A Data Driven approach for evaluating foundation skills of adults

Elena Salvatori[1], Lorenzo Gabrielli[2], Fosca Giannotti[2], and Dino Pedreschi[1]

[1] Department of Computer Science, University of Pisa, Pisa, Italy
{elena.salvatori,pedre}@di.unipi.it
[2] ISTI, National Council of Research, Pisa, Italy
{lorenzo.gabrielli,fosca.giannotti}@isti.cnr.it

Abstract. This work deals with quantifying skills use, reading, writing, numeracy, and ICT, at work and at home. It aims at contributing to the debate on developing policies transforming "better skills in better jobs, social inclusion and economic growth". Thanks to data collected by the Organization for Economic Cooperation and Development (OECD), we can observe variables such as the demographic status, education level and the ICT usage. Our goal is to understand how target variables (foundation skills) are linked to the use of reading, writing, numeracy and ICT skills at home and at work. The dataset covers around 4600 individuals representing the Italian population. We build a data-driven analytical process to help the national adult education system to evaluate determinants of literacy and numeracy skills of adults. In the future we would like to investigate further in two directions. First, we plan to consider additional factors that would allow an Italian adult to move towards a higher proficiency level. In addition, we aim at building a classifier to predict achievement levels and validate it based on the dataset resulting from a case study performed with a group of adult students.

Keywords: Data Mining, Clustering, Multidimensional Scaling, Fundamental skills, ICT skills, Adult Education

1 Introduction

European policymakers are aware of the importance of skills to sustain economic growth and social cohesion [4]. The EU 2020 strategy for inclusive growth, supported by the two initiatives "Agenda for new skills and jobs" and "European platform against poverty and social exclusion", has established the following objectives for education attainment by 2020: i) at least 40% of the population of the age-group 30-34 years-old, completing the third level of education (or equivalent), ii) reducing school drop-out rates below 10%.

"Better skills. Better jobs. Better lives" [7] contains the strategic approach to skills policies of the OECD to help countries to analyze their national skills systems, develop policies aiming at translating better skills into better jobs, social inclusion and economic growth. The underlying assumption is that good skills reduce the risk of unemployment, and are positively related to income, health, and social trust as well. In this framework, OECD has started a comparative survey on education called the Programme for the International Assessment of Adult Competencies (PIAAC). The OECD Working paper on "Measurement and Analysis of the Digital Economy, Skills for a digital world" [8], analyses the effects on skills for citizens and workers in the digital economy. Despite the title, authors suggests policy-makers to put in place all actions ensuring citizens to master digital skills as well as "strong foundation skills, higher order thinking competencies, and emotional skills to respond to greater levels of uncertainty".

With our work we would like to contribute to the debate on the skills a person should have in order to study, live, and work in the digital world, considering the perspective of adults' education. Thanks to a data-driven analytical process carried out on the national dataset on literacy, numeracy, and ICT skills usage, we were able to explore behaviors along several dimensions. Our results integrate the knowledge on the domain with visual representations of the pattern in the dataset obtained applying clustering and multidimensional scaling techniques. From our analysis, it emerges that relevant features related to high performing individuals in the sample are those related to: a) being employed or self-employed; b) intense use of writing, reading, numeracy, and ICT skills at work and at home or, if unemployed, c) to be young and use intensely writing, reading, numeracy, and ICT skills at home.

The work presented here is part of the PhD thesis entitled "Data analytics for educational process, a case study for assessing foundation skills of adult students"(Elena Salvatori, Universty of Pisa, XXX Ciclo, a.a. 2014/2015). The paper is organized as follows. After defining the research problem addressed in the analysis in Section 2, we will describe the dataset and the adopted methodologies in Section 3. In Section 4 we will show the results of the analytical process performed and Section 5 will present the conclusions and possible future works.

2 Problem definition

The results of the PIAAC analysis are well-known in Italy. In the literacy and numeracy ranking of the OECD countries, Italy alternates between the last and second last position. More than 70% of Italians scored at Level 2 or lower and only a minority, 30%, scored at Level 3 or above, against the OECD average of 47% where Level 3 is considered the level that matches the minimum necessary skills to live and work nowadays, see Section. 3.1.

Moreover, national figures about education attainment of Italians in the age group 18-64, show that 54% has no secondary education attainment, 34% has a secondary diploma and the remaining 12% has a tertiary qualification against the OECD average of 27%, 43%, and 29% respectively [2]. In Italy 17% of pupils quits the school at early stages, furthermore Italy with other 5 European countries, has one of the highest drop-out rate [3].

The national guidelines on adults' education [6] define the central role of the schools offering courses to adults for the development of life-long learning skills.

Nowadays in Italy, most of the job opportunities require a secondary diploma and adults' schools represents an unique opportunity for those individuals, often dropped-out from school in their teens, to achieve a secondary diploma.

The research questions driving the data-driven analytical process were the following:

Q1: Can we improve our knowledge on fundamental skills of Italians applying Data Mining techniques?
Q2: To which extent demographic status, education level and the use of ICT can determine a successful occupational outcome?
Q3: Which factors should be taken into accounts by adults' education in order to improve the provision of skills required by the labor market?

3 Materials & Methods

In this section, we briefly introduce the PIAAC Methodology, the dataset and the selected Data Mining methods of our study.

3.1 The PIAAC Methodology

In our datasets we used the measures of fundamental skills collected in 2012 in the PIAAC survey (Programme for the International Assessment of Adult Competencies) in Italy. PIAAC underpins the following definition of literacy, numeracy and problem solving in rich technological environment(PSTRE), [2].

Literacy is the ability to understand and use information from written texts in a variety of contexts to achieve goals and develop knowledge and potential. This ability is considered a requirement for developing higher-order skills and for "positive economic and social outcomes".

Numeracy is the ability to use, apply, interpret, and communicate mathematical information and ideas. This is an essential skill in the information society where a wide range of mathematical and quantitative information is accessible in the daily life.

Problem solving in technology-rich environments (PSTRE) "This refers to the ability in using technology to solve problems and accomplish complex tasks". The test does not measure "computer literacy", but rather how information is used and evaluated to solve a problem through higher-order skills.

The Literacy and Numeracy achievement levels are computed from the score (number 0-500), as follows: Below Level 1 (score < 175), Level 1 (score from 176 to 225), Level 2 (score from 226 to 275), Level 3 (score from 276 to 325), Level 4/5 (score > 326). For PSTRE, level values are four: Below Level 1 (score $<$ 240), Level 1 (score from 241 to 290), Level 2 (score from 291 to 340), Level 3 (score > 340).

For OECD, proficiency is reached with Level 3, in the case of literacy and numeracy, and with Level 2 for PSTRE. These correspond to the levels that match the minimum necessary skills to live and work nowadays. Other measures of interest in our study are derived from the non-cognitive module called **Skill Use Questionnaire** of the PIAAC Background Questionnaire. This module assesses the skills that respondents use at work and in their daily lives. The focus of the questions is on reading, writing, use of mathematical information and idea and information and communications technology (ICT).

3.2 Dataset

The dataset of this experiment is based on the OECD Public Use Files of the last Survey of Adult Skills that took place in Italy in 2012 [3]. The Dataset consists of 4600 records and 20 attributes. It stores demographic status, occupation, education, participation in formal education, and Skills Use of the Italians.

From [1], we computed also two new variables quantifying individual proficiency called *Proficency Literature Level* (PVLitLevel) and *Proficency Numeracy Level* (PVNumLvel). The variables contain the achievement levels in the two domain and were computed from the medians of the ten Plausible Values for Literacy (PVLit1:PVLit10) and Numeracy (PVNum1:PVNum10) [4] of the Italian sample. Table 1 displays the demographic, educational and occupational statistics of the sample. We report in brackets the corresponding values in the dataset.

The Skill Use variables are 14, they contain the values summarizing the answers to several questions of the Skill Use Questionnaire as described in Section 3.1. Skills Use at Home and at Work variables contain the level of use of reading, writing, numeracy, and ICT skills and their values vary from 1, lowest level of use, to 5, highest level. For Yes/No questions, the value 1 indicates the yes answer and 2 the no answer. Not reached questions, skipped answers and missing values have been coded with number 9.

3.3 Methods

The data-driven analytical process included two well-known Data Mining techniques: clustering and multidimensional scaling. These analytics do not make preventive assumptions on data and are actually aimed at extracting new knowledge from them for generating further hypothesis [5]. Both methods summarize data in clusters or segments composed of similar cases and are based on the definition of similarity/dissimilarity measures.

1. Clustering groups multdimensional data points in a way that points in the same cluster are very similar one to each other, and are as different as possible from points in other clusters. The data in a cluster are summarized by a

[3] Around 4600 adults aged 16 to 65 were surveyed and 1329 attributes were stored (14.7 MB)

[4] Plausible values (PVs) are multiple imputations of the unobservable latent achievement for each student, and are used to estimate population parameters in large-scale assessment programs.

	Percentage
Gender	
Male (1)	50.0
Female (2)	50.0
Age group	
16-24 (1)	14.4
25-34 (2)	18.9
35-44 (3)	24.4
45-54 (4)	21.8
55-65 (5)	20.5
Education	
Below secondary (1)	53.4
Secondary (2)	33.8
Above secondary (3)	12.1
Occupation	
Employed (1)	55.8
Unemployed (2)	9.0
Out of the labor's market (3)	34.5

Table 1: Demographic features of the Italian sample, PIAAC 2012; the sample is representative of the Italian population.

centroid, which is a vector containing for each variable, the mean value, the standard deviation, and other statistics. After the analysis, instead of inspecting individual records, one can look at their cluster centroids.

2. Multidimensional scaling (MDS), is a dimension-reduction technique that arranges data-points in a bi-dimensional scatter plot preserving distances between pairs of data-points.

4 Data Analysis

In our first analysis, we applied the K-Means clustering model to the Dataset. This was done after computing the optimal number of cluster k (7 in our case), with the silhouette method. The Fig. 2 shows the centroids matrix where the columns are the means vectors of the seven clusters, and the rows are the variables of the observation.

Analyzing the Table 2, we can observe some of the determinant variables that the Kmeans algorithm used to assign a record to a specific cluster. These were the answers to the questions ”Have you ever used a computer?”, ”Do you use a computer at work?”, ”Do you use a computer in your everyday life now (outside work)?”. Based on the analysis of the centroids matrix, we have characterized the seven clusters as follows.

- **Cluster 0** can be defined as the *mature skilled workers* cluster because it is mainly composed of workers, and the average age is 50,87 years. It has

Cluster centers	0	1	2	3	4	5	6
Age	50.87	31.54	50.63	34.01	40.90	60.44	21.36
Gender	1.42	1.39	1.47	1.48	1.80	1.66	1.57
Occupation	1.07	1.37	1.31	1.08	2.72	2.95	2.81
Age group	4.14	2.21	4.09	2.45	3.17	4.88	1.26
Education	2.14	1.60	1.42	2.24	1.59	1.38	1.63
Formal education?	0.04	0.32	0.21	0.14	0.04	0.00	0.73
Computer at work?	1.0	2.2	2.2	1.0	9.0	9.0	9.0
ICT level job?	1.8	9.0	9.0	1.8	9.0	9.0	9.0
Right ICT skills?"	1.1	9.0	9.0	1.0	9.0	9.0	9.0
Lack of ICT skills?	1.9	9.0	9.0	1.9	9.0	9.0	9.0
Ever used a computer?	9.0	1.3	1.7	9.0	1.3	1.6	1.0
Computer at Home?	1.2	2.2	5.4	1.1	3.6	6.0	1.2
ICT home	4	4	6	4	5	7	4
ICT work	3	9	9	3	9	9	9
Numeracy home	2	1	1	2	2	1	3
Numeracy work	3	1	1	3	9	9	9
Reading home	3	2	1	3	2	2	3
Reading work	3	1	1	3	9	9	9
Writing home	2	2	1	3	2	1	3
Writing work	3	1	1	3	9	9	9

Table 2: **Cluster 0** is the *mature skilled workers* cluster; the average age is around 50, and the variables indicating the usage of ICT at work and at home show high values. **Cluster 1** is the *young unskilled workers* group, the average age is around 30, they do not use ICT at work and their usage of fundamental skills at work and at home is low as in **Cluster 2**, the *mature unskilled workers* cluster whose average age around 50. **Cluster 3** is the *young skilled workers* group, the average age is around 34, and is quite similar to cluster 1. **Cluster 4**, *unemployed women*, and **Cluster 5**, *elderly people*, group individuals who, on average, do not use ICT at home and have a low level of usage of foundation skills. **Cluster 6**, *young adults*, group individuals who use ICT and foundation skills in their daily life and participates to formal learning.

almost the same number of males and females, with education above diploma, use a computer at work and outside work. On a scale from 1 to 5, the mean values for skills usage at home and at work, is around 3.

- **Cluster 1**, can be named as the *young unskilled workers* cluster. The average age is 31,54 and it has a majority of males, mostly employed, without secondary education who do not use a computer neither at work nor outside work. The mean values for skills usage at home and at work are below 2 or missing. A minority of them participated in formal adult education in the previous 12 months.
- **Cluster 2**, is made of *mature unskilled workers.* The average age is around 50 years and for the remaining variables, its centroid is quite similar to that of cluster 1.
- **Cluster 3**, shares the same mean values for many variables with cluster 0. But in this cluster, the average age is 34 years and it has the highest estimates for literacy and numeracy levels (around 2,6). It can be defined as the *young skilled and literate workers* cluster.
- **Cluster 4**, *unemployed women*, is mainly composed of unemployed women looking for a job; the average age is 41, education is below secondary, the majority have used a computer but do not use it in the daily life. Mean values for skills use at home are around 2 while the skills use at work values are missing.
- **Cluster 5**, is composed of those who do not work anymore and can be defined as *elderly people* cluster, the average age is 60, and the level education is below secondary. The majority has never used a computer but some of them use ICT at home. The skills use of writing, reading and numeracy at home is low, mean of 1 on a scale from 1 to 5.
- **Cluster 6**, *young adults*, is composed mainly by individuals out of the labor market, whose age is on average 21. Most participated in formal education during the 12 months before the interview and do not own a secondary diploma. They all use a computer and make an intensive use of ICT skills at home (4 out 5) as well as of reading, writing, and numeracy skills.

To have an estimate of the mean proficiency of each cluster, we computed the average values for the PVLITLevel PVNUMLevel variables (Tab. 3). Cluster 0 and 3, show the second and first highest values in the estimation of literacy and numeracy achievement levels so that their definition *mature/young skilled workers* can be further refined in *mature/young skilled and literate workers. Young unskilled-workers* and *mature unskilled workers*, have, on average, a value less than 2 both for literacy and numeracy achievement levels. For cluster 4 and 5, proficiency levels estimates are quite low (below 1,75 out of 4) while the proficiency estimators for cluster 6, *young adults* are, on average, above 2 so that we can better qualify this group as *literate young adults.*

We then performed a visual analysis of the dataset. Visualization techniques allow to plot high dimensional data in a bi-dimensional space making it is easier

Cluster centers	0	1	2	3	4	5	6
PVNUMLevel	2.46	1.96	1.74	2.62	1.73	1.59	2.14
PVLITLevel	2.38	1.95	1.85	2.60	1.87	1.69	2.38

Table 3: Clusters' mean of numeracy and literacy proficiency. The best performing cluster is cluster 3, *young literate and skilled workers*, followed by *mature literate and skilled workers* of cluster 0, and by *literate young adults* of clusters 6. The other four clusters, corresponding to *elderly people, young and adult unskilled workers, unemployed women*, on average, have worst results

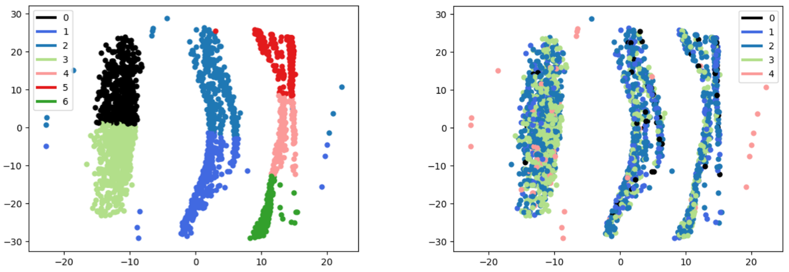
.

to inspect basic relations between variables. The dataset of this analysis, is the previous dataset extended with the new variable cluster identifier, and the techniques used was MDS.

In Fig. 1, (a) is the scatter-plot resulting from the processing of our dataset stratified by cluster id. Every data-point indicates an adult, represented after the multidimensional scaling. We can easily spot three separated and parallel large groups or clouds, each is further separated into subgroups. The central cloud is divided in two and is composed of data-points belonging to cluster 1 and 2, that is *young unskilled workers* and *mature unskilled workers.* The left plot is horizontally separated into two parts of black and pale green data-points belonging to cluster 0 and 3, *mature and literate skilled workers* and *young and literate skilled workers* cluster respectively. The right cloud displays three horizontally separated clusters: cluster 4, 5, and 6 and display green-data points for individuals belonging to the *young adults*, pink-data points for *unemployed women*, and red data-points for the *elderly people* cluster.

A second graphical analysis with MDS was performed to observe the literacy and numeracy proficiency level of each single test-taker. In Fig. 1 (b) every data-point is colored according to estimate of individual performance in literacy. The five colors represents the achievement Levels from 0, lowest level, to 4, highest level. In this image we have 3 well separated groups as before, although in each group we can not see the separation into two or three clusters as in the first visualization. Higher proficiency levels, represented as green and pink data-points, are denser in the left cloud and at the bottom and center of the left cloud. We know from the previous analysis that these data-points belongs to one of the following clusters: *mature and literate skilled workers*, *young and literate skilled workers* or *literate young adults.* The lowest proficiency level, represented by black data-points, is denser in the central and right groups as well as blue data-points, indicating level 1 and 2. The central area of the scatter-plot displays the data-points belonging to the *young unskilled workers* or *mature unskilled workers* cluster. The right part of the graph displays the *young adults*, *unemployed women*, and *elderly people* cluster.

The variables that separate, also visually, the groups are occupational status, ICT usage and age. Occupational status separates those out of the labor market,



(a) Clusters

(b) Proficency Levels

Fig. 1: The Multidimensional Scaling of PIAAC Dataset produces a bi-dimensional representation of the data obtaining three well separated groups. *Skilled workers* mainly belong to the left cloud, *unskilled workers* to the center cloud, *unemployed* belong to the right cloud. The first picture represent the clusters obtained (a). Clusters are well divided with respect of occupation status, use of ICT at work, and ordered from the left to right of the plot. Data-points are ordered from the bottom to the top according to age. The second picture represents the proficiency levels (b). The left group mainly contains individuals with higher levels of proficiency (Level 3 and 4), while the other two groups contain people with lower levels of proficiency (Level 0 and 1 and 2).

at the right, from those in the labor market, at the center and left. Workers using ICT at work are at the left, and those not using ICT at the center of the scatter-plot. Clusters are visualized according to the average age of its components; at the bottom of the plot, we find data-points from younger clusters while at the top those from older ones.

5 Conclusion and Outlook

The results of our data-driven analytical process integrate those of the PIAAC Italian Report [2]. Thanks to the analysis performed, we grouped the Italian sample in seven homogeneous sub-groups with common behaviors of skills usage. Both analysis' determinants were the ICT usage at work, at home and occupation. The clusters' inspection confirmed the positive relation between individual proficiency levels and skills use, occupation, age, and education in [7]. The visual analysis confirms spatially, the "digital division" of the Italian sample between those that use ICT at work and those that do not. Data-points are plotted according to use of ICT at work and age, and better achievement levels can be found in clusters corresponding to test-takers declaring to use ICT at work and at home, or, if unemployed and young, having high levels of skills' use at home, see Fig. 1. To conclude, the answer to Q1 "Can we improve our knowledge on fundamental skills of Italians applying Data Mining techniques ?" is affirmative.

Besides the demographic status and education attainment, the relevant features of high performing Italians are the use of writing, reading, numeracy, and ICT skills at work and at home. The answer to Q2: ”To which extent demographic status, education level and the use of ICT can determine a successful occupational outcome ?”, is that the clusters with the best occupational status are also characterized by high level of use of fundamental skills, and ICT skills at work and in daily life. Considering the importance of ICT skills, the answer to Q3: ”Which factors should be taken into account by adults’ education in order to improve the provision of skills required by the labor market ?” is that digital skills should not be given for granted in every age-group and that educational institutions should ensure e-skills at the Level 2 of the PSTRE domain. Also, adults’ education should make sure that every student, at the end of the secondary cycle, has a proficiency in literacy and numeracy at the Level 3 or above of the PIAAC scale as suggested by OECD.

In the future we would like to apply our data-driven analytical process to other OECD datasets, and to further investigate in two directions. First, we plan to understand additional factors that would allow an Italian adult to move towards a cluster with higher proficiency levels. In addition, we aim at building a classifier to predict achievement levels and validate it based on the dataset resulting from a case study performed on adults students.

Acknowledgments This work is partially supported by the European Community’s H2020 Program under the scheme ’INFRAIA-1-2014-2015: Research Infrastructures’, grant agreement #654024 ’SoBigData: Social Mining & Big Data Ecosystem’. (http://www.sobigdata.eu) WP:Well-being & Economy.

References

1. Public use files of the survey piaac, piaac international code book (2016), http://vs-web-fs-1.oecd.org/piaac/puf-data
2. DiFrancesco: Piaac-ocse, rapporto nazionale sulle competenze degli adulti. ISFOL, Temi e ricerche (2013)
3. European-Commission: Tackling early leaving from education and training in europe: Strategies, policies and measures. Eurydice and Cedefop Report. Luxembourg: Publications Office of the European Union (2014)
4. European-Commission: Europe 2020, the eu strategic growth strategy (2015), http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/priorities/inclusive-growth/index_en.htm
5. Michael R. Berthold, Christian Borgelt, F.H.F.K.: Guide to Intelligent Data Analysis, How to Intelligently Make Sense of Real Data. Text in Computer Science, Springer (2010)
6. MIUR: Linee guida per il passaggio al nuovo ordinamento a sostegno della autonomia organizzativa e didattica dei centri provinciali per la istruzione degli adulti cpia. Official Journal no. 130 of June 8, 2015 (2015)
7. OECD: Better skills, better jobs, better lives /content/book/9789264177338-en
8. OECD: Skills for a digital world /content/workingpaper/5jlwz83z3wnw-en