

When I saw the goal of this project, my thoughts first went to the basics, which meant the usual “random forest” sadly after experimenting with it a bit, it was quite slow and not that good, in the tests I did I tried a depth of 5 at first, then 50, then 500 and the results took an exponential time to appear, making the model unsuited for my needs.

Naturally the next step for me was to use some of the other models I keep hearing about, so I then used linear regression, and was quite saddened by the result I got, which was low compared to the other methods I tried and afterward I tried with K nearest neighbor, but no matter how I switched the number of points taken into account, the result was good, but still a bit under the 80% bar.

So, I then turned to google to try to see an algorithm that could work, and found one called Support Vector Machine.

To be perfectly honest, I barely understand the theory behind the algorithm, only that it was considered robust and good at dealing with large amounts of unclassified data (according to Wikipedia and the Sklearn site).

And this algorithm worked on the first try, obtaining the 86% accuracy which was asked of me.