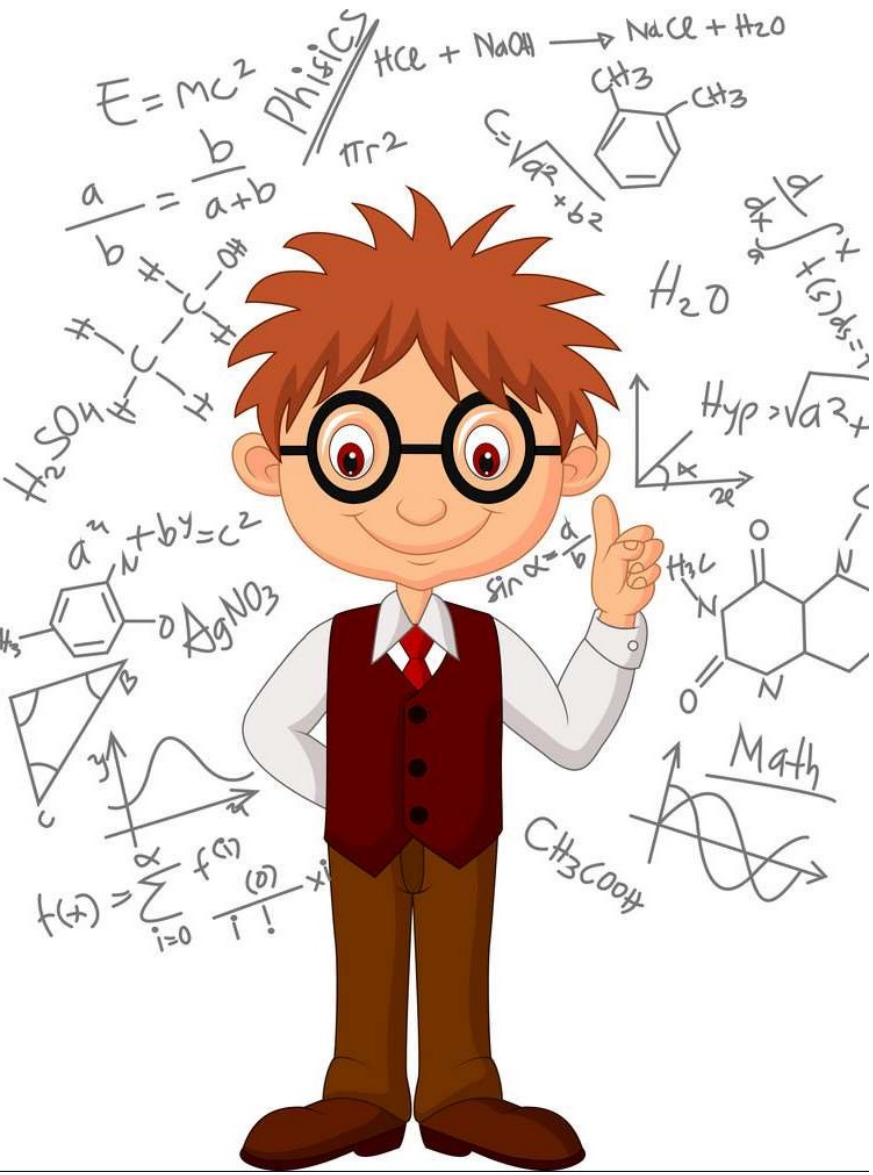
Predictive Data Analysis

Lecturer: Marina Iantorno
E-mail: miantorno@cct.ie
July 2022



In today's class we will cover:

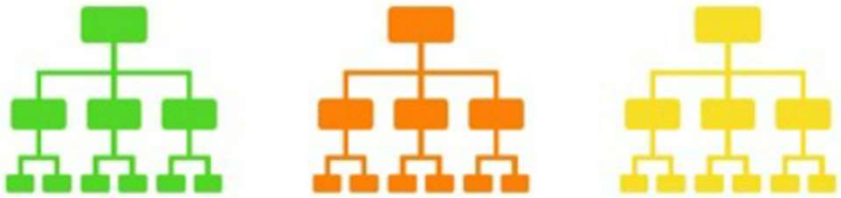
- ❑ Random Forest
- ❑ Evaluation of the model
- ❑ Practice in Python

Regression

RANDOM FOREST

Random Forest

The concept is similar to the DT, but the Random Forest works with Ensemble Learning. This concepts happens when we take multiple algorithms in multiple times, and we put them together to make a model more powerful than the original.



Random Forest

We can analyse the Random Forest with some steps.

1. Pick randomly “k” datapoints from the dataset or the training set (if we divided the data into Training and Test)
2. Build the DT associated to these “k” datapoints.
3. Choose the number of tress that you want to build and repeat the steps 1 and 2. So we will keep building and building (multiple algorithms).
4. Prediction: For a new datapoint make each of the trees predict the value “y” for a specific datapoint, and assign to the datapoint an average across all of the predicted Y values (multiple predictions).

Random Forest

Statistically speaking, when we have an average of the result of the same event happening multiple times, we are closer to get the real result. This is very similar to the concept of the sample size. When the sample size is bigger, it will be more alike the population, and therefore, the results that we get in that analysis will be closer to the “truth”. That is why we say that using Random Forest is a way to be close to the real result when we make the prediction, because we multiply the algorithm many times.

Random Forest

Let’s try it in Python

Regression

REGRESSION MODEL SELECTION

Model Selection

We have covered so far some models to make predictions, and it is important to understand what we are counting, and ultimately choose the best model to predict it.

The best way to know what is the best model is by checking the performance of the model. There is a coefficient that helps us to do it, and it is the “Determination Coefficient”, known as well as r^2.

We already used this coefficient before, now we will use it for all our models and using the library sklearn.

Model Selection

In order to check the performance of our model, we will use the dataset called "Performance.csv" on Moodle.

This is a large dataset, with approximately 10 thousands rows. Just for you to know, there are no missing values in this dataset. But imagine that the last column is our dependent variable and we want to evaluate the performance of all the models.

You will also find a folder called "Templates" and there you will find files with a template for all the models that we have covered, so it will be easier for us to decide which one is the model that best fit to our data.

Let's try this together in Python!

THAT'S ALL FOR TODAY

THANK YOU