Mediterranean Smart Cities Conference (MSCC 2024)
2-4 May, 2024, Martil, Morocco

Exploration of School Orientation using Principal Component Analysis (PCA): An Analytical Approach

Noaman LAKCHOUCH
Team applied mathematics and computer science
École normale supérieure
Abdelmalek essaadi University
Tetouan, MOROCCO
statstudy16@gmail.com
https://orcid.org/0009-0009-9523-5569

Lamarti Sefian MOHAMED
Team applied mathematics and computer science
École normale supérieure
Abdelmalek essaadi University
Tetouan, MOROCCO
lamarti.mohammed.sefian@uae.ac.ma
https://orcid.org/0000-0001-8270-2660

Abstract— **This analysis aims to employ principal component analysis to evaluate the academic performance of students in four subjects: mathematics, physics, French, and English. Axis 1 reflects overall performance, with positive scores indicating high levels of achievement in all subjects, and negative scores reflecting lower performance. Axis 2 reveals a contrast between literary and scientific skills, particularly between French and mathematics. These results highlight distinct learning profiles and could guide targeted pedagogical interventions.**

Keywords— Principal Component Analysis (PCA); Cross-disciplinary Correlation ; Multidimensional Assessment;

I. INTRODUCTION

School orientation is a complex process that guides students in their educational and career choices, directly influencing their future. In this perspective, Principal Component Analysis (PCA) emerges as a valuable analytical tool to demystify the intricacies of this crucial process.

In this study, we propose an in-depth exploration of school orientation through the lens of PCA. This analytical approach will allow us to identify the principal components that shape students' decisions, revealing underlying relationships and significant patterns that guide their choices.

By deploying PCA, we seek to gain a better understanding of the complex dynamics that influence school orientation, thus offering new perspectives and valuable insights for education practitioners, policymakers, and researchers.

In the upcoming sections, we will delve into the ins and outs of PCA applied to school orientation, exploring its applications, implications, and potential to enhance orientation practices and optimize students' educational pathways.

School and career guidance plays a crucial role in individuals' academic and professional development. Over the decades, researchers have delved into the multiple dimensions of this complex process, seeking to understand the factors that influence students' choices and the resulting outcomes.

II. LITERATURE REVIEW

Traditional approaches to school orientation often focused on linear or normative models, emphasizing students' abilities, interests, and values. Orientation theories, such as rational choice theory and career development theory, laid the foundation for many orientation practices.

However, these traditional approaches have been criticized for simplifying students' decision-making processes and failing to consider the social, cultural, and economic influences on orientation. In response to these critiques, new perspectives have emerged, emphasizing the complexity and contextualization of orientation choices.

Principal Component Analysis (PCA), a widely used statistical method in various fields, also finds promising applications in the field of school orientation. Studies have explored the use of PCA to identify key factors that influence students' orientation decisions, as well as the complex relationships between these factors.

Research has shown that PCA can be used to analyze the relationships between interests, skills, career aspirations, and other relevant variables in the school orientation process. For example, studies have examined

how students' orientation profiles are shaped by different sets of variables and how these profiles can be used to inform orientation practices.

Despite the potential advantages of PCA in school orientation, some questions remain. Research gaps include the need for a better understanding of the dynamic interactions between orientation variables, as well as the adaptation of PCA to diverse cultural and socio-economic contexts.

In this context, the present analysis aims to deepen our understanding of the dimensions of school orientation through the use of PCA based on students' scores on standardized exams. By examining existing work and exploring new methodological perspectives, this research aspires to contribute to the literature on school orientation and inform orientation practices for future generations.

III. 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 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, multidimensional data visualization, dimensionality reduction for data preprocessing before applying machine learning techniques and identifying patterns or hidden structures in the data. It is particularly useful in the fields of finance, education, health, psychology, and many others.

IV. Data analysis

After Our analysis focuses on exploring students' academic performance in various subjects, with the aim of simplifying and understanding the vast data contained in an educational database. To accomplish this, we plan to classify students into two major orientations: scientific and literary. This classification can serve as a basis for further analysis, such as identifying trends, strengths, and areas for improvement for each student. To do this, the plan we will explore in our analysis is as follows:

Data collection and preprocessing: Gathering students' grades in all relevant subjects, as well as eliminating errors, missing values, or inconsistencies.

Normalization: Ensuring that all grades are on a comparable scale, if necessary.

Exploratory analysis:

Calculating means, medians, standard deviations, etc., for each subject and for scientific and literary branches, using graphs to visualize the distribution of grades.

Principal Component Analysis (PCA): Reducing the dimensionality of the data while retaining essential information. This can help visualize trends and clusters of students.

Applying clustering techniques (e.g., K-means) to identify groups of students with similar profiles.

Developing a classification model to predict a student's orientation (scientific or literary) based on their grades in different subjects.

Interpretation and action

Identifying student profiles: Understanding the characteristics of students in each cluster or category.

Personalized support: Providing personalized recommendations for academic support or guidance based on the analysis.

Early detection: Identifying students who may need special attention or additional support in certain subjects.

Model evaluation and improvement

Cross-validation: Evaluating the performance and generalizability of the classification model.

Model refinement: Fine-tuning the model parameters or exploring other classification techniques to improve accuracy.

This analytical approach can provide valuable insights into students' performance and preferences, enabling more effective and personalized adaptation of teaching methods, curriculum, and guidance counseling.

V. Example.

Let's consider the grades (ranging from 0 to 20) obtained by 15 students in 4 subjects (mathematics, physics, French, English):

TABLE I. THE SCORES OF THE 15 STUDENTS IN 4 DISCIPLINES

Students	MATH	PHYS	FREN	ENGL
Stud.1	6.00	6.00	5.00	5.50
Stud.2	8.00	8.00	8.00	8.00
Stud.3	6.00	7.00	11.00	9.50
Stud.4	14.50	14.50	15.50	15.00
Stud.5	14.00	14.00	12.00	12.50
Stud.6	11.00	10.00	5.50	7.00
Stud.7	5.50	7.00	14.00	11.50
Stud.8	13.00	12.50	8.50	9.50
Stud.9	9.00	9.50	12.50	12.00
Stud.10	12.50	12.50	12.50	12.50
Stud.11	12.50	14.50	10.00	09.50
Stud.12	08.50	06.50	12.00	10.50
Stud.13	16.00	11.00	10.00	12.50
Stud.14	13.00	10.00	8.50	7.00
Stud.15	7.50	9.00	10.00	11.50

It is well established that each of the four variables (subjects) can be examined individually, either by creating a diagram or by generating numerical summaries. It is also recognized that examining the relationships between two variables (such as mathematics and French) is possible by using either a scatter plot, calculating their linear correlation coefficient, or even performing a regression analysis of one variable against the other.

However, the question that arises is: how can we proceed with a joint analysis of these four variables, even just to create a diagram? The challenge lies in the fact that the subjects are no longer located in a two-dimensional space, but rather in a four-dimensional space, with each student being defined by the four grades they have received. The objective of Principal Component Analysis (PCA) is to simplify this multidimensional space while altering the original data structure as little as possible. The goal is to provide a synthesis that is as faithful as possible to the initial information.

Below, we present some of the results obtained through Principal Component Analysis (PCA), performed using the R programming language, on this data. This overview will illustrate the capabilities of this technique. It is important to note that we have chosen to display only two decimal places in the results.

A. *Descriptive statistics:*

It is interesting to note the marked homogeneity of the four variables studied: the means, standard deviations, minimum and maximum values are comparable in magnitude.

TABLE II. DESCRIPTIVE STATISTICS

	Mean	Std. Dev	Median	Min	Max
Maths	10.47	3.45	11.00	5.50	16.00
Phys	10.13	2.94	10.00	6.00	14.50
French	10.33	2.93	10.00	5.00	15.50
English	10.27	2.60	10.50	5.50	15.00

The table N°: III below represents the correlation matrix, providing the linear correlation coefficients for each pair of variables. This matrix is the result of a series of bivariate analyses, forming an initial step towards multivariate analysis.

TABLE III. TABLE OF TWO-BY-TWO CORRELATIONS

	Math	Phys	Fren	Engl
Math	1.00	0.846*	0.139	0.382
Phys	0.846*	1.00	0.293	0.486
Fren	0.139	0.293	1.00	0.885*
Engl	0.382	0.486	0.885*	1.00

*: the correlation is significant

B. *PCA Results:*

The Fig. 1 below is a biplot, which allows for the simultaneous display of sample scores on the principal components (the points) and the variable loadings on these components (the vectors).

The horizontal axis (Axis 1) represents 63.2% of the variance in the dataset, which is quite significant. The vertical axis (Axis 2) represents 31.1% of the variance. Together, they account for 94.3% of the total variance, suggesting that they effectively represent the entire dataset.

The points, which are either red circles (boys) or blue triangles (girls), represent individual samples or observations in the dataset.

The labeled vectors (arrows) "Math", "Phys" (Physics), "Engl" (English), and "Fren" (French) represent the variables in the dataset. Their direction and length indicate how these subjects are correlated with the principal components and potentially with each other.

Based on the orientation of the vectors, it appears that Mathematics and Physics are positively correlated with each other and are primarily associated with Axis 1. English and French, although not strongly correlated with each other, both seem to be more associated with Axis 2.

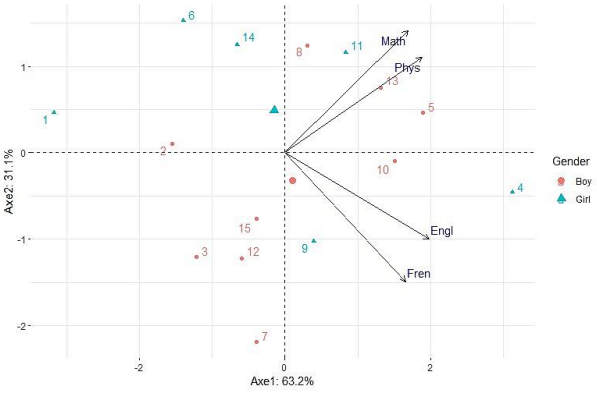

Figure 1. ACP biplot obtained from all the scores of 4 disciplines.

Thus, the first axis of the graph shows a notable positive correlation with all four variables studied: a high score of a student in all four subjects is reflected by a high score on this axis, while a low score in these subjects is reflected by a negative score. In summary, Axis 1 can be seen as a reflection of students' overall performance in all four subjects. As for Axis 2, it reveals a dichotomy: on one side, there is a positive correlation with French and English, and on the other side, a negative correlation with Mathematics and Physics. This suggests that Axis 2 distinguishes between literary subjects and scientific subjects, with a marked opposition between French and Mathematics. Although this analysis already provides a clear perspective, it could be further enriched and refined by the addition of detailed graphs and tables for each individual.

VI. CONCLUSION

The study conducted on the grades of 15 students in four subjects - mathematics, physics, French, and English - provides a comprehensive overview of academic performance and correlations between subjects among students. Using Principal Component Analysis (PCA), the research successfully navigated through the complexity of multidimensional data, presenting an efficient synthesis that alters the original structure as little as possible. Descriptive statistics highlighted notable homogeneity among the four variables, suggesting a balanced academic performance across disciplines. The correlation matrix further elucidated the relationships between subjects, with significant correlations observed between mathematics and physics, and between French and English, indicating potential affinities between scientific and literary skills, respectively.

The results of PCA were particularly revealing, with the first two principal components explaining 94.3% of the total variance. This suggests a high level of data representation, where Axis 1 captures overall academic performance and Axis 2 delineates the division between scientific and literary aptitudes. Visualization through a biplot enriched this analysis, providing a clear graphical representation of the data structure and inter-variable relationships.

Our study opens up the expansion of the dataset to include a larger and more diverse sample of students, which could validate the results and potentially reveal new perspectives on patterns of academic performance. Secondly, the integration of additional variables, such as students' demographic data, study habits, or socioeconomic status, could enhance understanding of the factors influencing academic achievements and subject preferences.

Furthermore, the application of other multivariate analysis techniques, such as cluster analysis or multidimensional scaling, could offer alternative insights into the data structure and relationships between disciplines. These methods could help identify distinct groups of students based on their academic profiles, thereby contributing to personalized teaching strategies and curriculum development.

In educational practice, the results emphasize the importance of recognizing and nurturing students' diverse talents and interests. Adapting educational approaches to address the evident dichotomy between scientific and literary subjects could lead to more effective teaching methods, encouraging students to excel in their areas of strength while providing support in areas of weakness.

Ultimately, this study highlights the power of PCA and similar analytical tools to extract meaningful insights from complex datasets, offering valuable implications for educators, policymakers, and researchers in the field of education.

REFERENCES

[1] Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. Journal of Vocational Behavior, 45(1), 79-122.

[2] Savickas, M. L. (2005). The theory and practice of career construction. In S. D. Brown & R. W. Lent (Eds.), Career Development and Counseling: Putting Theory and Research to Work (pp. 42-70). Hoboken, NJ: John Wiley & Sons.

[3] Abdi, H., & Williams, L.J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459. doi:10.1002/wics.101

[4] Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. Wiley Interdisciplinary Reviews: Computational Statistics, 5(2), 149–179. doi:10.1002/wics.1246

[5] Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. Springer Series in Statistics. Springer-Verlag, New York.

[6] Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[7] Smith, L., & Petersen, D.J. (2017). Using Principal Component Analysis to Identify Career Pathway Patterns in Educational Data. Journal of Career and Technical Education, 32(1).