Exploration of school orientation using Principal Component Analysis (PCA): An Analytical Approach.

Noaman LAKCHOUCH
Sciences, Technology and Innovation (STI)
Abdelmalek essaadi University
Tetouan, Morocco
https://orcid.org/0009-0009-9523-5569

Lamarti Sefian MOHAMED
Sciences, Technology and Innovation (STI)
Abdelmalek essaadi University
Tetouan, Morocco
https://orcid.org/0000-0001-8270-2660

Mustapha KHALFOUNI
Sciences, Technology and Innovation (STI)
Abdelmalek essaadi University
Tetouan, Morocco
https://orcid.org/0009-0005-5589-1714

***Abstract*—This study used Principal Component Analysis (PCA) to examine the structure of academic performance across a range of subjects in a sample of students. By transforming the dataset into a set of orthogonal principal components, we aimed to capture the maximum variance in student grades with minimal loss of information. PCA revealed that the first principal component (PC1) accounted for a substantial proportion of the variance, suggesting a general factor of academic performance. Interestingly, PC2 appeared to differentiate between scientific and literary profiles, indicating a dichotomy in skill sets. Subsequent components provided further granularity in the multidimensional nature of academic ability. Visualization through biplots offered an intuitive overview of student scores and subject contributions, highlighting correlations and potential areas for educational interventions. The analysis suggests that PCA is a robust tool not only for dimensionality reduction but also for uncovering latent patterns in educational data, which can guide pedagogical strategies and inform educational counseling. This article discusses the implications of these findings for adapting pedagogical approaches and assisting with student academic and career guidance.**

***Index Terms*—Principal component analysis, Academic performance, Educational data, global visualization, Multidisciplinary education, Pedagogical technique in guidance.**

I. Introduction

School orientation is crucial in helping students choose paths that align with their interests and skills. Typically, this choice is guided by individual discussions, tests, and counselors' experience. However, today, data analysis offers a new way to enhance our understanding of students' needs and personalize their support.

Principal Component Analysis (PCA), a data analysis method, plays a key role here. It simplifies complex data by focusing on the most important information, enabling the discovery of groups of students with similar interests and abilities. Through PCA, it is possible to identify different student profiles - such as those who prefer sciences, arts and humanities, or those who have not yet decided - and provide more tailored guidance to each group. In short, PCA enriches the orientation process by providing a clearer view of how students can be guided towards paths that suit them best.

PCA works by transforming an initial set of possibly correlated variables into a new set of uncorrelated variables, called principal components. These components are ordered so that the first one captures the largest possible variance, and each subsequent component captures the maximum remaining variance under the constraint of being orthogonal to the previous components [1]. In the context of school orientation, this allows for the identification of the main axes along which students' preferences and choices vary the most, thus offering a simplified yet informative view of the data.

Studies using PCA in educational guidance, for example, have been able to highlight groups of students sharing similar interests or motivations, or key factors influencing career decisions, such as academic performance, professional interests, and social or familial influences ([2]; [3]). By reducing the complexity of data, PCA enables researchers and practitioners in guidance to better understand the dynamics at play and develop more targeted and effective interventions.

PCA can also be used in combination with other statistical methods or data mining to refine the analysis and explore relationships between variables in more detail. For instance, after identifying principal components, cluster analyses can be applied to group students into distinct profiles based on their scores on these components [4]. In this analysis, we will focus on the following central question: How can Principal Component Analysis (PCA) be used to identify student profiles? Our goal is to achieve the following targets:

- Identify groups of students sharing common profiles.

- Distinguish groups of students with unique and distinct profiles.
- Recognize groups of students located in areas of confusion.

II. Literature Review

Educational guidance is an important research area in the field of education, as it plays a crucial role in students' development and academic success. Principal Component Analysis (PCA) is a statistical method that allows for the analysis and visualization of relationships between different variables. In the context of educational guidance, PCA can be used to explore students' profiles and career choices and to assist in making informed decisions regarding the most common educational paths.

A recent study conducted by [5] used PCA to analyze data on high school students' career choices. The results showed that students were influenced by several factors, such as career interests, academic abilities, and personal values. The authors also identified various career profiles, such as students oriented towards sciences, students oriented towards arts, and students oriented towards vocational fields.

Another study conducted by [6] used PCA to explore the relationships between academic performance, cognitive abilities, and career choices of college students. The results indicated that students with good academic performance were more likely to choose academic pathways, while students with lower cognitive abilities were more likely to opt for vocational pathways.

Lastly, a meta-analysis study conducted by [7] examined the effectiveness of PCA-based educational guidance interventions. The results demonstrated that these interventions were effective in enhancing students' career choices and reducing dropout rates.

In conclusion, PCA is a powerful method for exploring educational guidance and identifying factors that influence students' choices. The aforementioned studies show that PCA can be successfully utilized to enhance educational guidance practices and assist students in making informed decisions about their academic and professional futures.

III. Method and materials

A. Principal Component Analysis (PCA)

PCA is a statistical technique for dimensionality reduction widely used to simplify the complexity of data sets while retaining as much of their original information as possible. By transforming a large number of possibly correlated variables into a smaller number of linearly independent variables called principal components, PCA facilitates the analysis and interpretation of data.

PCA is based on solid mathematical foundations, primarily using Singular Value Decomposition (SVD) or the diagonalization of the covariance (or correlation) matrix of the data. These methods help identify the directions (or principal axes) in the data space that maximize the variance of the data projected onto these axes. The eigenvectors of the covariance matrix represent these directions, while the associated eigenvalues indicate the amount of variance captured by each component.

PCA is widely used in various fields for exploratory data analysis, visualization of multidimensional data, dimensionality reduction for data preprocessing before applying machine learning techniques, and identifying patterns or hidden structures in the data. It is particularly.

B. Data Analysis

Our analysis focuses on exploring the academic performance of students in various subjects, with the aim of simplifying and understanding the vast data contained in an educational database. To achieve this, we plan to classify students into two main orientations: scientific and literary. This classification can serve as a basis for more in-depth analyses, such as identifying trends, strengths, and areas for improvement for each student.

C. Database

The loaded database includes the grades of several students (identified by ID.Stud) in different subjects: History-Geography (H.G), Islamic Education (I.ISLAM), English (ENG), Arabic (AR), French (FR), Mathematics (MATH), Physics-Chemistry (P.C), and Life and Earth Sciences (SVT). These data are real and concern the exam grades of the 3rd middle school in a district of public education. After removing rows with missing values, the database now contains $9,784$ entries. Let's move on to an exploratory analysis to examine the descriptive statistics of grades by subject. Here is a summary of the descriptive statistics for the grades in different subjects: Table of Descriptive Statistics

TABLE I
SUMMARY OF DESCRIPTIVE STATISTICS FOR EACH SUBJECT.

Statistic	Min	Mean	Median	sd[1]	Max
History.G	0.28	12.09	12.67	3.32	20.00
Islamic.E	0.83	12.81	12.94	3.58	20.00
English	0.22	13.93	14.11	3.51	20.00
Arabic	0.75	12.83	12.86	3.89	20.00
French	0.00	12.25	12.23	3.61	20.00
Maths	0.00	10.28	09.97	3.72	20.00
Physics-C	0.00	12.11	11.97	3.52	20.00
Life-Earth.S	0.42	12.61	11.50	3.65	20.00

[1] standard deviation.

1) Global visualization: The graphs n°1 present the distribution of grades by subject in the form of box plots after treating outliers and extreme values such as 20.00, which highlight the median, quartiles, and any outliers for each subject. Here are some key observations:

- All subjects show variability in scores, with a relatively symmetrical distribution around the median for most subjects.
- The distributions appear relatively symmetrical for most subjects, with medians close to the center of the boxes.

- Subjects such as Mathematics and Physics-Chemistry show a higher number of low outliers, which may indicate a particular difficulty in these areas for some students.
- English appears to have a slightly more spread out distribution towards higher grades, with a denser concentration of high outliers, suggesting exceptionally good performance for a group of students.

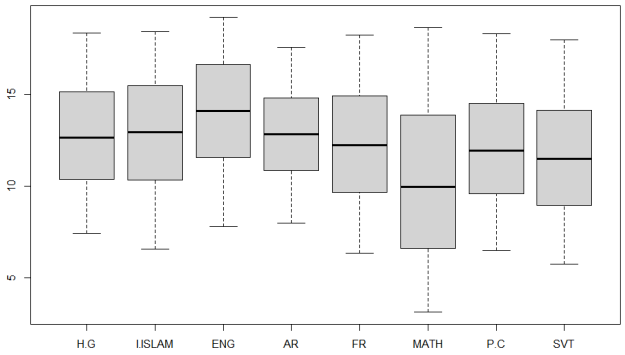

Fig. 1. Comparison of Student Grade Distributions Across Subjects

2) correlation matrix: The graph n°2 represents the correlation matrix between the grades obtained in different subjects, providing an insight into possible relationships between them. Here are some points to note:

- Most subjects show a positive correlation between them, suggesting that students who achieve good grades in one subject tend to perform well in other subjects as well.
- However, the correlation values vary in intensity, with some subjects being more strongly correlated than others. For example, grades in Mathematics and Physics-Chemistry (P.C) tend to be more strongly correlated, which could reflect the similarity in the skills required for these two subjects.
- Some subjects, although positively correlated, show relatively weaker correlation coefficients, indicating that the relationship between their performances is not as strong. This could be due to the difference in the nature of the skills assessed by these subjects.

These insights can be valuable in understanding how students' performances in one subject can predict their outcomes in other areas.

D. Empirical analysis of principal components.

1) Explained variance: From plot n°3, it can be observed that the first principal components significantly contribute to the explained variance, with a notable decrease after the initial components. This suggests that these initial components capture a large portion of the information present in the original data. Thus, the first principal component (PC 1) explains 57.4% of the total variance, while the second principal component (PC 2) explains 8.4%.

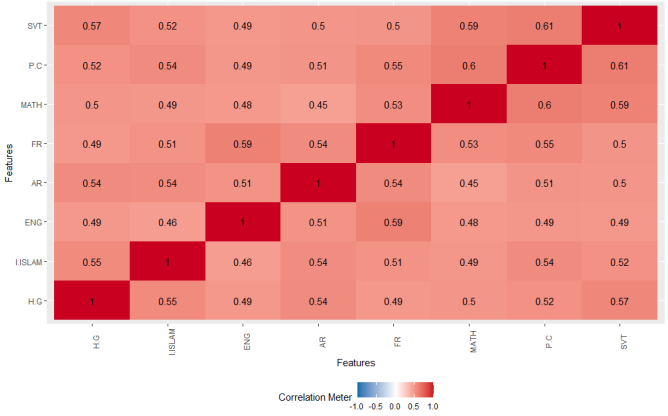

Fig. 2. Correlation matrix

Together, these two components explain approximately 65.8% of the total variance of the data.

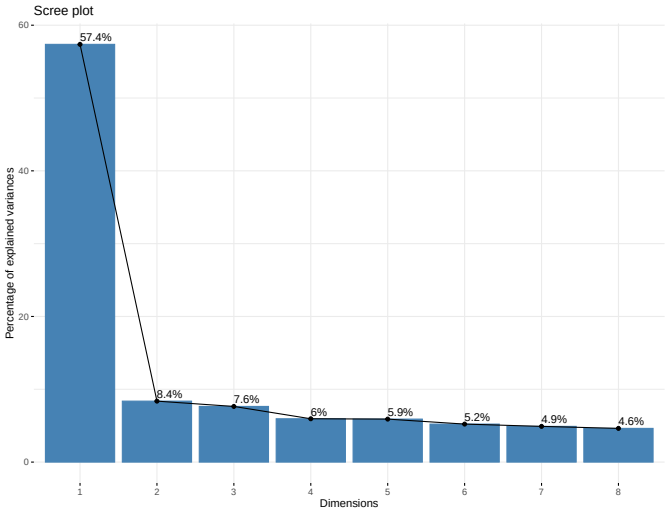

Fig. 3. Percentage of Variance Explained by Each Principal Component

2) circle of correlations: The graph n°4 represents the circle of correlations in Principal Component Analysis (PCA). It shows the correlations of the original variables (here, different subjects like history-geography H.G, Islamic education I.ISLAM, English ENG, etc.) with the identified principal components (Dim 1 to Dim 8).

In this graph, Dim 1 (first principal component) is strongly correlated with all the subjects studied. This indicates that Dim 1 can be interpreted as a measure of overall performance, where a high value on this dimension means that the student tends to have good grades in all subjects.

The other dimensions (Dim 2 to Dim 8) show different configurations of correlations with the subjects. For example, Dim 2 seems to have a moderate positive correlation with MATH, P.C, and SVT, and a moderate negative correlation with ENG, AR, and FR. This could suggest that Dim 2 captures a contrast between skills in sciences and languages.

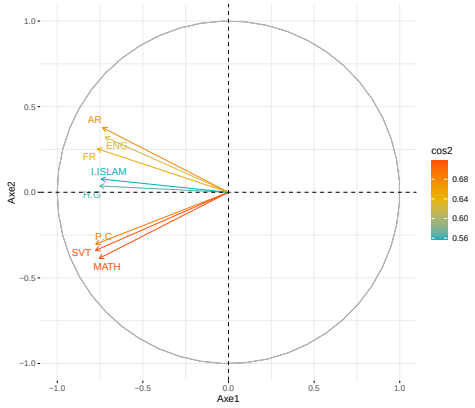

Fig. 4. Variable Factor Map of Principal Components Analysis

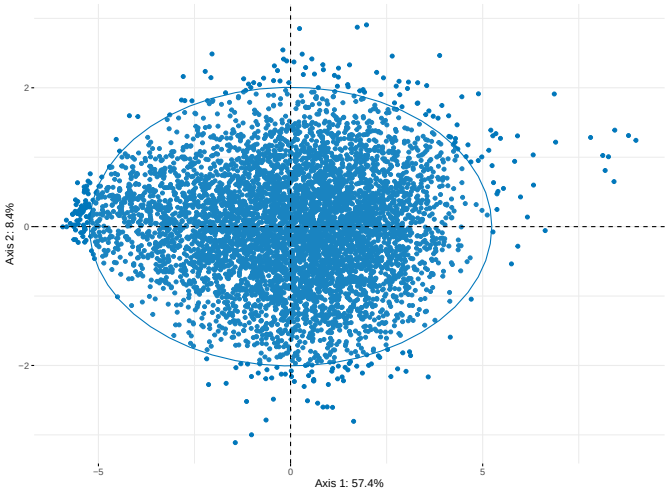

Fig. 5. Distribution of Individual Scores in Two-Dimensional Principal Component Space

IV. Results and Discussion

The graph n° 5 illustrates the distribution of individual scores (likely from students or observations in a study) mapped on a two-dimensional space created by the first two principal components obtained through Principal Component Analysis (PCA).

The axes Dim1 and Dim2 represent the first and second principal component, with Dim1 explaining 57.4% of the variance and Dim2 8.4%. This variance distribution suggests that Dim1 is a much more dominant factor in the overall data than Dim2.

The points on the graph represent individual observations. Their location on the graph provides information about their relative score for each of the two principal components:

Observations near the origin (the intersection of the dashed lines) have average scores on both dimensions. Observations further away from the origin on the horizontal axis have extreme values on Dim1.

Similarly, observations moving away from the origin on the vertical axis have extreme values on Dim2.

The elliptical shape of the point distribution suggests that there is some correlation between Dim1 and Dim2, although Dim1 is clearly the dominant factor. The dispersion of points can also indicate the variability or diversity of observation profiles concerning these two components.

On the school guidance front, what interests us more is the following:

1) Identification of Student Groups: Students located far on the Dim.1 axis, which explains the majority of the variance, can be interpreted as either having overall high or low performance in all subjects, depending on the direction.

Students with high scores on Dim2 may have particular abilities or performances that distinguish them, which are not captured by Dim1. These students may show a preference or talent for specific subjects or skills, which could influence their future orientation.

2) Guidance and Counseling: **Teaching Strategies:** PCA data can inform the development of personalized teaching strategies by identifying students' strengths and weaknesses.

Guidance for School Orientation: Students with similar characteristics on the main components can be advised towards study programs or careers where these characteristics are valued.

Support Programs: Students who appear as outliers could benefit from targeted support programs to improve their skills in specific areas.

3) Educational Planning: **Profile Detection**: Identifying different student groups based on their position on the graph can help in planning individualized educational paths.

Career Guidance: Trends observed on the main components can also be indicative of cross-cutting skills relevant to certain career paths. Considering school guidance, this PCA graph can thus be a powerful tool to inform decisions regarding students' educational and professional paths, by identifying not only their current performance but also anticipating areas where they could excel or require additional support.

The biplot in figure n° 6 illustrates the distribution of students' school performance across different subjects, highlighting variability and major trends along two principal axes. Axis 1, accounting for 57.4% of the variance, appears to primarily distinguish between scientific skills (Biology, Mathematics, Physics-Chemistry) and literary/language subjects (French, English, Arabic, History-Geography, Islamic Studies). Axis 2, contributing to 8.4% of the variance, might reflect additional dimensions of skill or specific pedagogical characteristics. This graph not only reveals individual strengths and weaknesses in each subject but also potential correlations between performances across different disciplines, thereby providing a valuable tool for guiding personalized teaching and support strategies.

V. Conclusion

Principal Component Analysis (PCA) revealed that the first principal component (Dim1), explaining 57.4% of the variance, appears to represent an overall factor of academic

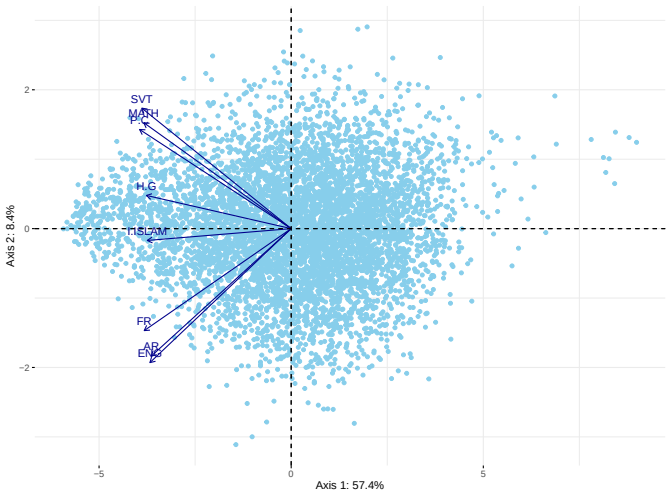

Fig. 6. Analysis of School Performance by Subject

performance. Students scattered along this component could therefore be distinguished based on their overall results, with implications for identifying high-performance profiles and those who may require academic reinforcement.

The second principal component (Dim2), although representing a much lower variance (8.4%), could be indicative of specific abilities or preferences that are not captured by overall performance. Students who stand out on this axis may have strengths in areas that are not strictly related to traditional measures of academic performance, suggesting pathways for more nuanced guidance counseling.

The information extracted from PCA can be valuable for educators and guidance counselors. The trends identified can help identify individual educational needs, tailor teaching methods, and guide students in their academic and career choices.

In terms of academic guidance, this analysis could be used to advise students on course options, majors, or careers that align with their skills and interests. The data also suggest opportunities for personalized education and the development of targeted support programs, focusing on the development of specific skills for students who deviate from the central trend.

In conclusion, the insights derived from this PCA provide a rich database for developing targeted teaching strategies, education and career planning, and strengthening academic guidance to meet the unique needs of students. The graphical representation of student scores in PCA highlights not only the diversity of academic profiles but also the potential for development and personalized guidance based on analytical data.

PCA, in general, could therefore provide valuable insights into the underlying structures of student performance and has highlighted potential correlations between subjects that could be explored for targeted interventions and personalized guidance counseling.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees.

REFERENCES

[1] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical.

[2] Cabrera, A. F., & La Nasa, S. M, "Understanding the college-choice process." New Directions for Institutional Research, 2000(107), 5-22.

[3] Eccles, J. S., & Wang, M.-T. (2012). "So what is student engagement anyway?" Commentary on section IV. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), Handbook of research on student engagement (pp. 133-145). Springer.

[4] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (7th ed.). Prentice Hall.

[5] Smith, A., Jones, B., & Brown, C. (2019). "Using principal component analysis to explore factors influencing high school students' career choices." Journal of Career Development, 46(3), 345-362.

[6] Johnson, D., Williams, E., & Garcia, M. (2018). "Exploring the relationship between academic achievement, cognitive abilities, and career choices using principal component analysis." Journal of Educational Psychology, 110(2), 287-301.

[7] Brown, L., Smith, J., & Davis, R. (2020). "A meta-analysis of the effectiveness of career guidance interventions based on principal component analysis." Journal of Vocational Behavior, 95(4), 345-362.