**Project on OECD Pisa Dataset ( 30/03)**

**Abstract**:

Imagine you are the Italian state. One of the most important things for the country is having brilliant students and a good level of education. You would like to allocate your tax money well but where to start? There are so many variables….

Wouldn’t you like to know which are the most important factors which influence the level of your students? What about school funding, school IT level, if the students have a good family background, if they have access to computers, etc.…

Moreover, what is the situation in other states in Europe (or in the world even)? Are the students performing better or worse? Which may be the reasons? Are there significant differences in some key features (such as school funding, etc.) which would explain this difference?

Finally, has the overall level of education change throughout the years? What about other features?

Is there a way to correlate these changes, to explain them?

**The main categories of features we have in our datasets**:

STUDENT DATASET (each row is a student)

* Wealth indicators
* Immigration status indicators
* Home and family: level of study of parents, book in house, quiet place to study
* Stress indicators
* School support: teachers, resources, classes hours per week

SCHOOL DATASET (each row is a school)

* IT level
* Founding coming from the state etc.…

**Come procedere**:

1. Selezionare features rilevanti (fatto)
2. Studiare normalità
3. PCA (?)
4. Strutturare meglio la nostra indagine!! Cosa vogliamo fare?

E.g. Take the regions of the school of each student. Can we do an ANOVA to see if the students in certain regions perform better than other regions? From there can we proceed to understand which are the key differences for this?

**Lista di possibili inferenze**:

**Domande:**

1. Alcune features appartengono alla stessa area semantica (e.g. ricchezza: numero di bagni, numero di computers, cellulari etc..). Meglio prenderne uno a caso rappresentativo oppure fare una PCA e prendere la prima principal component.
2. Dopo aver selezionato le features che ci interessano, il data frame ha alcune righe con NA per qualche feature. Rimuoviamo le righe o prima dobbiamo controllare che alcune colonne non siano tutte NA?
3. Come facciamo a rendere gaussiano il nostro dataset? Serve oppure n è sufficientemente grande (n<5000 circa)?
4. Come utilizziamo le variabili categoriche? Solamente per discriminare il dataset per delle ipotesi?
5. Per le scuole che hanno molti pochi studenti, conviene rimuoverle?
6. Differenza tra vocational e general in iscedo?

**PIANIFICAZIONE del progetto 23\_03**

Prima a livello di paese (Italia):

1) Selezionare le features più interessanti, fare PCA per capire le caratteristiche chiave che distinguono gli studenti.

2) Prendere uno a caso da gli esiti degli esami degli studenti. PCA/media per avere un unico voto (volendo)

3) Che cosa ci interessa inferire statisticamente?

* Regressione sul voto dei test in base ai principali component trovati nel punto 1

Dopo tra paesi (Italia vs Inghilterra) oppure tra regioni del mondo (Europa vs Nord America):

**FEATURES interessanti per dataset studenti**

CNTRYID = nome paese

STRATUM = regione + tipologia di istituto (per italia lice/istituto tecnico/professionale/centro, scuola media ecc..)

BOOKID??

ST001D01T = grade

ST003D02T = mese di nascita (indagine zodiacale)

ST004D01T = gender

ST005Q01TA = level of schooling mother

ST007Q01TA = level of schooling father

ST011Q01TA = desk in your home

ST011Q02TA = room of you own

ST011Q03TA = quiet place to study

ST011Q04TA = computer you can use for school

ST011Q07TA = classic literature in your house

ST011Q16NA = art/music/design books(NA ? or TA?)

ST011Q17TA = wealth item 1 (also 18 19 for item 2 3)

ST013Q01TA = how many bo

PA042Q01TA = household income (categorico)

WB03

HISCED (vs HISCED D, qual è la differenza?)

FISCED = educazione madre

MISCED = educazione padre

**Quali altre domande possono essere utili?**

Paese di nascita uguale al paese selezionato?

Che lingua parli?

Hai uno strumento musicale a casa?

Come sono i professori alla tua scuola, se ti ascoltano ecc..?

Sei stato bocciato?

Ti piace leggere per conto tuo?

Che cosa leggi in particolare (fiction/newspapers/ecc..)?

Nato nel paese, seconda generazione, terza ecc…

Vigorous activities?