# Exercise 5

Authors: Lavi.Lazarovitz (065957383) & Aharon Sharim (052328523)

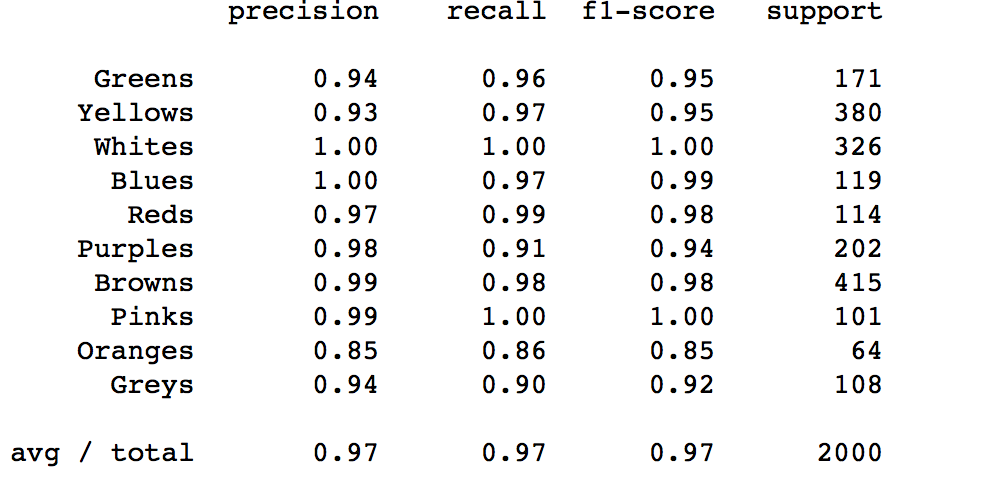
Our final submission includes the following elements:

* Votepedict.py – the module that loads the data, train it on the labeled datasets and predict the new dataset.
* Coalitions.py – the module that analyzes the voters’ clusters.
* EX\_5 & coalitions python notebooks
* Vote\_predicton.csv – the predicted voted of the new data (as specified in the assignment)

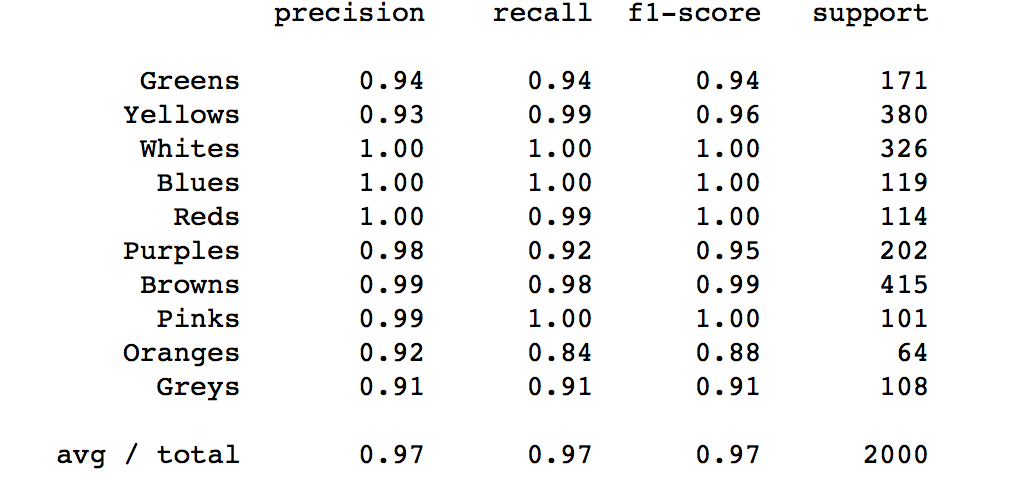
# Vote Prediction

In our previous exercises we used OneVOne classifier as it demonstrated the best results. However, after playing around a bit more with the data and after learning about the Ensemble methodology, we tried Random Forest classifier and the AdaBoost. The results we got after training and cross validating the data were great:

Using Random Forest we got:



And using AdaBoost we got:



Those results are better by the best results we got previously. Therefore, we decided to go with the Random Forest methodology which is just clearer to us.

After this quick classifier adjustment, we ran our script for data cleansing we built for previous exercises to prepare the new data (outliers, columns filter and more)

Now that we have the data ready, we ran the predict function on the features (after removing the ID column) and saved it together with the ID to a csv.

**The prediction showed that the BROWN party has won the most votes with 2255 of the votes which is 22.55%.**

The percentage break down according to our prediction is as follows (can be seen in the EX\_5 notebook as well):

Reds: 3.710000

Greens: 10.200000

Whites: 3.120000

Yellows: 5.260000

Greys: 5.580000

Oranges: 7.900000

**Browns: 22.550000**

Blues: 8.310000

Pinks: 12.940000

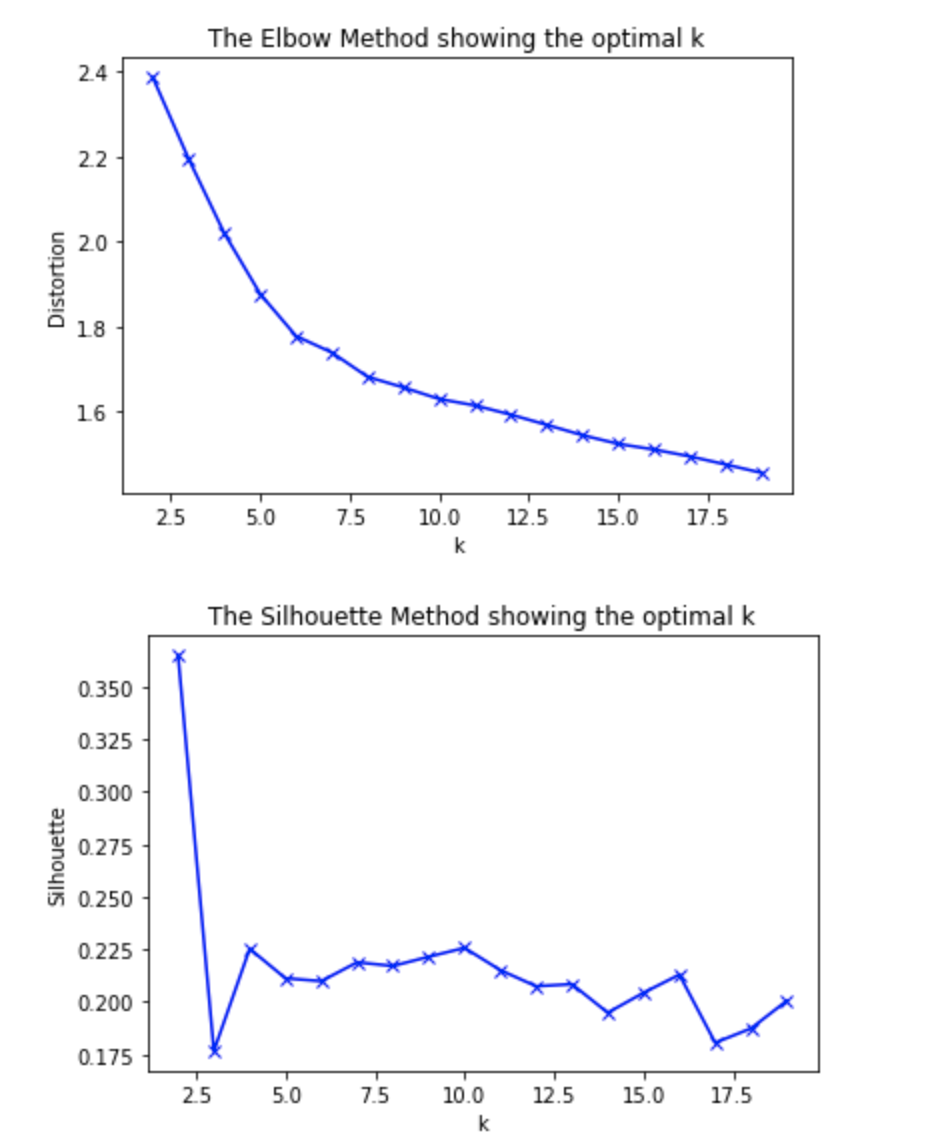
Purples: 20.430000

# Coalitions:

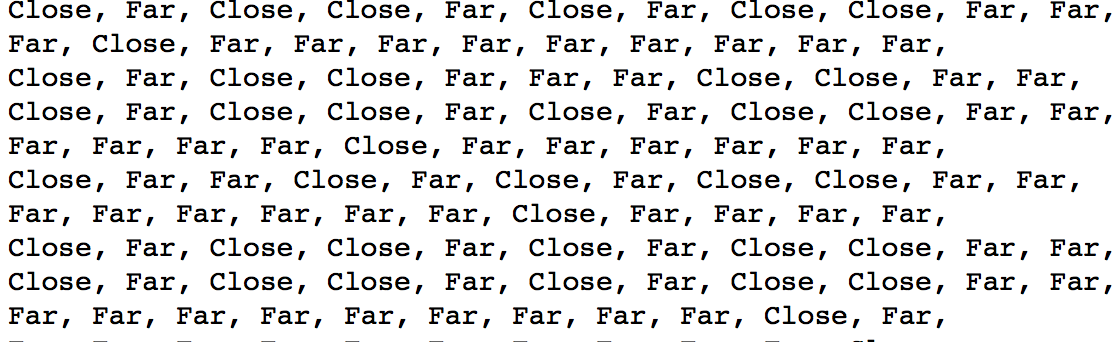
In our previous analysis we have found that the White-Brown-Purple-Pink (WBPK) coalition leads in the number of votes and the correlation between those parties’ voters. Even though the new data positioned the Browns on top, the coalition WBPK seems to remain the strongest based on the analysis we ran in the coalitions analysis.

Because the analysis is very much similar to the one we ran in the previous exercise, the following review the brief.

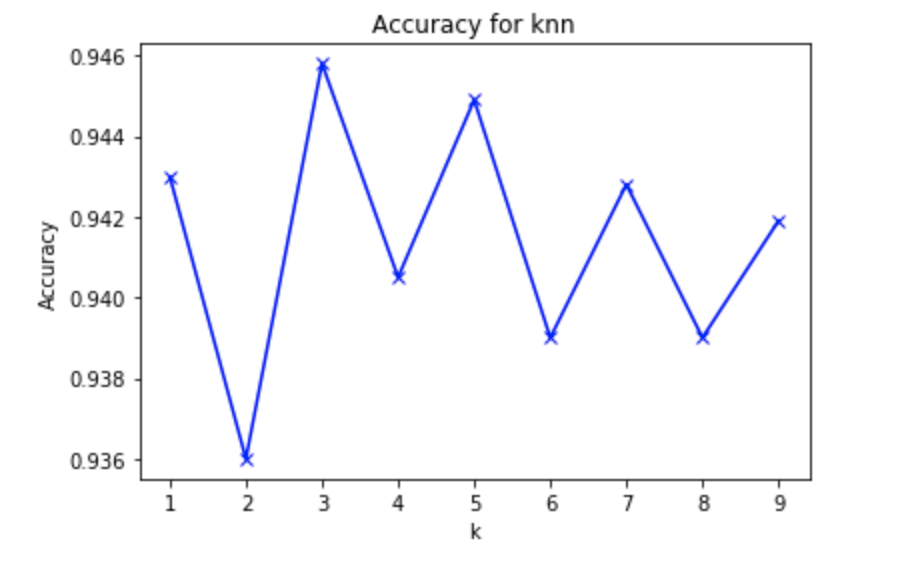
At first, we ran KNN and Kmeans to learn the voters distributions across the different clusters. As in the previous exercise we chose k to be 11 for the Kmeans based on:



We chose k to be 5 for the KNN based on:



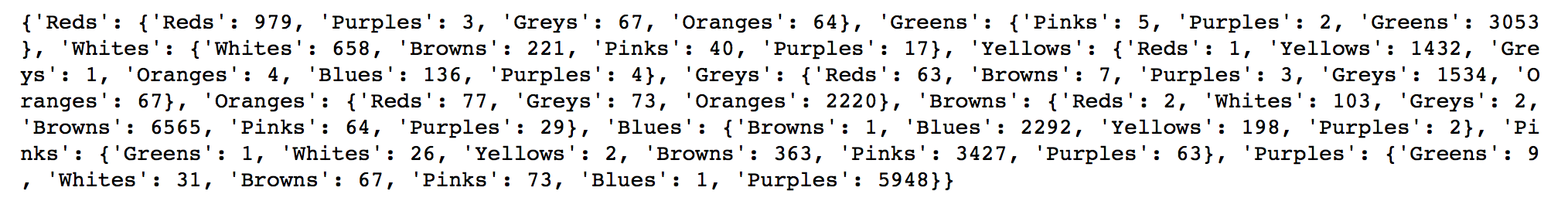
Shows that 5 neighbors separate the clusters pretty good (the 5th column shows that the centers are pretty fat from each other). And k = 5 also provide good accuracy:



Both algorithms showed 3 major coalitions:

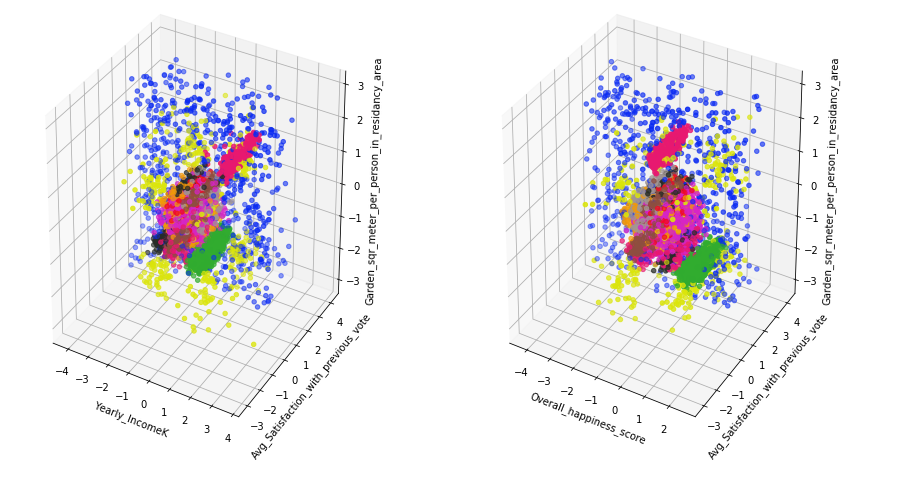
* **White-Brown-Purple-Pink**
* Red-Greys-Oranges
* Blue-Yellows

The data that supports this is printed when the collation.py is executed or via the collation notebook:



Out of those 3 coalitions only one, WBPK, is a valid coalition that have 58% of the voters. The 2 other coalitions have less than 50% and other parties have relatively distinct features that do not allow the coalition to extend.

So… As can be seen from our previous visual modeling:



Together with the fact that the WBPK coalition seem to cluster together and naturally bond more than 58% of the voted, make WBPK our natural selection.