# Kaggle Python Tutorial on Machine Learning

Leonel Fernando Ardila, lfardilap@unal.edu.co
Juan Sebastián Flórez , jsflorezj@unal.edu.co

April 2, 2018

## 1 Chapter 1: Getting Started with Python

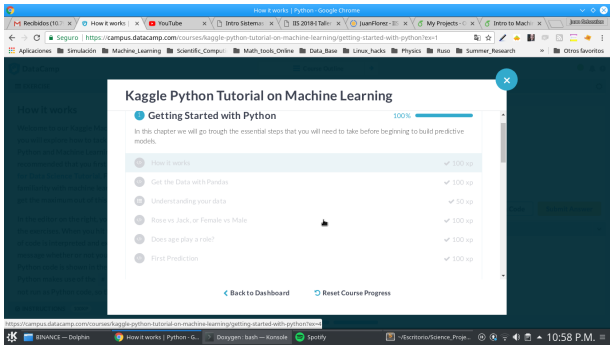


Figure 1: 100 % on the Chapter 1

## 2 Chapter 2: Predicting with Decision Trees

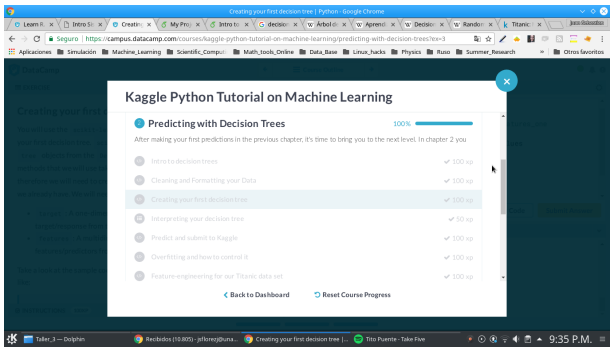


Figure 2: 100 % on the Chapter 2

We created the first prediction as was proposed in the tutorial and got a score of 0.71770,

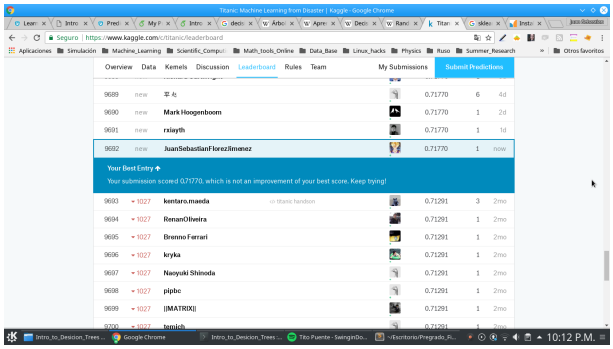


Figure 3: Score of the simple decision tree, which is better than a random classifier

The **max_depth** and **min_samples_spli** features were used to overcome the overfitting of the previous tree and a better score was obtained, in this case the score was 0.76076,

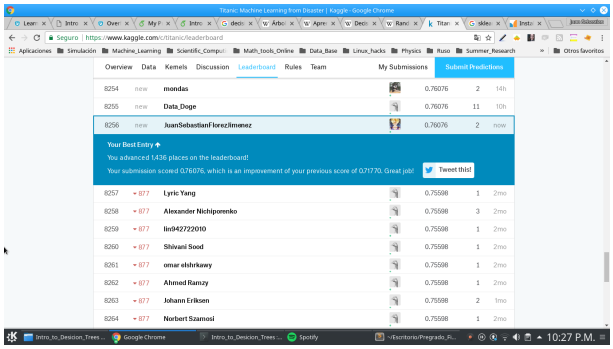


Figure 4: Score of the simple decision tree with the features **max_depth** and **min_samples_spli**

Although a better score for the training set is obtained when using the new variable called **family_size**, the test set gets a lower score i.e. 0.64114, hence, probably this feature makes the overfitting even stronger,

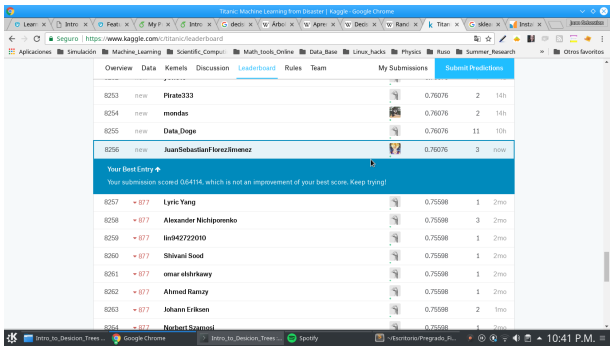


Figure 5: Score of the simple decision tree with the new variable **family_size**

# 3 Chapter 3: Improving your predictions through Random Forests

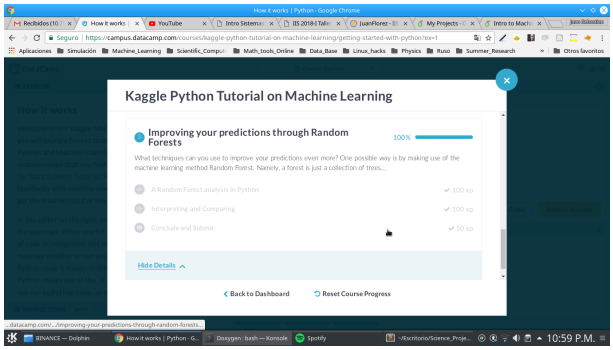


Figure 6: 100 % on the Chapter 3

A prediction almost as good as the one obtained with the modified decision tree is obtained when using the random forest with the parameters set as proposed in the tutorial, here the score of the prediction is 0.75119,

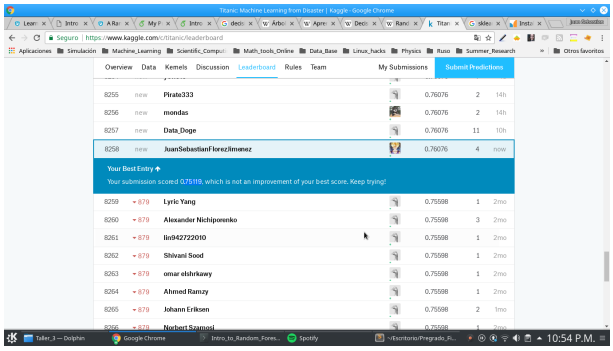


Figure 7: Score at Kaggle when using the random forest with the proposed parameters of the datacamp tutorial