Data Mining Kaggle Competition

**Team name**: Cristobal Zamorano Astudillo

**Team members**: Cristobal Zamorano Astudillo

**The highest private score**: 0.835016835016835

**The highest public score**: 0.77990

**Please describe how you improved the accuracy of your model step by step and what the accuracy was after each optimization**:

1. At the beginning I tried using a Naive Bayes Classifier. By far that was the worst score. Then I starting reading the documentation of `from sklearn.model\_selection cross\_validate`about the method. So I used a way to benchmark each model easily to visualize. I came up building Dtaframes with metrics using recall, precision, accuracy, and F1.

I know the competition was looking for accuracy only but I thought that I wanted to see how the models behave in other metrics. Also the Sklearn method of cross\_validate makes it easy to just copying and pasting the code of one and just changing the parameter of scoring. In addition, I used a VotingClassifier object to help pick the best model.

1. Then I went I ahead pick some the best previously mentioned models benchmark to search for their most optimal parameters by hand and Gridsearch. Probably this is the where I took the most time to work on.
2. The finally I tried using a BaggingClassifier but it wasn’t even good to myy previous gridserach models. So I ended up using a randomforest with the optimal parameters of gridsearchcv

**What sorts of people were more likely to survive?**

**It seems that Age, Cabin and Fare were the most important features in that order for my model. If we look at some of the pivot tables I made in my notebook, people that were rich or had a nice cabin, and younger had the highest priority to get out of the ship and use the escape boat. This makes sense if we remember the well-known movie Titanic where in fact rich kids are the ones who were prioritize the most.**