Where are the happiest students?

Analysis of students' well-being through the PISA dataset

Marco Galliani, Giulia Mezzadri, Ettore Modina, Federico Angelo Mor, Beatrice Re

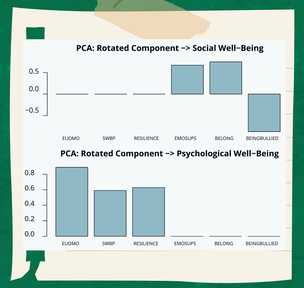
tutored by Chiara Masci and Alessandra Ragni

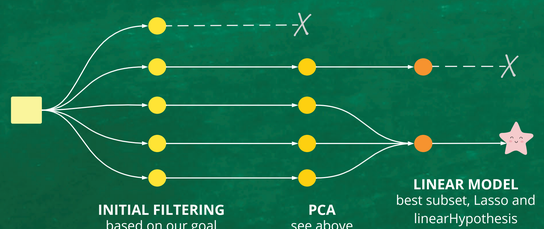
TEAM 12!

Features Extraction

Our goal was to analyze the well-being of the students, but being that a wide concept, there were no variables that measured it directly in the dataset. Moreover, the dataset had a very large dimensionality (there were recorded more than 1400 questions, i.e. possible covariates) which hindered a close approach. To solve these problems we followed this procedure:

- We selected variables with a feasible number of NA and also discarded countries with a large number of NAs;
- Starting from a series of psychological measurements, we computed a score for well-being through PCA: this resulted in two different scores measuring the social well-being and the psychological well-being;
- We further reduced the dimensionality using PCA on groups of similar variables.





Selected Variables (Linear Mixed Model)					
Perceived competence of the students in using ICT devices	Teachers' skills in motivating and stimulating students	Total learning time	Educational and cultural possession at home	Attitude towards learning activities	Performance in PISA test (reading)

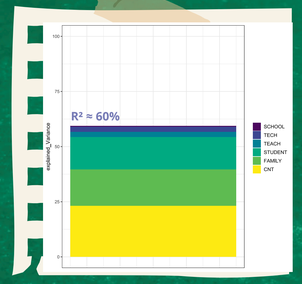
Introduction

This project aims to predict the well-being of students across various European countries by analyzing a range of several factors measured by OECD PISA questionnaires. PISA dataset has a two-level hierarchical structure, where students are nested in schools and schools are nested in countries. In our analysis we grouped students by their belonging schools by averaging student-level variables and account for the grouping induced by countries using mixed effects techniques.

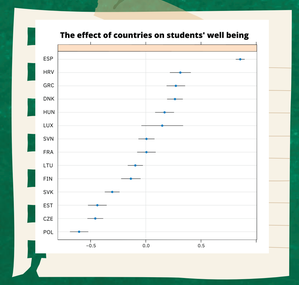
Linear Mixed Models

Models:
We decided to focus on predicting the **social well-being score**. We followed two different approaches to deal with the influence of the countries in the prediction:

1. Assuming independence between observations, we built a simple linear model that used dummy variables to describe the effect of the countries;
2. Denying the independence assumption, we used linear mixed models to account for the dependence between observations induced by the grouping structure.



Variable selection:
We applied forward and backward selection to select the best variables for the LMM. We implemented the two algorithms aiming to maximize the **AIC** of the mixed model. The algorithms converged to the same subset of variables, which we considered optimal. We used an already implemented exhaustive search algorithm for the simple linear model, considering **R^2 adjusted** and **BIC** as performance metrics.

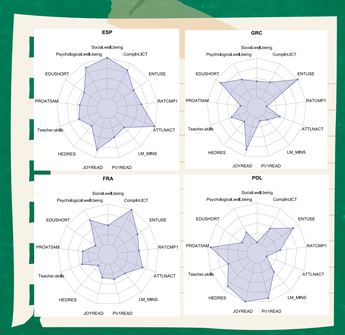


Assumptions:
The normality of the residuals assumption was not fully satisfied as there were some issues in the tails of the qqplot. To fix this we tried by transforming the original variables and by removing outliers.
The latter was the most effective approach, which we implemented using Cook's distance, leverages, standardized, and studentized residual plots. Even if deviations from Gaussianity were shrunk, issues still remained in the tails and this led us to consider also other approaches such as Mixed Effect Random Forest.

Comparison:
Comparing models (1) and (2) we noticed that even if the two are comparable in terms of performance (measured by MSE), model (1) does not fully satisfy its assumptions. Indeed, the independence of the observations is not satisfied, as the variance explained by the grouping induced by countries cannot be neglected.

Manova and Anova

The focus quickly moved into analysing the differences among countries in relation to our set of variables. Using first MANOVA we confirmed the presence of distinct variations in a subset of the selected variables.
Radar plots showed how the sources of variation varied across countries (using the τ_i from ANOVA), revealing noteworthy patterns.
Building on these observations, we further classified the countries into three distinct levels based on significant differences in means. This approach enabled us to effectively differentiate between good, medium, and bad states for each covariate.



References

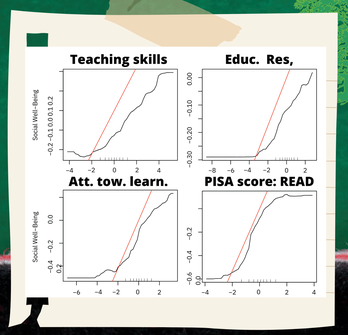
- Game (Julia): Bezanson, Jeff and Edelman, Alan and Karpinski, Stefan and Shah, Viral B: A fresh approach to numerical computing, https://julialang.org/
- MERF: M.Pellagatti, C.Masci, F.Ieva, A.M.Paganoni: Generalized Mixed-Effects Random Forest: a flexible approach to predict university student dropout
- Rotated PCA and Factor Analysis: Procedures for Psychological, Psychometric, and Personality Research
- Distinction between SocialWB and PsychologicalWB: Govorova E, Benítez I, Muñiz J. Predicting Student Well-Being: Network Analysis Based on PISA 2018. Int J Environ Res Public Health. 2020 Jun

Mixed Effect Random Forest

To construct a model that accounted for the **dependence between observations induced by the hierarchical structure of the data**, we built a Mixed Effect Random Forest model, which is composed of two parts:

- a *fixed effect part*, where a random forest is used to estimate **non-linear dependences** between features and target;
- a *random intercept*, that described the **effect of the grouping** induced by the countries.

The difference between LMM and MERF in estimating fixed effects can be appreciated by the partial plots, where the black line represents the marginal effects estimated by the MERF, and the red lines represent the slopes estimated by LMM.



Conclusions

Take home messages:

- Countries matter! As shown by both LMM and MERF, a high portion of the variability of the targets (PVRE over 25% for both models) depends on the countries. This can also be seen by the effects of dummy variables representing countries in the linear fixed effects model.
- Effects of the variables: All our models showed that several factors positively affects students' well-being, such as the attitude toward learning activities, the educational resources available at home, the skills of the teachers in motivating and stimulating the students, and the performances of the students in the reading section of PISA test.

Limitations and future developments:

- From grouping by school to include school effects: We decided to group the dataset by schools to compare as many countries as possible with a feasible number of observations. However, doing so we discarded interesting information about potential school effects, so an extension of our analysis including that may be considered.
- Understanding the impact of countries: As said, the most interesting result of our analysis is the relevance of countries in predicting students' well-being. This result may be futher analyzed gathering country-level data, in order to explain which aspects determined such a difference. Unfortunately such data are not present in PISA dataset, so different sources are needed. Finally, the robustness of the result should be assessed with respect to changes in the computation of the well-being indices.

Let's Play!

To share our results we developed a game: you have to choose your character, the **state** to play with, and how to spend your **budget** on the different categories that we defined with our analysis.
Like if you were the **"Minister of Education"** in that state.
We will then build up a **global scoreboard** with also everyone else who will play the game.
We coded it in Julia and implemented it as a bot on Telegram. Reach it through the QR code or at **@AppStat_ProjectGame_bot**