**Reply from Matteo:**

Dear group 23:  
  
> The task’s difficulty (compared to the intended project scope).  
  
The project idea seems perfectly compatible with the project scope, and also interesting wrt to the results you can obtain - although you should be ready to expect some "obvious" results.  
  
> The available data – WHO attributes, World Happiness Report. Maybe we will not find reasonable attributes?  
  
I think some useful attributes will be there. Actually, some attributes will probably be very strong predictors, in my opinion, and sometimes correlated - so PCA or related methods might be useful to try. Unhappy countries often have a combination of bad health system, wars, bad economy, ... But you might have to deal with some missing data, which is good as being part of the project.  
  
I assume/hope that there will be some aspects not captured by the attributes (such as cultural aspects) that might lead to some misclassified points. These would be very interesting to look at, if any appears.  
  
> The sample size – only 150 countries on average participate for 8 years. In total there are therefore 8\*150=1200 records (too little?)  
  
That is ok as a size. But you will have to pay attention not to build overcomplicated models.  
  
> The chosen data mining task (labeling) for assigning a score (rational nr. between 0-10) of happiness. Maybe we need to change it to three values, for example? - High happiness, middle happiness, low happiness? because labeling can’t handle the big number of labels (a rational number).  
  
That's a good question that I leave to you to reflect upon during the development of the project. You can have a look at the section of your book about discretisation, that gives some alternatives. We will expect you to "defend" your final choice.  
  
> Using decision-tree algorithm to create a model that can predict happiness. Are there other, more suitable, methods to create better models?  
  
For this type of project I would say that decision trees are a very appropriate tool. You will have to deal with overfitting. I would suggest to use ensembles of trees - that is, random forests, but you can start with simple trees.  
  
I also think that you will have some time left, especially if you divide the work effectively among the three of you, and I suggest to try at least one other type of classifier and make a quick comparison, ultimately arguing which one you think is the most appropriate at the end.  
  
I hope I have answered your questions.  
  
Best,  
Matteo

*ORIGINAL MESSAGE:*

Hi Matteo!

We don’t really have a complete idea on what we want to do but in general we want to see if other factors then those used to explain “the happiness score” in the “World happiness report” (<https://worldhappiness.report/ed/2019/#read>) can be used to explain a country’s happiness and to what extent they are applicable to do that. It would be interesting to see if there are “hidden factors” that normally aren’t considered affecting happiness.

We will use data supplied by the “world happiness report” 2012-2019 and for “other factors” we will use factors such as homicide rates, sickness, literate rate, etc. and those will primarily be extracted from WHO. If we find other interesting world-wide measurements for the period 2012-2018/9 we will use those too. (We are lacking in this area because we haven’t decided on which attributes to use yet.)

**The course of the project would be:**

1st to determine which other factors correlate to happiness from a pre-defined sample (for example homicide rates, sickness, literate rate). 2nd to create a model (maybe tree-based) to label a country’s happiness score based on the chosen attributes.

**Our questions to you are:**

Do you think this is at all possible with respect to:

* The task’s difficulty (compared to the intended project scope).
* The available data – WHO attributes, World Happiness Report. Maybe we will not find reasonable attributes?
* The sample size – only 150 countries on average participate for 8 years. In total there are therefore 8\*150=1200 records (too little?)
* The chosen data mining task (labeling) for assigning a score (rational nr. between 0-10) of happiness. Maybe we need to change it to three values, for example? - *High happiness, middle happiness, low happiness?* because labeling can’t handle the big number of labels (a rational number).
* Using decision-tree algorithm to create a model that can predict happiness. Are there other, more suitable, methods to create better models?

Thanks in advance!

**Group 23**

**Felix Lux Dryselius**

**Ulrika Eriksson**

**Simon Wahlström**