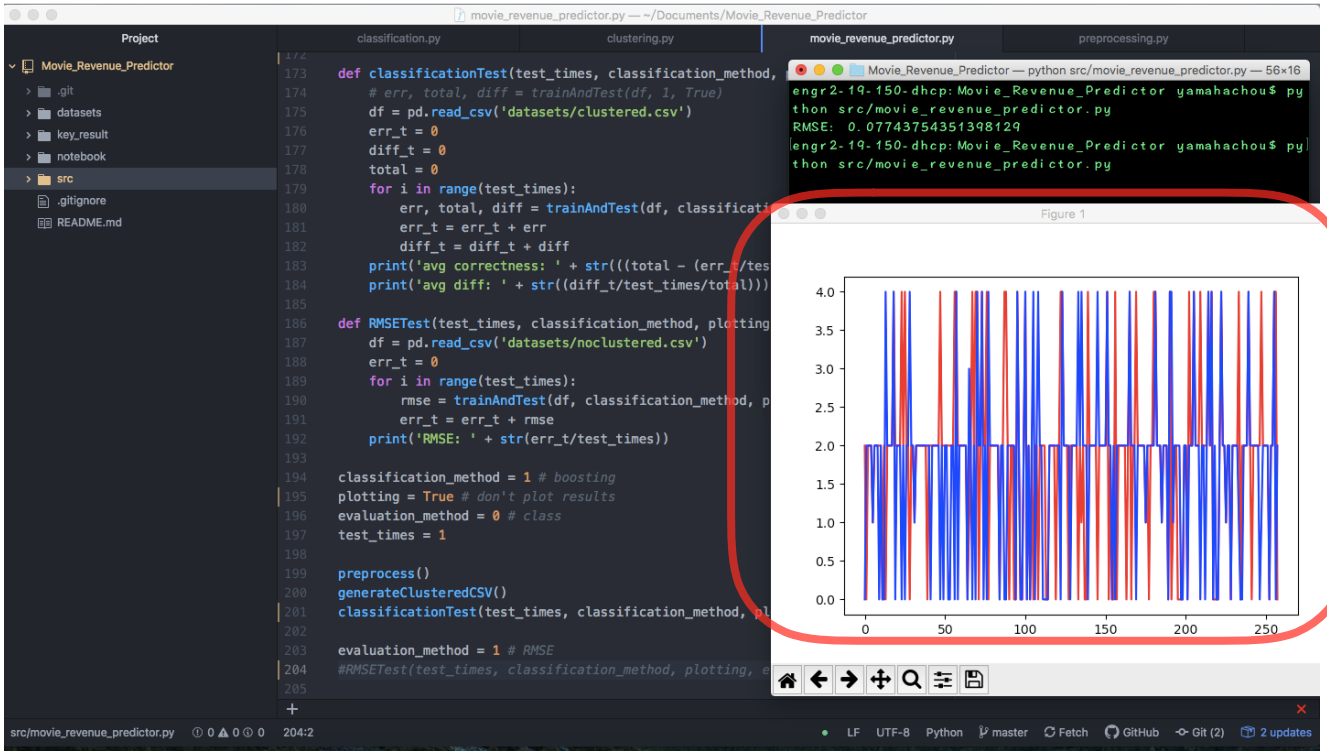
2. After it complete the calculation, it would show the plotting picture



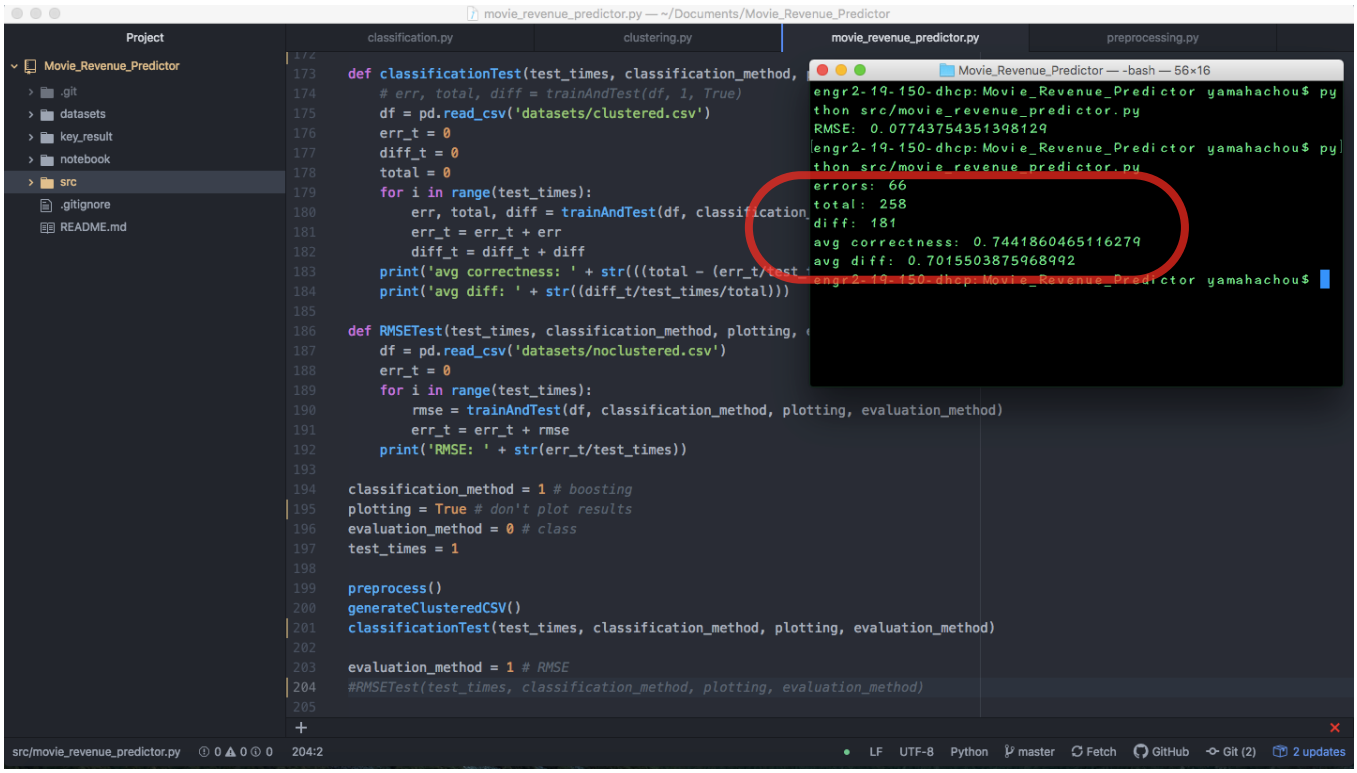
3. After close the plotting picture, the terminal would show the computation result of RMSE



4. Start to run the classification error evaluation, it would show the plotting picture



5. After close the plotting picture, the terminal would show the computation result of Classification error method

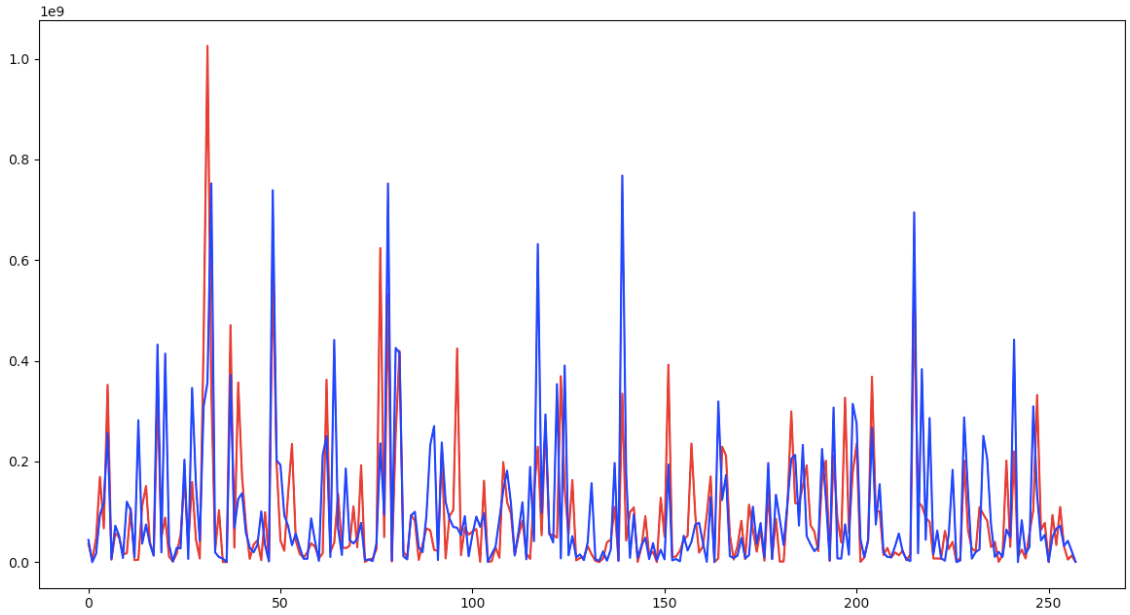


II. Evaluation result for each classification

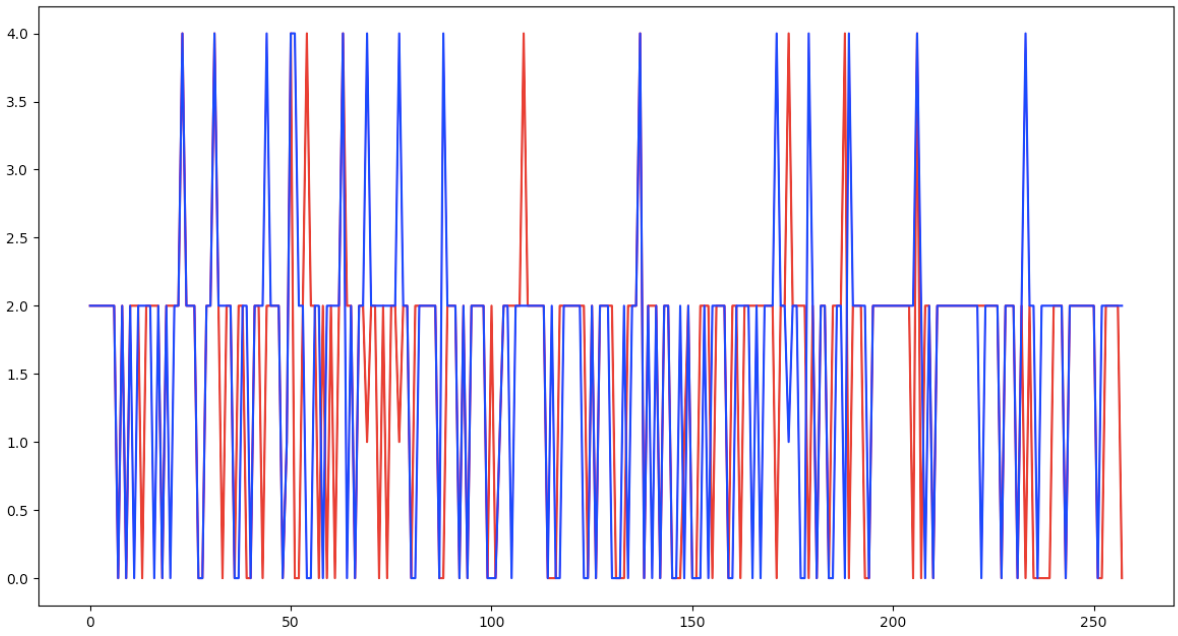
After we ran the evaluation method on each classifier, we would get the plotting picture for each classifier, the red part means **our prediction**, and blue part means the **real situation**.

1. Decision Tree Classification

* RMSE

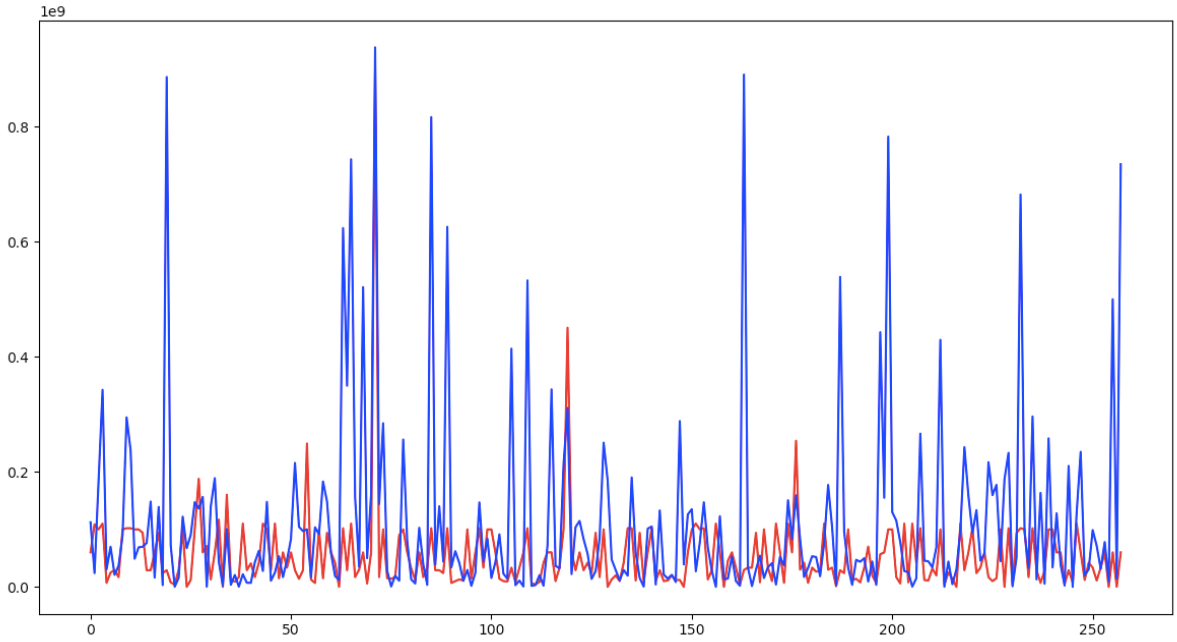


* Classification error

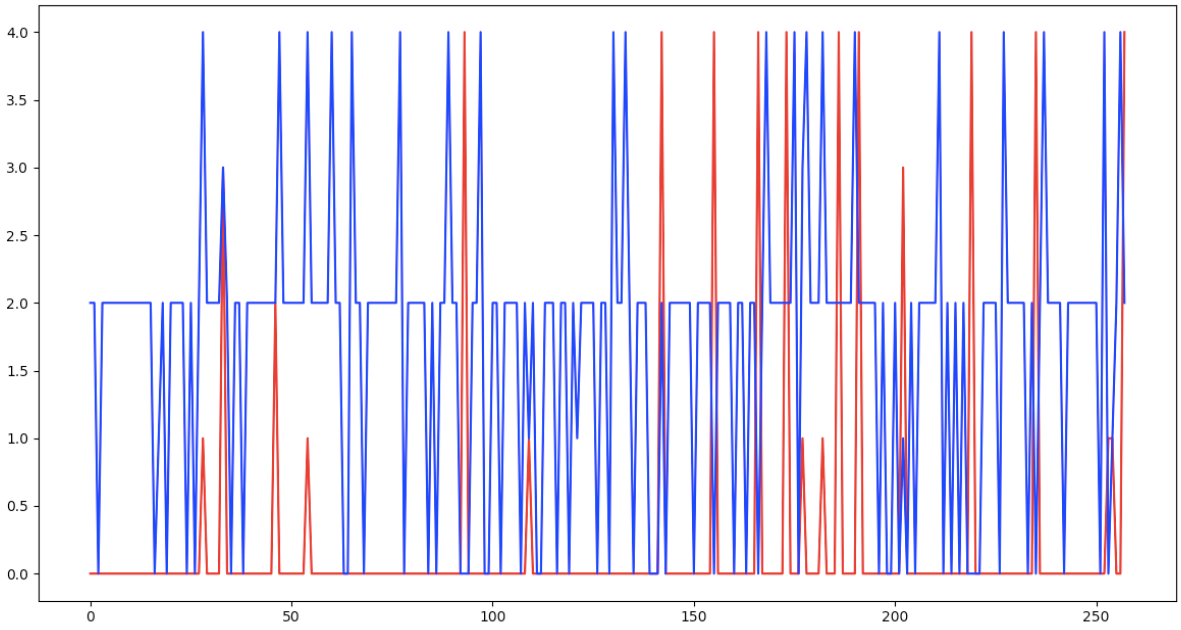


2. Gaussian Naive Bayes Classification

* RMSE

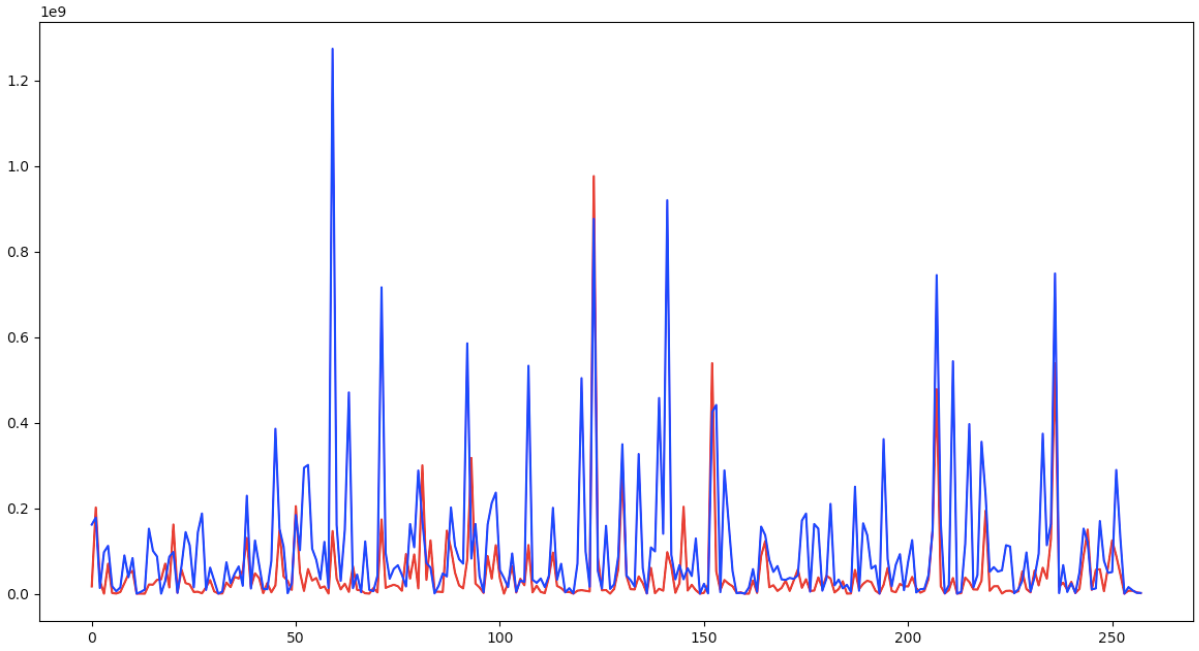


* Classification error

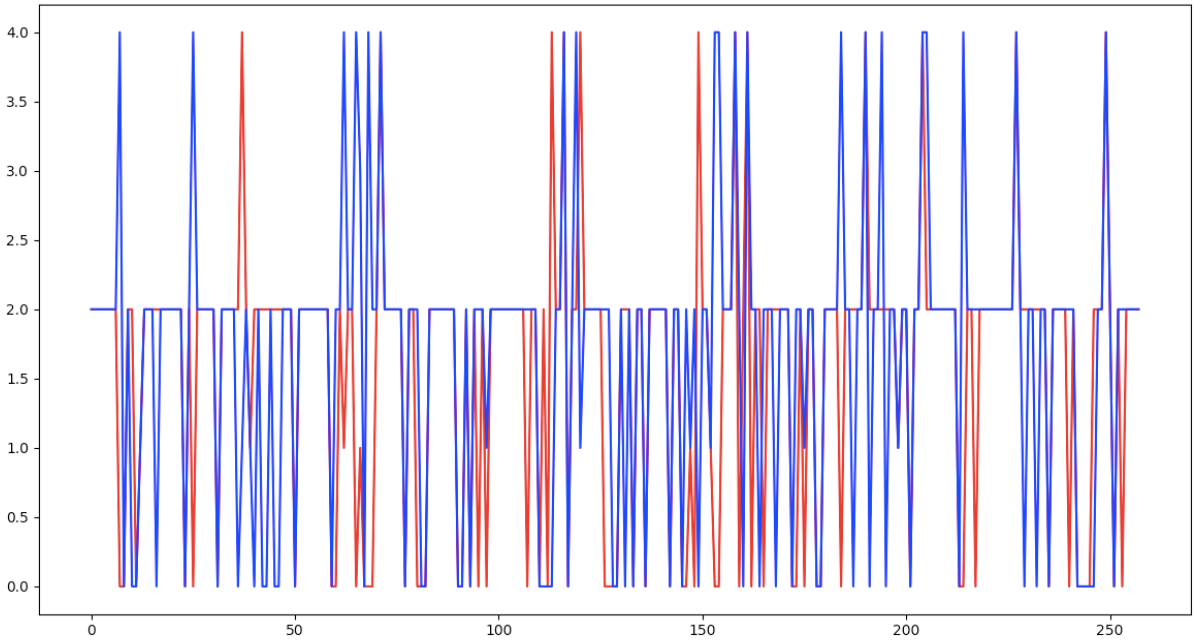


3. Random Forest Classification

* RMSE

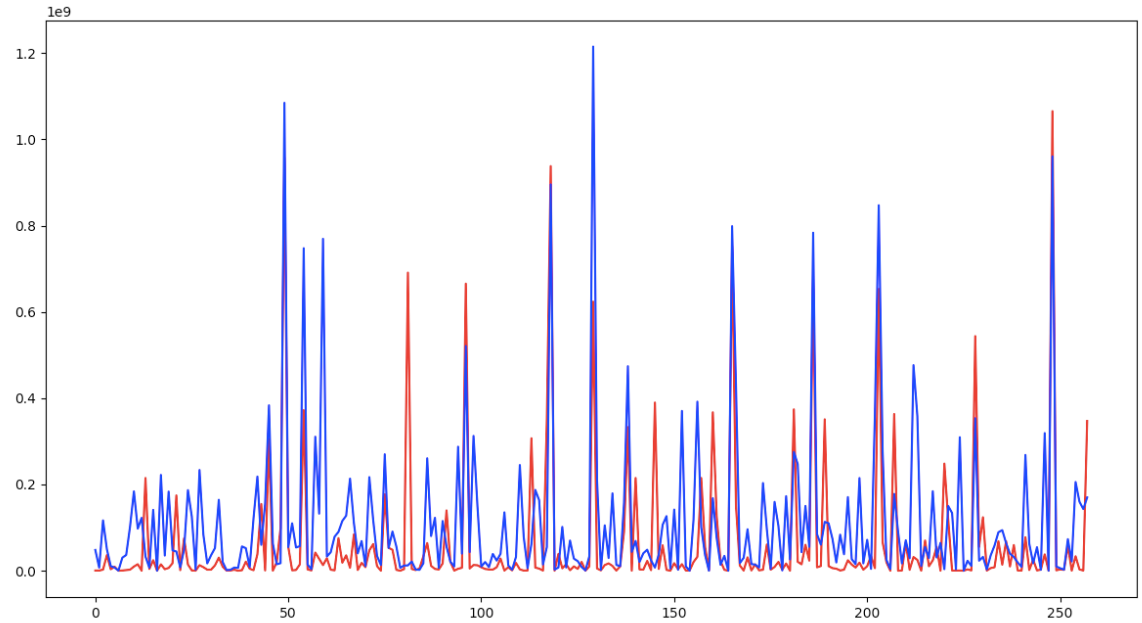


* Classification error



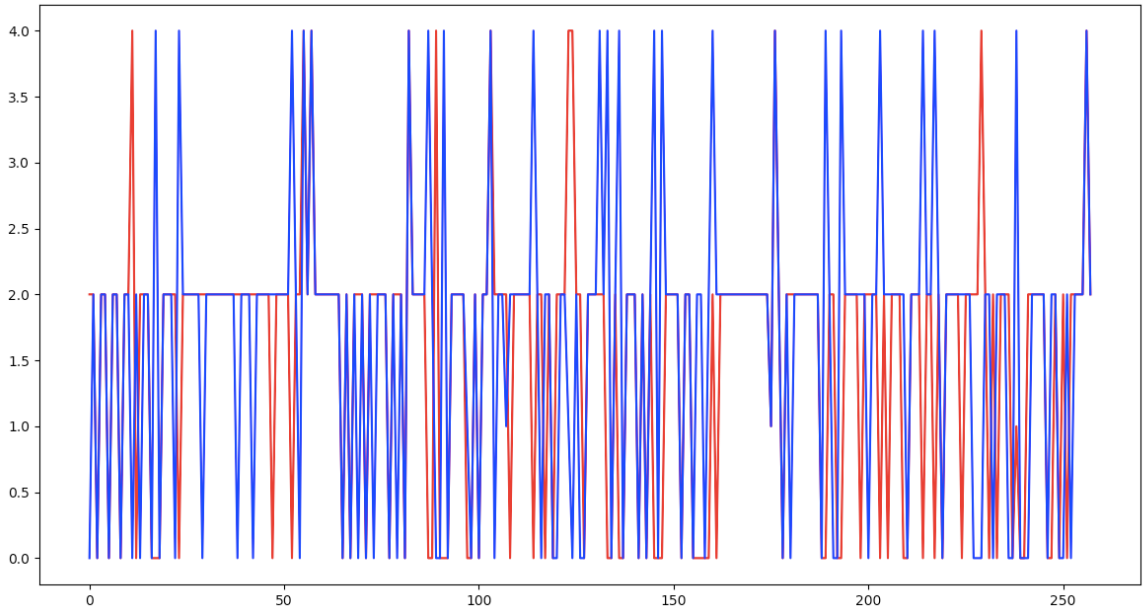
4. Bagging (with Decision Tree as base estimator)

* RMSE



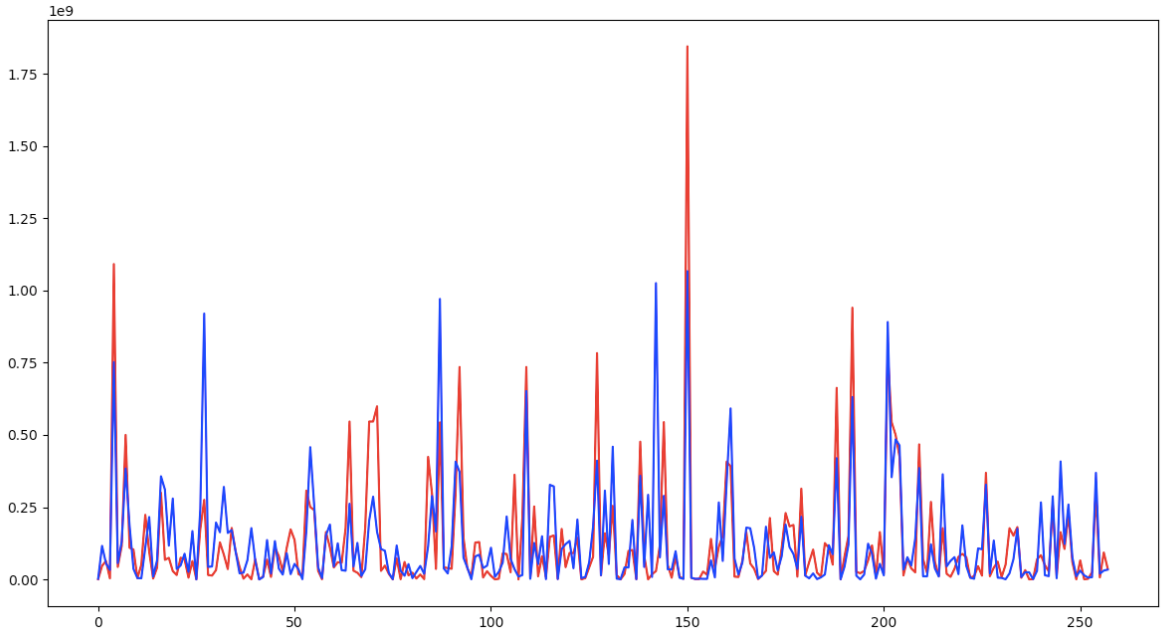
*

* Classification error

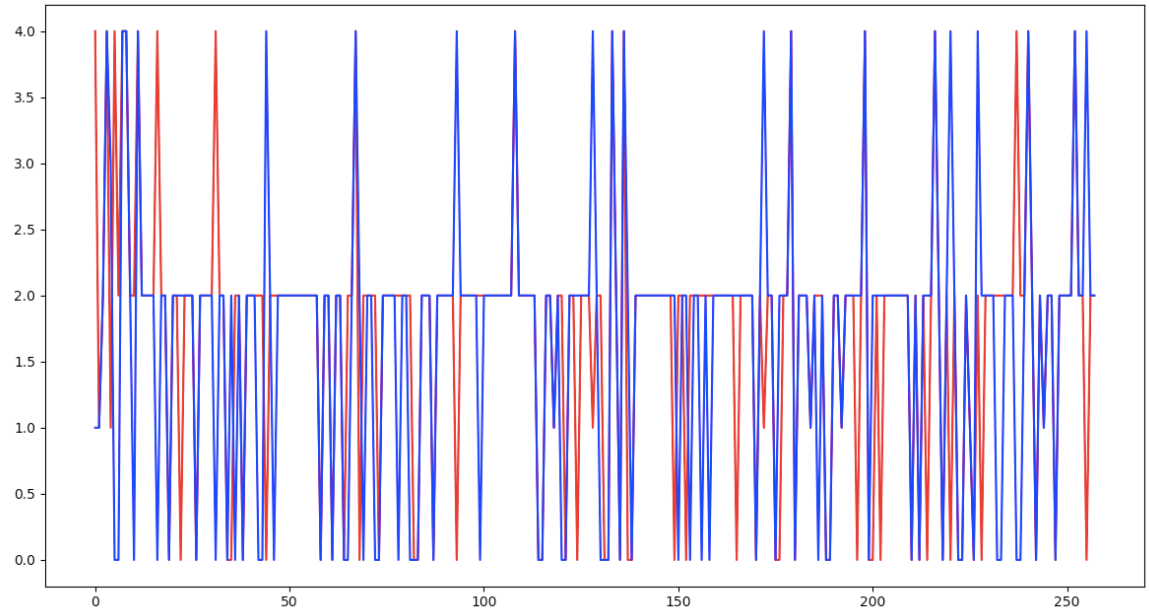


Bagging (with Random Forest as base estimator)

* RMSE

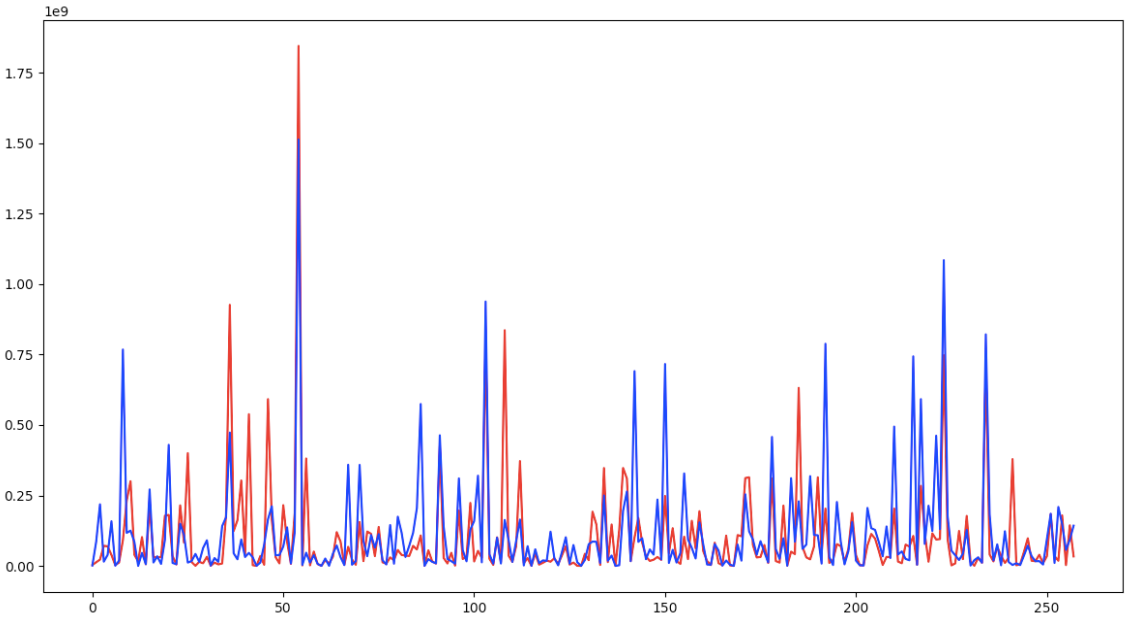


* Classification error

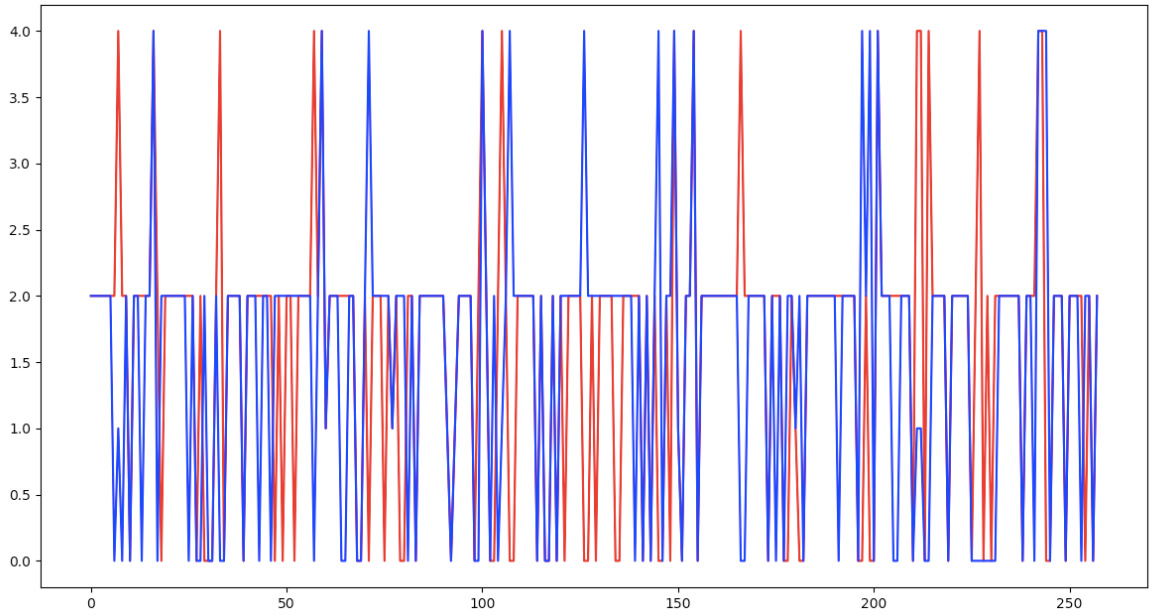


Boosting (with Decision Tree as base estimator)

* RMSE

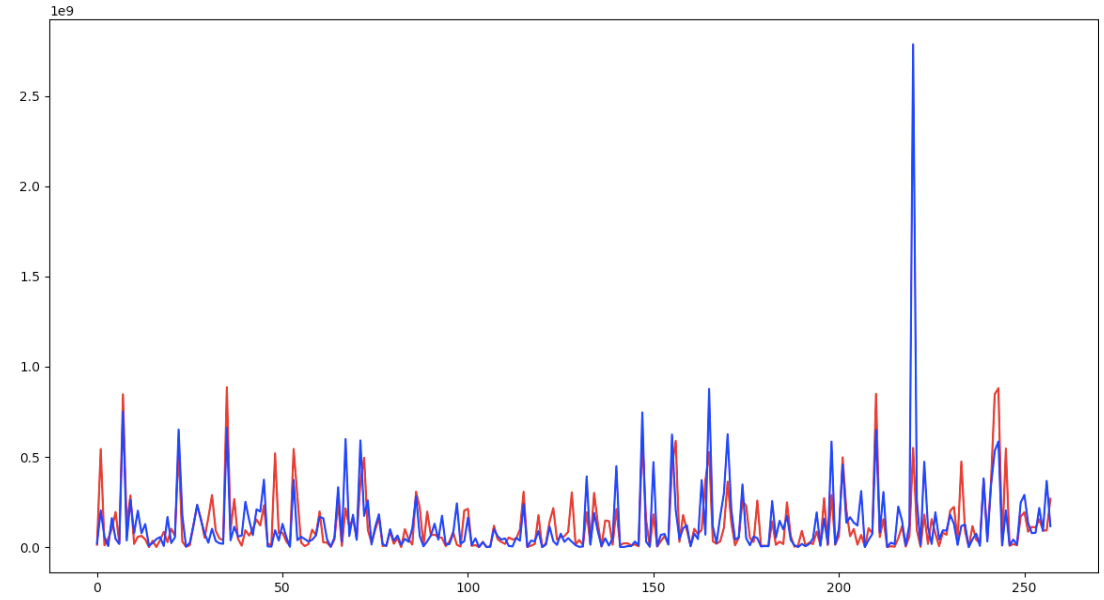


* Classification error

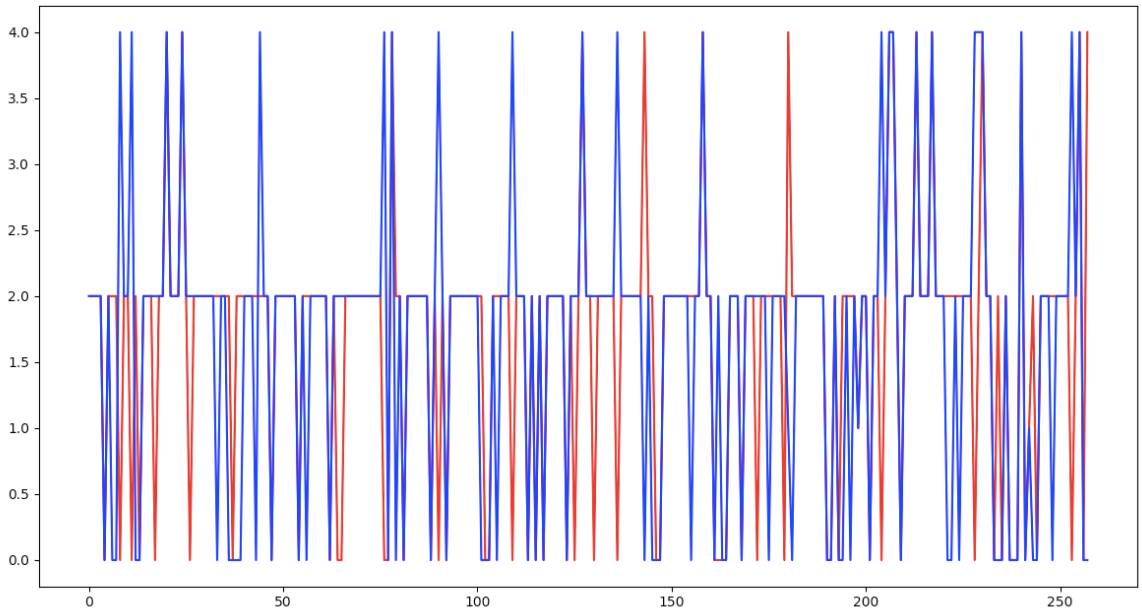


Boosting (with Random Forest as base estimator)

* RMSE



* Classification error



III. Comparison of the evaluation result between each classifier

In this evaluation, we would split our original dataset into a training set and testing set. After we trained our training set we implemented two ways to test our result in testing set. First one is RMSE, it is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modeled. RMSE serves to aggregate them into a single measure of predictive power. We also normalized the result of RMSE since we wanted to facilitate the comparison between different datasets, and we used the range (defined as the maximum value minus the minimum value) of the measured data.

Second evaluation method is classification error method. In the beginning, we used K-Means clustering to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Then we trained training set and tested the result in testing set to see whether it can fall in the right partition, if it fall into the right section, we calculated how far it is from the right section, and divided total data number to see the diff rate. Below is a comparison table which record the two result from each classifier algorithm, the red part is the best result.

	Decision Tree	Gaussian Naive Bayes	Random Forest	Bagging		AdaBoosting		w/o Hashtags
base estimator				Decision Tree	Random Tree	Decision Tree	Random Tree	Random Tree
RMSE	0.131	0.175	0.087	0.113	0.0868	0.098	0.0706	0.0764
Classification								
avg correctness	0.75	0.2	0.79	0.74	0.81	0.79	0.82	0.75
avg diff	0.55	1.75	0.51	0.65	0.45	0.51	0.45	0.55