1. **Abstract (what we want to do)**

In this study we can observe data related to students of 15 year of age coming from European countries among which there’s a significant number of immigrant students. How do immigrant students face up when compared to native students? Which are the main differences in their scholastic, familiar, psychological characteristic? Which features are most important (*relevant*) when it comes to their scholastic success? How can we

take example from the best European countries with respect to integration to help these students?

**(Alternative)** (Abtract è più presentare il contesto/problema e meno quello che si vuole fare 🡪 sezione obiettivi)

High School years are a fundamental stage of development for each individual, and it’s in the interest of every country to ensure that all students have access to the best possible education and are well integrated into the social system. In reality we observe that immigration is a huge factor in denying many foreign students from achieving the same results of their native peers. This trend is common to all countries across Europe but some of them show significantly larger gaps with respect to the others.

In this study we aim at investigating quantitatively the reasons behind all these differences.

1. **Introduction/objectives (how we want to do it)**

The dataset we focused on is based on some tests given to students, their parents and their schools, regarding both scholastic knowledge and social questions.

After a general visualization of the dataset and an analysis of it through Manova tests and Clustering methods, we focused on the students’ scores in math and reading noticing that there were differences between native and immigrant students in term of their results.

We proceeded with a deeper analysis and the build of some models, which highlighted that the score of a student is affected by some features as his socioeconomic state, the amount of time he dedicates to study, his class’s size and some others, besides his hard work and study. The obtained results suggested us some ways to help the lagging students, who most of the times were the immigrant ones, understanding the points that could be improved.

Then, focusing on schools and proceeding with a Multinomial Regression, we clustered schools with respect to some features that affected the results of their students, highlighting what makes a school the best in the sense both of results and of integration.

1. **Materials & Methods (how we did it)**

* Data selection

Starting from the data of Pisa’ OECD program of 2018, which contains student and school answers to standardized questionnaires, we restricted our analysis to 10 selected countries which had a sufficient sample of immigrant students: Austria, Belgium, Switzerland, Germany, Denmark, Spain, Great Britain, Italy, Luxemburg, Sweden.

Chart, sunburst chart

Description automatically generated

Since the number of covariates available was more than 1000 (one for each questionnaire answer), we selected the most relevant ones related to our focus on immigration. Some of the questions answered by the students were already aggregated by Pisa’s group and resulted to be the most complete.

The features selected range over many fields: ESCS status, Immigration status, teacher support, sense of belonging, class size, at home weekly learning time, etc.

* Clustering

We first qualitatively visualised our dataset trough clustering. Using k-means, we came up with 3 clusters which are not perfectly divided and meaningful but seem to confirm our hypothesis that immigrants are generally less successful at school.

Indeed, the cluster linked to the best performances contains only the 23% of the total number of immigrants, while the worst one contains the 54% of them. Moreover

* Manova

We then used Manova to get statistical evidence for our hypothesis of there being differences between native and immigrant students in math and reading scores. This was confirmed for all European countries (except Great Britain). By ranking them, we see that Great Britain is the country with least score difference, while Denmark is the worst among the selected countries.

After confirming this hypothesis, we further investigated differences between these groups in ESCS status and learning time.

Learning times for immigrants are consistently more in all countries with respect to natives’ times.

* Linear models

After assessing the presence of differences in scores between immigrants and natives we tried to understand which features are most important for a student’s scholastic success, to find the optimal way for schools to help lagging students and to have a better integration system.

By focusing our aim on three datasets (European aggregated dataset, Great Britain as the most performing country with respect to integration, and Denmark as the worst one) and using as our target variable the scores of math and reading, we selected the best model through backward selection.

* LMM

To measure the effect of living in a certain country on the students’ performance, we implemented a linear mixed model aggregating the observations through countries. To better focus on the aim of our analysis we also added a random effect associated with the variable immigration.

# results: ha effetto sia il country che immig (passo da PVRE del 2.5% a 4.4%), noto che nel modello finale con anche immig GBR ha effetto positivo rispetto alla media se sei un immigrato, mentre DNK effetto negativo

We performed two other mixed models in order to detect the differences across schools in Great Britain and Denmark. Indeed, we aggregated the students through their schools trying to detect the main reasons why Great Britain is an effective model of integration as opposed to Denmark.

* Multinomial regression models

We divided the schools in three clusters, ranked trough the effect on the scores as highlighted by the mixed models.

To explore a posteriori the school clusters we applied a multinomial logistic model, which allowed us to identify the different characteristics of the best and worst schools

1. Results ( what were the results)
2. Conclusions (What do we learn from the results)