Examining Pupils' Achievement in Primary and Secondary Schools in Greece

Ilias Papadogiannis, Manolis Wallace, and Vasileios Poulopoulos

***Abstract* — This study examines the academic performance and demographic characteristics of Greek students in elementary and secondary school. It is an extension of an earlier study that was done in a different time frame to support related conclusions. The dataset includes all students in the last two grades of primary school and the first three years of secondary school. The academic success levels, as well as their longitudinal dimension and divergence by demographic factors, were identified. The results of our previous study, which showed that there are four consistent levels of academic achievement across time, were confirmed by the current research. In addition, the significance of demographic characteristics such as gender and guardian occupation were verified. Last but not least, the importance of early identification of low-performing students was emphasized, as was the likelihood of a significant improvement in their performance. We think that one of the biggest problems with making good educational policies is that it's hard to help students who aren't performing well.**

***Key words* — academic performance; clustering; elementary and secondary education; unsupervised learning; X-means algorithm.**

I. Introduction

This paper attempts to confirm the results related to academic achievement in primary and secondary education in Greece. A specific number of academic levels emerged in our earlier article [1]. Additionally, it has been shown that diverse demographic characteristics are not independent of students' academic achievement and that the level of performance is constant over time. These results provide some proof of Coleman's findings [2] and Bourdieu's views regarding the importance of socioeconomic and educational background as mechanisms for reproducing social inequalities, respectively [3]. In addition to the importance of students' social characteristics, there is an extensive amount of study on the variables that affect student achievement. The non-cognitive components include student attitudes and motivation [4], 5[], self-concept [6], self-regulation [7], self-esteem [8], [9], and goal orientation [10], that are further aspects of emotional intelligence. Leadership [11], school culture [12], school climate [13], educators' expectations [14] and parental involvement [15] are all aspects of the learning environment.

The importance of socioeconomic factors was supported by a significant meta-analysis of 2,138 studies [16]. They were thought to have a considerable impact on performance. Expectations of instructors and school climate also had a moderating effect. Minor influences came from leadership, learner attitude and self-efficacy, stress, motivation, and family support.

II. Educational Data Mining

A. Field Review

Inferences from educational data analysis can be made using the rapidly growing field of "educational data mining" (EDM). This also aids decision-making at many levels, from the classroom to the top of the educational pyramid. Many sub-fields have emerged as the number of EDM studies has increased and many literature reviews have been conducted [17]–[23]:

- The focus of Academic (AA) and Institutional (IA) Analytics is the collection, analysis, and visualization of educational data pertaining to courses, curricula, and other research for institutional use. It emphasizes political or economic issues as a result.
- The analysis of teaching activities is known as Learning Analytics (LA). It is mostly a tool for teachers who want to use the results to come up with, carry out, and evaluate educational projects [24].
- Big Data in Education (BDE), or the process of drawing enlightening findings from a significant volume of educational data [25].
- The field of Educational Data Science (EDS) uses educational data to address educational problems [26].

Numerous new aims have been introduced. Evaluation of educational theories, pedagogical methods, and cooperative learning, curriculum analytics, visual analytics for learning, deep learning, causal relationship discovery, early warning systems, self-regulated learning, emotional analytics for learning, assessment of the effectiveness of intervention, game analytics for learning, interpretable and explanatory learner models, multimodal analytics for learning, sentiment analysis, transfer learning, and comprehension of navigation paths.

The number of publications is increasing, going from 1.000 articles in 2012 to 6.000 in 2018 [20]. Analyzing academic outcomes is a typical method of gauging the success of interventions and a common aim of many studies. For educational data mining, unsupervised, supervised, and semi-supervised approaches are used. Studies that use supervised methods, such as regression and classification, are widely used in the scientific literature and consistently produce positive results [22]. One distinguishing feature of the research is its focus on quantitative data and higher

Submitted on October 19, 2022.
Published on December 29, 2022.

I. Papadogiannis, ΓAB LAB – Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, Greece.
(e-mail: i.papadogiannis@uop.gr).
M. Wallace, ΓAB LAB – Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, Greece.
(e-mail: wallace@uop.gr)
V. Poulopoulos, ΓAB LAB – Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, Greece.
(e-mail: vacilos@uop.gr)

education. Less frequently used are unsupervised or semi-supervised machine learning techniques [20]–[23]. Higher education is where most studies on unattended learning are done [27]. Academic achievement levels and student performance in higher education is a common field of study.

The main benefit of adopting an unsupervised learning algorithm in this study is that it allows for data-driven conclusions to be drawn without the involvement of researchers. The x-means clustering algorithm [30] allowed for the evaluation of academic achievement levels without making any assumptions about their number. Clustering is also used to first divide the levels of performance into variables, which are then used in the analysis that follows.

B. Research Questions

We wanted to corroborate the results of our earlier study in this one over a longer time span. The goal of this study was to determine the levels of academic achievement in primary and secondary education, to investigate the stability of these levels over time, and to assess the variation in achievement levels based on specific demographic factors. This study's objectives were in line with the research questions of the previous study [1]. Particular attention was paid on four research questions (RQs):

RQ1: Identifying the number of student achievement levels and their frequencies.

RQ2: Examining the trends of students' achievement over time.

RQ3: Analyzing the impact of demographic variables.

RQ4: Comparing the findings from the three study questions to the previous survey's results [1].

III. Data and Context

In Greece, primary and lower secondary education, since we also have high schools, are both mandatory and last nine years, as long as the student has not reached 16 years of age. Along with this constitutional requirement, the government also requires two additional years of preschool instruction beginning in kindergarten. Information on the students is gathered and recorded at the three levels mentioned above. Students' data has been methodically entered into a centralized MIS since the 2015–2016 school year in order to guarantee that their profiles are consistently updated from kindergarten through high school graduation. The management information system is called "My School" and is related to both elementary and secondary education.

Data entry and user access quality control are ongoing processes. The school, which is the main data entry entity, must also enter data from the various levels of the educational pyramid. The information system gathers academic data on students (grades, absences, attitude), as well as socioeconomic status information (occupation of guardians, nationality, region of residence, etc.). Additionally, information on the teaching staff is provided (qualifications held, classes taught, years of service, teaching hours, contact details, etc.). The infrastructure of the educational units is also recorded in a systematic way. This includes the building, the teaching materials, and the equipment.

In this way, a large amount of data is gathered for use in decision making. Due to this, only the Greek Ministry of Education's central services have access to information, primarily using pre-made reports and queries. The type of data utilized in relation to student personal information is currently regulated by strict European and national guidelines (GDPR) [29].

Data on the MIS is consequently restricted in access. To fulfill the requirements for decision-making at each administrative level, the types of information that are made available are expressly mentioned and constrained. Data is typically restricted from usage for research purposes and only made available upon request from the research institution. This fact meets the legal requirements while imposing limitations on studies, as it did in this case.

Access to full, but limited, student data from "My School" was obtained for the sake of our current (and previous) research. We were provided with data on all pupils nationally, but not all the requested characteristics. It is the biggest set of student data that has ever been used in a study.

The student demographic information from the MIS used for the current study is summarized in table 1. The pseudo_id field is an anonymized field that doesn't exist in the MIS and is used to track student performance from year to year. Guardian's occupation was matched to the categories based on the International Standard Classification of Occupations (ISCO) categorization [33]. School's administrative region and gender are some of the other features provided.

TABLE I: Demographic Characteristics

Element	Type	Description
Pseudo_Id	text	Anonymous field (pseudo_id) for each student,
Educational Directorate	text	12 administrative services for primary and secondary education (50 addresses + (6 regions of Attica and 2 regions of Thessaloniki) × 2
Gender	Boolean	Boy/Girl
Occupation of guardian	text	Free text (to be completed by the school, in Greek)

The datasets included information about student accomplishments, such as the average grade for each student, the number of absentees from each class, and subject-specific marks. We were provided three years of data but without the matching class for each year. Thus, we had to run some queries (join inner, left, or right) on the primary school data to identify which class matched each year. In secondary school, it was simple to differentiate between courses because they were taught in different classes.

There were two datasets created. The first one contained students' data that started fifth grade in 2019 and graduated from their first year of high school in 2022. Students' data from the first high school class of 2019 who graduated from high school in 2022 is contained in the second dataset. As there are students who leave the Greek educational system because they move to another country, repeat a class for academic reasons, or don't attend enough classes, it is obvious that these datasets represent a subset of the whole data.

A. *Data Pre-processing*

Checking for completeness and integrity is important in order for the data to be suitable for processing. Various entries were found to be lacking grades in some courses due to the size of the dataset. The missing scores frequently have to do with subjects that aren't taught to particular student groups, like religious education, or subjects that aren't taught because of a lack of teachers. The results of several missing value imputation approaches were studied, without significantly different results. For this reason, the use of the average score of the remaining courses was selected as the simplest one. This approach was only applicable to a single missing value in a lesson. Following a strict strategy, when more than one lesson grade was absent, the entire instance was erased.

Other fields, without strict system checks, contained more incomplete data. The Guardian occupation field is a common occurrence of the phenomenon of missing data or "other" or "do not answer" entries.

The common “psevdo_id” values for the three successive years of each set were kept after the missing score-values were removed. In the first dataset, 79.561 records (72.91%) and 80,601 records (77.00%) of the original records were still present.

The first dataset examines students' the step from primary to secondary education, a critical stage in their school lives. The academic performance of each student in each class was evaluated using the x-means algorithm [30] in order to identify levels of students' achievement without predefining their number. The stability of students' performance through classes was also studied. A new variable connected with the students' performance category has been created. Three new variables have been added to each dataset. Additional analysis was conducted using these ordinal variables as a basis.

In particular, it examined:

a) the number of level-clusters for each class, and their relevant frequencies,

b) the continuity of learners' performance and the transition through levels,

c) the average and standard deviations of grades (GPA) per performance level,

d) the statistical significance of the differentiation of the average score (GPA) using non-parametric tests (Kruskal-Wallis), due to the lack of homogeneity in the variables,

e) the effects of the mentioned demographic characteristics on students' academic performance (These included the students’ gender, the guardian's occupation, and their place of residence.), and

f) the findings of the current analysis, which covers the three years 2019–2022, compared with those of our previous study, which covered the years 2015–2018.

B. *Clustering without Pre-determining the Number of Clusters*

Data is grouped into various clusters through the process of clustering, with the goal of achieving the highest level of similarity between data inside a cluster and the lowest amount of similarity between clusters. For many years, the industry standard for metric data was K-means [30]. Its simplicity and local-minimum convergence characteristics make it interesting. The cluster's centroid serves as the starting point for the group of clusters in the K-Means algorithm. The closest point to the starting cluster's center point determines where the center point of each group or cluster is.

The identification of the initial cluster center point, however, is the shortcoming of the K-Means approach in this case. This is a result of the lack of a strategy for selecting and detecting the cluster center point. The cluster center is arbitrarily chosen from a given set of data. The K-Means algorithm's clustering results are not perfect in every trial and are usually suboptimal. Therefore, the cluster's original centroid plays a role in both the positive and negative effects of clustering.

A method for determining the number of clusters was developed by Dan Pelleg and Andrew Moore of Carnegie Mellon University in Pittsburgh, Pennsylvania. The Bayesian Information Criterion (BIC) is used to calculate how many clusters to split the data into. However, MDL and AIC can also be utilized [28], [30]. The genuine value of K is therefore calculated using this method simply on the dataset and without any monitoring. Only the highest and minimum feasible X values must be determined during the initial stage of X-means grouping. The academic performance of the students was divided into distinct groups for each dataset using this approach. For further analysis, the generated achievement clusters were used as ordinal variables.

IV. Results

A. *First Research Question*

The first research question examined whether it was possible to group students' academic performance. A previous study demonstrated that it was possible to divide academic achievement into four distinct levels. It was discovered once more that there were four levels of achievement from the analysis of the present dataset, which included three school years. Clear evidence for student achievement was given by this grouping. We observe a notable variation in performance at the elementary and junior high school levels. In middle school, pupils in the high level ("A") performed on average better than 19 out of 20 and close to the top of the grade scale (10) in elementary school. In contrast, students in the low-performing category had GPAs below 8.5/10 in elementary school and close to 13/20 in high school. The relevant standard deviations are quite small. Students with intermediate achievement levels also have significant differences in their GPA.

Overall, we reach the conclusion that the occurrence of four levels in both primary and secondary education is not a coincidental discovery. For the remainder of this essay, we shall refer to these categories as very strong, strong, weak, and very weak.

Fig. 1 displays the percentage distribution of the performance levels' frequencies. In terms of the frequency of high achievers, there is initially a difference between educational levels, with the percentage falling from 55% in primary school to 30.6% in secondary school. The percentage of students in the low-achieving category has practically quadrupled (from 4.8% to 16.8%). Clear stability is shown when comparing the percentage frequencies between the various grades. In particular, when compared to the various averages, the standard deviations are low. The category of weak students in elementary school has a higher variance with a 0.6% mean and a 12.5% coefficient of variation.

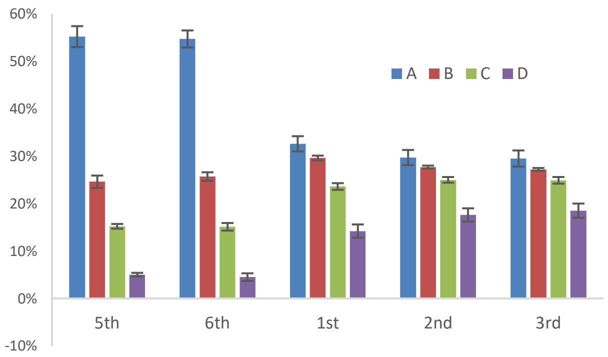

Fig. 1 Frequencies per achievement level.

The declining proportions of very strong students serve as a stark reminder of how challenging junior high school courses are and how challenging it is for pupils to adjust. It is also possible that it stems from the different methods of evaluation used in primary schools, which take into account traits like effort, initiative, creativity, peer cooperation, etc., in addition to performance. Last but not least, we notice that the group of students that perform very poorly shows an increase when they move from one grade to the next. This indicates how more and more pupils are falling behind as the years pass.

Overall, the findings are in line with our prior research's conclusions on the consistency of grouping student performance across levels and our observations about the gradual decline in all learners' performance as the difficulty rises from grade to grade.

B. *Second Research Question*

1) *First dataset*

The second research question focused on the three-year longitudinal study of student achievement. Only data that was consistent across the three subsequent school years was kept in the datasets. Studying the grouping of pupils according to their academic attainment levels was made possible in this way. The first dataset, which covers grades from fifth grade through first grade of high school, covered students' transition from elementary to secondary education. The second dataset studied the changes in students' academic performance throughout the course of the three junior high school grades.

Table II shows the variation in performance levels in primary schools. The high-achieving category of students in the sixth grade of primary school was shown to vary. In the sixth grade, 53.44% of the highly competent fifth graders continue at the same level of achievement, while 45.36% now fall into the second level. A transition from level to level corresponds to a small difference in averages between "A" and "B" levels (9.97 and 9.59 out of ten). Students who earned a "B" in fifth grade also demonstrated similar changes, with a 43.66% increase in their level. However, the small percentage of students (3.56%) who fall to very weak level is still present. Similar to the fifth grade, many weak children either perform at the same level or improve by 32.94% to achieve a "C" in sixth grade.

TABLE II: CLUSTER FREQUENCIES OVER TIME—FIRST DATASET (A)

6thclass	5th class level*			
level	A	B	C	D
A	53.44%	43.66%	18.03%	3.56%
B	45.36%	40.25%	19.38%	3.90%
C	1.13%	14.94%	50.42%	32.94%
D	0.06%	1.15%	12.16%	59.59%

Comparing achievement levels between the fifth grade of primary and first grade of junior high school demonstrates the lack of volatility in achievement. In the first grade of junior high school, 68.19% of the very strong kids retain their characteristics, and just 0.10% fall to the level of extremely weak children. In contrast, only 0.43% of very weak pupils are considered to be very strong, with 76.67% remaining at the same level.

TABLE III: CLUSTER FREQUENCIES OVER TIME—FIRST DATASET (B)

1stclass	5thclass HS level*			
HS level	A	B	C	D
A	68.19%	21.55%	4.82%	0.43%
B	28.43%	56.33%	39.85%	6.57%
C	3.28%	21.84%	54.41%	16.03%
D	0.10%	0.28%	0.92%	76.97%

Increased variability is evident among students with average achievement (B and C), as was also found in our earlier study. Of the students who received a "B" in fifth grade, 56.33% continue to perform at the same level in first grade, although close to 21% are classified as "very strong" or "weak". The performance levels of the weak students (level C) improved by 54.41 percent, while 54.41% remained at the same level.

It is also found that there are significant differences between high and extremely poor-performing students in the distribution of GPAs between the sixth and first grades. Additionally, the statistical significance of the variation in average grades was investigated. Tables IV and V display the Kruskal-Wallis statistics for each level of achievement. The statistical significance for all possible combinations of performance levels tends to zero.

TABLE IV: GPA, BASED ON INITIAL CHARACTERIZATION – FIRST DATASET (A)

	6th Class GPA ES			
5th Class PS Rank	Mean	Std. Deviation	Median	Kruskal-Wallis statistic
A	9.9624	0.061	10.0	230.84
B	9.6482	0.1299	9.6	-94.95
C	9.1727	0.1912	9.2	-149.21
D	8.3991	0.4094	8.5	-96.44

TABLE V: GPA, BASED ON INITIAL CHARACTERIZATION – FIRST DATASET (B)

	1st Class GPA HS			
5th Class PS Rank	Mean	Std. Deviation	Median	Kruskal-Wallis statistic
A	17.7056	1.5901	18.0	163.84
B	15.6408	1.8482	15.7	-63.13
C	13.9935	1.9009	13.9	-109.36
D	12.5261	1.8403	12.5	-73.81

DOI: http://dx.doi.org/10.24018/ejeng.2022.1.CIE.2958

2) *Second dataset*

In the second dataset, the same analytical approach was used. The key finding is that there is significant stability in the high school between very strong and extremely weak students. In junior high school, 77.31% of very strong students continue at the same level in the next grade, and in the third grade, that number rises to 87.63%. On the other hand, very few students fall into the category of extremely weak pupils. When very weak students are included, the tendency is reversed; in the secondary school, 75.50% stay at the same level, and 56.77% in the third class. Students who have an average performance move from one level to the next without following a certain pattern.

TABLE VI: CLUSTER FREQUENCIES OVER TIME—SECOND DATASET

		A	B	C	D
A	2nd class	77.31%	13.62%	0.91%	0.20%
	3rd class	87.63%	25.78%	2.25%	0.18%
B	2nd class	21.07%	56.48%	16.88%	2.16%
	3rd class	11.79%	61.18%	34.24%	4.06%
C	2nd class	1.54%	28.00%	59.36%	22.13%
	3rd class	0.52%	12.47%	54.75%	38.99%
D	2nd class	0.07%	1.90%	22.85%	75.50%
	3rd class	0.05%	0.56%	8.76%	56.77%

The GPA based on the initial grouping of students is analogous to the first dataset's depiction of GPA differentiation. We notice that the mean and median GPAs for very strong students are 18.899 and 19.1, respectively, whereas the corresponding values for third grade students are 18.274 and 18.5. The very weak first-graders, in contrast, had average GPAs of 13.327 in the second grade and 12.233 in the third. A Kruskal-Wallis test was used to determine the statistical significance of the variation in total grades for each performance level, and it was discovered that the p-values tend to zero for all combinations of performance levels.

TABLE VII: GPA, BASED ON INITIAL CHARACTERIZATION – SECOND DATASET (A)

1st Class	2nd Class HS			
HS Rank	Mean	StDev	Median	Kruskal-Wallis statistic
A	18.899	0.943	19.1	207.82
B	16.937	1.27	17.0	-11.66
C	15.073	1.347	15.1	-132.27
D	13.327	1.273	13.2	-129.55

TABLE VIII: GPA, BASED ON INITIAL CHARACTERIZATION – SECOND DATASET (B)

1st Class	3rd Class HS			
HS Rank	Mean	StDev	Median	Kruskal-Wallis statistic
A	18.274	1.28	18.5	200.76
B	15.959	1.538	16.1	-10.25
C	13.944	1.461	13.8	-128.58
D	12.233	1.321	12.1	-126.32

C. *Third Research Question*

The academic community has been concerned with the connection between students' socioeconomic status and academic performance for decades [2], [3]. It has been suggested that the educational system serves as a breeding ground for social injustices [3]. On the other hand, the Greek constitution states that "education is a basic state mission and aims at the moral, spiritual, professional, and physical education of the Greeks." Based on this viewpoint, the education system is a structure that aids students by offering them equal chances. However, because every student has a unique social history, it is impossible to anticipate that the non-individualized method of offering the same educational services will produce a balanced result. As a result, student demographics might not be unrelated to achievement, supporting Bourdieu's theory to some extent [3].

Guardian occupation, gender, and residence area were linked to socioeconomic profile in the data available to the researchers. Due to the General Data Protection Regulation's (GDPR) rules, they could only offer a few demographic factors.

1) *Profession of Guardians*

The field related to the guardian's occupation in the MIS "My school" is free text. As a result, the data needed to be categorized using a reliable classification. The International Standard Classification of Occupations (ISCO) was utilized for the classification [31]. After the occupations were classified according to the ISCO classification and missing values were removed, an χ^2 statistical test was carried out to detect statistically significant variations in student performance levels according to their respective caregiver occupational categories. The results of the statistical test revealed statistically significant differences, with the statistics' values and p-values being displayed in Table VIII.

TABLE IX: X^2 STATISTICS (OCCUPATION ILO)

	Class	X^2	p-values
1st dataset	5th ES	2.748	0.0000
	6th ES	1.911	0.0000
	1st HS	3.029	0.0000
2nd dataset	1st HS	4.057	0.0000
	2nd HS	3.724	0.0000
	3rd HS	3.938	0.0000

Students with guardians from the "professionals" category clearly do better in both datasets. In the first dataset, the average percentage of professionals in the high-performing category is 9% (st. dev 4%), whereas in the second dataset, it is 11% (st. dev 1%). The outcomes are in line with the average of 20.72% from our prior survey. In junior high school, the overperformance rate is lower. In both datasets, students whose guardians worked as salespeople or service providers came in second (mean = 7%, st. dev = 4% and mean = 9%, st. dev = 1%, respectively). Students with guardians who fell into the categories of being unemployed, skilled or unskilled, and manual workers made up the majority of those with low performance at the other end of the performance spectrum. For the categories of skilled and unskilled workers, the rates are lower in primary school, where they range from 3% to 5%, and rise in secondary school.

These findings are consistent with the earlier study we conducted [1]. In this way, it appears that our idealistic hypothesis that equal learning chances would be provided is erroneous. Similar to what we found in our earlier research, we found that having guardians with the same professions is linked to better performance, while having guardians with the opposite professions is linked to very poor performance.

DOI: http://dx.doi.org/10.24018/ejeng.2022.1.CIE.2958

2) Gender

We had the chance to check the results on differentiation using the χ^2 test between the very high and very low achievement levels based on the students' gender. The test applied to both datasets gave similar outcomes. Statistically significant differences showed up in each case. Girls exhibit a higher frequency of occurrence at the high-performance level and a lower frequency of occurrence at the low-performance level. In particular, for the high achievement level in the first dataset, the average percentage difference between actual and theoretical frequency in favor of girls was 5.43%, and for the second dataset, it reached 8.13%. On the other hand, girls show a significant difference in the very low performance category, where they show up 5.7% less often than expected.

In our earlier study [1], it was also noted that girls performed better than boys (8.15%). It is a rather constant situation in the Greek educational system. Using information from the Program for International Student Assessment (PISA), the phenomenon of differentiation in student performance in Greece has also been observed [32]. Academic success gaps between girls and boys provide an opportunity for research on the qualities of girls or boys that contribute to this divergence. Further insight into the numerous aspects of this issue can be gained by closely examining the nature of gender inequalities in academic achievement between distinct social groupings. In earlier research [33], it was shown that "social class and parental education are still the most reliable indicators of a child's educational achievement" and that "in general, boys' underachievement in school is strongly class and race-based." Some more ideas for the issue connect boys' underachievement to cultural norms of masculinity and femininity and to male school culture. In this study, the results are clear from the data alone, which backs up all of the above and eliminates the need to make a model of possible factors that affect behavior.

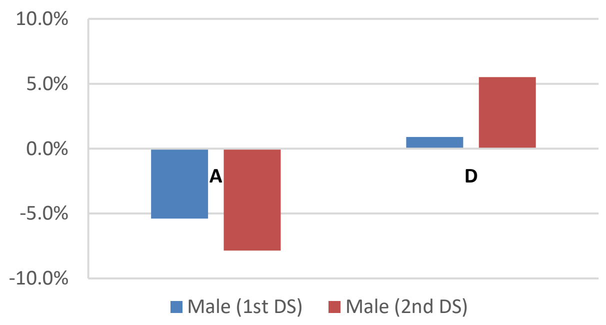

Fig. 2 Male differences High and Low achievers.

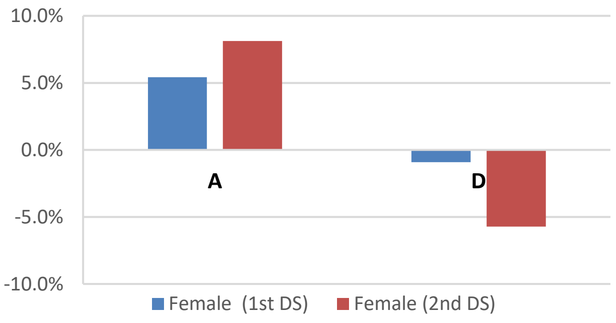

Fig. 3 Female differences High and Low achievers.

3) Region

Following the extraction of the performance levels, any discrepancies were examined using the region of residence data. For primary and secondary education, the field generally refers to the county or metropolitan area of residence. Like in our earlier survey (χ^2= 174.000, p-value= 0.42), statistically significant differences were discovered after the χ^2 test (χ^2= 4056.810, p-value=0.00). Consequently, in both primary and secondary education, there were areas with a higher incidence of very high performance and areas with a higher incidence of very low performance. The following tables show the top and bottom six districts in the category of very strong and very weak students.

TABLE X: TOP AND BOTTOM 6 REGIONS (FIRST DATASET)

	Region	A (ES)	D (ES)
Top 6	Chios	14.58%	-3.53%
	Xanthi	13.95%	-2.41%
	Rodopi	13.36%	-4.54%
	Karditsa	12.37%	-3.99%
	Grevena	11.15%	-4.27%
	Larissa	10.17%	-5.61%
	….		
Bottom 6	Evritania	-12.52%	5.87%
	Heraklion	-3.80%	5.96%
	Halkidiki	-6.08%	7.62%
	West Attica	-8.87%	8.12%
	Lasithi	-3.37%	8.54%
	Rethimno	-9.59%	8.55%

TABLE XI: TOP AND BOTTOM 6 REGIONS (SECOND DATASET)

	Region	A (HS)	D (HS)
Top 6	Chios	17.54%	-1.53%
	Fokida	9.03%	-0.18%
	Arkadia	7.96%	0.29%
	Athens (2nd area)	7.89%	-1.24%
	Evritania	7.79%	-0.10%
	Lesvos	6.95%	-0.10%
	…..		
Bottom 6	West Attica	-2.06%	1.48%
	Lakonia	3.57%	1.51%
	Imathia	-5.19%	1.56%
	Messinia	-2.71%	1.92%
	Karditsa	3.91%	2.24%
	Lasithi	-6.26%	3.30%

In both education levels Chios region comes out on top. Chios is an island in the eastern Aegean that looks to be at the top of elementary and secondary education, although it barely accounts for 0.48% of students in our dataset. The placement of West Attica in the bottom 6 for both levels is a common finding between this and the prior survey. There are numerous minorities in this area. In fact, we have seen a decline in the percentage of students who are overrepresented in this area from 8.12% in elementary school to 1.48% in high school. We can only assume that this is more due to the student attrition of students from this social group, which has been identified and has been the subject of extensive educational interventions, as we did not have data available to us detailing the socioeconomic status and student attrition of Roma students.

The percentage frequencies of representation for the very high and very low performance categories were also compared, and the correlation between them was examined. The table below shows that there is a significant negative correlation between the frequencies of districts with more extremely good students and the corresponding percentage

DOI: http://dx.doi.org/10.24018/ejeng.2022.1.CIE.2958

of weak students in the same districts (r = -0.7521, p-value = 0.000). A positive correlation was also found between districts with low performance in primary and secondary education. This means that districts with low performance in primary are likely to have low performance in secondary education as well (r = 0.5225, p = 0.000).

TABLE XII: CORRELATIONS BETWEEN HIGH AND LOW ACHIEVER' PERCENTAGES

	A (ES)	D (ES)	A (HS)	D (HS)
A (ES)				
D (ES)	-0.7521*			
A (HS)	0.204856	-0.26827		
D (HS)	-0.31207	0.5225*	-0.20853	

D. Forth Research Question

By comparing the findings of the two studies, it is feasible to draw some broad conclusions regarding the problem of student achievement and how demographic factors affect it. The four levels of academic success distinctions for primary and secondary schools were initially found to be stable in the first study [1]. In both time periods examined, the same number of clusters that met the BIC criterion was obtained. The results about performance are common, despite the fact that the two middle performance groups essentially overlap to some extent. Additionally, the GPA for each performance level in both surveys maintains the same values. Since the numerical achievement scale was first used in the fifth grade, there has been a clear picture of very strong and very weak students. A very strong student has a very high probability of remaining at the same level and very low probability of dropping to a lower level. As a result, there are particular types of students who can manage the course's demands rather easily. The opposite picture is also extremely obvious. Very weak students are likely to maintain their low performance levels in both surveys. On the other hand, the likelihood of succeeding at some point in the future is little to none. Unfortunately, it is a well-known "doom" that these children will do poorly in school, which often leads to failing classes and having to start over.

The overlap between "B" and "C" performance categories is combined with a gradual decline in performance as the difficulty increases, especially in secondary school. Our findings highlight a significant challenge that has to be addressed, namely consistently raising students' cognitive levels and, consequently, achievement, to the extent that student achievement is an indicator of the effectiveness of the educational system.

The occupation of guardians is a criterion for differentiating performance. Both high and low achievement categories confirmed the overrepresentation of students with particular guardians' or parents' occupations. Jobs requiring manual labor have a higher prevalence of low performance levels. The same is true for students whose parents are unemployed. In contrast, children of parents who fall under the "professional" category (such as doctors, lawyers, engineers, government employees, teachers, and others) frequently achieve at the highest levels. The finding is really alarming to the extent that occupation reveals socioeconomic status. Children from a certain social setting struggle to rise above an "invisible ceiling" of achievement. Theories have offered a variety of explanations for the phenomenon, including the diminished value placed on knowledge or the inability to support pupils who come from poor families [33]. Unfortunately, the researchers didn't have enough demographic data to be able to look at the problem in more depth, so it can only be brought up in passing.

It was confirmed that performance differed according to gender. Female students did better than male classmates in all grades and years. Numerous investigations conducted worldwide support this conclusion. The location of a residence is a factor that can differentiate performance. However, there is no consensus across research in certain regions. In both studies, West Attica is the only place where the rate of very low achievement is higher, especially in primary education.

V. CONCLUSIONS

Conclusions about the use of data mining in educational data, the efficacy of educational policy, and the study field in general were made based on research findings. The value of using unsupervised methods to analyze educational data was confirmed. Analysts can conduct analysis without making prior assumptions thanks to a variety of clustering techniques and the use of optimization for algorithms' parameters. In this approach, the data "speaks" without the prior implementation of a theory. In fact, the opposite approach is taken, lowering the likelihood of researcher bias. In this study, data provided support for a theory, such that of the school acting as a reproductive social mechanism [3].

The outcomes of the analysis and any new variable can be used for additional research using statistical or non-statistical methods. The results of the current study were combined with common statistical methods like the χ^2 and Kruskal-Wallis tests. Alternative methods include categorization, regression, and clustering. Additionally, there are currently many algorithms and cost-free programming tools that can be used. In general, we found that the DM tools were easy to use when testing the study hypotheses in the context set by educational research.

Through the perspective of the decision-making and policy implementation processes, we found that Academic Analytics can be especially useful. Specific issues with education policy were highlighted. Students' achievement levels have remained stable over time, highlighting a somewhat static educational framework that does not really benefit weak students. The issue of fair opportunities in education is brought into the spotlight by the non-independence of academic achievement from factors like the guardian's profession and place of residence. Because education in this country is a responsibility of the central government, there are centralized and uniform policies that are put into place, yet these policies don't seem to be able to discriminate against and help weaker students. The principle of equality vs. equal opportunities for all is used, and more ambitious and targeted policies need to be made and put into place.

The move from primary to secondary education was also associated with a significant decrease in the percentage of achievers. This has shown a distinct, more impersonal evaluation system for secondary school students, based solely on performance. The students' gradual drop in grades shows that they are getting less and less able to handle harder subjects.

The need for a thorough study into the factors that led to these conclusions is made clear by an appraisal of the research findings. Additional research is crucial for the development of tailored policies, and it can be supported by both secondary data from the MIS and, when necessary, primary data. In our opinion, an in-depth study of the factors that contribute to the academic achievement gaps between boys and girls as well as the differences between excellent and extremely weak students is important. Both studies call for an increase in students' characteristics, which were not provided for this particular research. Examining learning styles, socioeconomic factors, and the value placed on education can shed light on the causes of the phenomenon.

The need for a framework for providing educational data to researchers was found. We faced an ad hock application to the General Data Protection Regulation both times that we requested data from the Ministry of Education. A far wider range of demographic characteristics might be offered by creating a framework and anonymizing certain characteristics beyond the student ids, including city of residence and other characteristics. The Ministry of Education was in charge of and took care of all the indexes, so the results of the analyses could be compared to real characteristics and specific conclusions could be made.

From a research point of view, this study confirmed without a doubt the results of our previous study. This shows that the results are not random and are based on a stable situation in the Greek educational system. As long-term data is collected in the information system, it is expected that a long-term study of students' performance will be possible.

Acknowledgment

The authors would like to thank the central services of the Greek Ministry of Education for responding quickly to their request for information about the number of students in the country as a whole.

Funding

This study was not financed. The Greek Ministry of Education provided the data. The analysis was carried by using open source software.

Conflict of Interest

Authors declare that they do not have any conflict of interest.

References

[1] Papadogiannis I, Wallace M, Poulopoulos V, Karountzou G, Ekonomopoulos D. A First Ever Look into Greece’s Vast Educational Data: Interesting Findings and Policy Implications. *Education Sciences*. 2021;11(9):489.

[2] Coleman, J. Equality of Educational Opportunity; U.S. Department of Health, Education, and Welfare. U.S. *Government Printing Office: Washington, DC*.1966.

[3] Bourdieu P. Cultural Reproduction and Social Reproduction. In. *Power and Ideology in Education*. J. Karabel, & A. H. Halsey. Oxford University Press. 1977, pp. 487–511.

[4] Islam S; Baharun H, Muali C,Ghufron I; Bali I. Wijaya M, Marzuki, I. To Boost Students' Motivation and Achievement through Blended Learning. *J. Phys. Conf. Ser*. 2018,1114, 012046.

[5] Ozen O, The Effect of Motivation on Student Achievement. In *The Factors Effecting Student Achievement*. 2017.Springer: Cham, Switzerland, pp. 35–56.

[6] Marsh W; Pekrun R, Murayama K, Arens K, Parker. D; Guo J,Dicke, T. An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Dev. Psychol*. 2018;54:263–280.

[7] Lai L, Hwang J. A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Comput. Educ*. 2016;100:126–140.

[8] Cvencek D, Fryberg A, Covarrubias R, Meltzoff. N. Self-Concepts, Self-Esteem, and Academic Achievement of Minority and Majority North American Elementary School Children. *Child Dev*. 2018;89:1099–1109.

[9] Yang Q, Tian L, Huebner S, Zhu X. Relations among academic achievement, self-esteem, and subjective well-being in school among Elementary school students: A longitudinal mediation model. *Sch. Psychol*. 2019;34:328–340.

[10] Geller J. Toftness. R,Armstrong I, Carpenter S. K, Manz, C.L, Coffman C.R.; Lamm M.H. Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory* 2018;26:683–690.

[11] Day C, Gu Q, Sammons P. The Impact of Leadership on Student Outcomes. *Educational. Administration.* 2016;52: 221–258.

[12] Ohlson, M.; Swanson, A.; Adams-Manning, A.; Byrd, A. A Culture of Success—Examining School Culture and Student Outcomes via a Performance Framework. J. Educ. Learn. 2016;5:114.

[13] Konold T, Cornell D, Jia Y, Malone M. School Climate, Student Engagement, and Academic Achievement: A Latent Variable, Multilevel Multi-Informant Examination. *AERA Open* 2018;4:233285841881566.

[14] de Boer H. Timmermans A.C, van der Werf C. The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis. *Educ. Res. Eval*. 2018;24:180–200.

[15] Sebastian J, Moon M, Cunningham M. The relationship of school-based parental involvement with student achievement: A comparison of principal and parent survey reports from PISA 2012. *Educ. Stud*. 2017;43:123–146.

[16] Karadag E. *The Factors Effecting Student Achievement—Meta-Analysis of Empirical Studies.* Springer: Cham, Switzerland, 2017.

[17] Baker R. S, Yacef K. The state of educational data mining in 2009: A review and future visions. *JEDM Educ. Data Min*. 2009;1:3–17.

[18] Romero C, Ventura S. Educational Data Mining: A Review of the State of the Art. *IEEE Trans,* 2010;40:601–618.

[19] Papamitsiou Z, Economides A. Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educ. Technol*. Soc. 2014;17:49–64.

[20] Baker R. S, Inventado PS. Educational Data Mining and Learning Analytics. In *Learning Analytics*; Larusson J, White B.Eds.; Springer: New York, NY, USA, 2014:61–75.

[21] Dutt A, Ismail MA, Herawan T. A. Systematic Review on Educational Data Mining. *IEEE Access* 2017;5:15991–16005.

[22] Papadogiannis I, Poulopoulos V, Wallace M. A Critical Review of Data Mining for Education: What has been done, what has been learnt and what remains to be seen. *Int. J. Educ. Res. Rev*. 2020;5:353–372.

[23] Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2020;10(3):e1355.

[24] Prieto L P, Sharma K, Dillenbourg P. Studying teacher orchestration load in technology-enhanced classrooms, In *Design for Teaching and Learning in a Networked World, ser. Lecture Notes in Computer Science,* G. Conole, T. Klobuar, C. Rensing, J. Konert, and E. Lavou, Eds. Springer International Publishing, 2015;9307:268–281.

[25] Daniel Ben Kei. Big Data and data science: A critical review of issues

DOI: http://dx.doi.org/10.24018/ejeng.2022.1.CIE.2958

for educational research. *British Journal of Educational Technology*, 2019;50.1:101–113.

[26] Romero C, Ventura S. Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2017;7.1: e1187.

[27] Lang C, Siemens G, Wise A, Gasevic, D. The Handbook of Learning Analytics; *Society for Learning Analytics Research (SoLAR)*: Ann Arbor, MI, USA, 2017.

[28] Pelleg D, Moore A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 15–18 August 1999; Association for Computing Machinery: San Diego, CA, USA, 1999;1:727–734.

[29] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L 119, 2016, 1–88.

[30] Duda R.O, Hart P.E, Stork D.G. *Pattern Classification,* 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2001.

[31] ILO. *International Standard Classification of Occupations 2008: ISCO-08.* International Labour Office: Geneve, Switzerland, 2012.

[32] Domnech - Betoret F, Abelln R. Self-Efficacy, Satisfaction, and Academic Achievement: The Mediator Role of Students' Expectancy-Value Beliefs. *Front. Psychol*. 2017;8:1193.

[33] Rezaeinejad M, Azizifar A, Gowhary H. The Study of Learning Styles and its Relationship with Educational Achievement among Iranian High School Students. *Procedia-Soc. Behav. Sci*. 2015;199:218–224.

Ilias Papadogiannis is born in Tripolis in 1975. He holds two BA’s, a BA in Management from Athens University of Economics and Business (Management Dept), Athens, Greece, 2007, a BA in Accounting from The Technological Institution (TEI) of Patras (Accounting Dept), Patras, Greece, 1996, and a postgraduate degree (Msc) in Governance from the University of Peloponnese, Tripolis, Greece, 2015.He is a PHD Candidate in the field of Educational Data Mining in the University of Peloponnese, Tripolis, Greece.

He has worked in the financial sector as security broker for Eurobank Security (2000-2005). He has 15 years of experience in administrative positions in educational fields. He is currently HEAD OF THE ICT DEPARTMENT of the Regional Education Directorate of Peloponnese, which resides in Tripolis, Greece and an Erasmus+ promoter in the region of Peloponnese. He has also participated as a lecturer in seminars concerning educational issues and he has published in educational conferences. Some of his previous publications include: Digital citizenship in Greek Primary schools in Peloponnese, Citizenship Education: a problematic concept or a myth, Warsow, Poland, CiCeA, 2018. His main scientific interest is Data Mining, Statistics and Research Methodology, and he has participated in several educational research projects.

Manolis Wallace was born in Athens in 1977. In 2001 he received a diploma in electrical and computer engineering and in 2005 a PhD in intelligent knowledge-based systems in uncertain environments, both from NTUA's School of Electrical and Computer Engineering. Since 2007 he is a faculty member at the Department of Informatics and Telecommunications of the University of Peloponnese, while at the same time and up to 2013 also a senior researcher at the Foundation of the Hellenic World. Before that, He was at the Athens Campus of the University of Indianapolis, where I served as the chair of the Department of Computer Science. His research interests lie in the meeting of computing and humans, specifically in areas such as cultural informatics, educational informatics, smart cities, personalization and so on, and since 2002 He have (co-) authored around 150 papers in these fields. Most of them can be found at Google Scholar. He serves, or have served in the past, as associate or guest editor in numerous journals and as general, local, program committee or publicity chair in numerous conferences.

Vasileios Poulopoulos was born in Kalamata in 1982 received his diploma from the Computer Engineer and Informatics Department of the University of Patras in 2005. He obtained his MSc and PhD from the same department in 2007 and 2010 accordingly in data mining and analysis from heterogenous sources of the web and especially big data Recently elected as an Assistant Professor at the department of the Digital Systems of the University of Peloponnese, while being amember of Knowledge and Uncertainty Research Lab of the University of Peloponnese from 2017. In 2019 I completed my post-doc performing research on the role of Big Data in Cultural Informatics, continuing the research on it. He has worked for CTI-DIOPHANTUS (Research Institute) on several EU projects from 2002-2010 when he decided to turn to the private sector and especially decided to follow the “Greek start-up wave”. Being a founding member of Hellenic Start-up Association and working for several projects, companies and start-ups from 2010 until 2015 a year that found me back to the University classes teaching for the Technological Educational Institute of Peloponnese. His research interests include among others: data mining, knowledge extraction, big data, cultural heritage, personalization techniques, clustering techniques, categorization techniques as well as innovative web and mobile applications that could make our everyday life easier andbetter. He has published more than 50 papers in the aforementioned fields. Detailed information about the publications can be found in Google Scholar.