VNU Journal of Science: Education Research
Journal homepage: https://js.vnu.edu.vn/ER

Original Article

Introducing a Cutting-Edge Dataset: Revealing Key Factors in Student Academic Outcomes and Learning Processes

Nguyen Thi Kim Son[1,2,*], Nguyen Hong Hoa[3], Hoang Thi Thu Trang[3], Tran Quynh Ngan[3]

[1]*Hanoi University of Industry, 298 Cau Dien, Bac Tu Liem, Hanoi, Vienam*
[2]*Vietnam Academy of Science and Technology, Graduate University of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam*
[3]*Hanoi Metropolitan University, 98 Duong Quang Ham, Cau Giay, Hanoi, Vietnam*

Received 18th June 2024
Revised 17th September 2024; Accepted 09th October 2024

Abstract: This paper presents an educational dataset that consolidates various aspects of educational science. The purpose of this study is to illuminate the factors that impact students' learning experiences before and during their time at university. This dataset was designed to support research in educational science, including the application of machine learning and deep learning models to predict student outcomes. The primary objective is to improve educational methodologies, empower students with informed decision-making, and enhance overall learning effectiveness. The dataset comprises 992 samples across 89 fields and was collected through direct methods like questionnaires and indirect methods involving training management units. These samples are categorized into Personalized factors, Factors affecting learning outcomes, and Learning outcomes, encompassing both general education performance and university module achievements. Following collection, the dataset was subjected to thorough processing, cleaning, and statistical analysis, using techniques such as Pearson correlation analysis, analysis of variance, Std. Error, Std. Deviation, and tests of homogeneity of variance.

Keywords: Educational science, Data science, Dataset, Learning outcomes, Influencing factors, Educational data mining.

1. Introduction

The Fourth Industrial Revolution is occurring in a new era marked by transformative digital technologies, with big data emerging as a crucial element [4] (Sarmiento et al., 2021). Progress in science and technology, particularly in artificial intelligence and deep learning for data analysis, has revolutionized decision-making across various sectors, including educational science [7] (Taylor, 2021). Educational data mining has

* Corresponding author.
E-mail address: sonntk@haui.edu.vn.

https://doi.org/10.25073/2588-1159/vnuer.5157

seen a rise in popularity, with its applications spanning a wide range of areas, such as enhancing learning processes, improving course completion rates, aiding course selection, creating student profiles, identifying reasons for dropout, understanding student objectives, refining curricula, and predicting learning outcomes [1] (Zorić, 2020).

Recent studies have focused on datasets related to students' learning activities. For instance, Tran Trung et al,. conducted a survey capturing Vietnamese students' learning habits during the COVID-19 pandemic [14] (Tran Trung et al., 2020). Similarly, Dien Thi Bui et al., curated a dataset on online learning activities among secondary school students in Vietnam, collected through Google Form surveys [2] (Dien et al., 2022).

In the midst of current educational trends, there has been a noticeable increase in student dropout rates and academic warnings [15] (Lan et al., 2003). Scholars like Tuti Hartati and colleagues advocate for data-driven methodologies to personalize and enhance the quality of education, highlighting its significance in educational institutions [13] (Hartati et al., 2023). Conversely, Misiejuk et al., raise concerns regarding the scarcity of educational data systems, citing deficiencies in sources, regulations, and standardized data [6] (Misiejuk et al., 2023).

Educational systems worldwide accumulate extensive datasets, which necessitate robust data management frameworks. The systematic integration and use of data across various disciplines poses significant challenges to the current educational landscape. Consequently, harnessing big data technologies holds promise in revolutionizing pedagogical approaches and optimizing educational experiences to enhance the effectiveness of educational systems [10] (Omarova et al., 2024).

This article introduces the development of an extensive dataset containing student learning records, covering the factors that affect learning outcomes and university course grades. To establish a clear focus for our analysis, this research addresses three fundamental questions. First, it seeks to identify the key factors influencing students' academic performance and evaluate the crucial attributes in the input data that significantly affect their outcomes. Second, the study aims to detail the essential data preprocessing steps required to construct a robust training dataset, enabling the effective application of advanced data analysis techniques such as statistical methods, machine learning, and deep learning. Finally, the research will explore how educational administrators can leverage this dataset to enhance educational management practices and make more informed, data-driven decisions.

This paper analyzes data from nearly a thousand students, providing comprehensive insights into the factors affecting their learning outcomes through meticulous statistical analysis. This dataset facilitates the application of advanced data analysis methodologies, such as machine learning and deep learning, thereby aiding decision-making in educational endeavors.

This paper stands out for its novelty and contributions, distinguishing itself from previous studies through several key highlights. Firstly, the research team developed a highly detailed survey tool specifically tailored for students majoring in Mathematics Education and Physics Education. This tool, which includes 89 attributes related to personal information, family factors, and academic performance, represents a comprehensive and innovative approach within the field. Moreover, we compiled and structured a dataset of 992 students spanning 10 cohorts from 2014 to 2023, providing a robust foundation for in-depth analysis of factors influencing academic success. Utilizing advanced statistical methods, the paper meticulously examines the determinants of academic performance, identifying the most critical factors and drawing well-founded scientific conclusions. Finally,

from these analyses, the research team proposed targeted solutions to enhance the academic outcomes of students in Mathematics and Physics Education, paving the way for new and practically significant research directions. In essence, the paper not only introduces a novel approach and methodology but also makes a substantial contribution to improving educational quality in the fields of Mathematics and Physics Education.

2. Literature View

In recent years, numerous global studies on educational data have been conducted, particularly focusing on factors influencing academic performance. Superby et al., [5] developed a survey to collect personal information, students' behaviors, and learning perceptions. Farooq [9] found that students' academic performance is influenced by various factors, which can be categorized into two groups: internal and external factors. Internal factors mainly pertain to individual students, their interest in learning, and their time management, whereas external factors are beyond students' control and planning. Marcus Credé et al., [3] emphasized the significance of factors such as educational programs, faculty quality, access to services, training environment, and university facilities in influencing students' academic performance. Singh et al., [12] highlighted the impact of psychological, economic, social, personal, and environmental factors on students' academic outcomes. According to Irfan Mushtaq and Shabana Nawaz Khan [8], a model with four hypotheses was proposed that affects students' academic performance: the use of technological devices related to supporting software and media, activities during the learning process, students' cognitive abilities, motivation, and attitudes, and classroom characteristics and the learning environment. Ali et al., [11] identified student-related factors, including their efforts, age, learning motivation, interests, entry-level qualifications, academic performance, and learning environment in previous educational stages.

Domestically, Vo Van Kiet et al., [17] identified seven primary factors affecting academic performance: the level of interest in learning, facilities, peer pressure, intellectual capacity, and family. Nguyen Thi Thu An et al., [16] analyzed student characteristics such as gender, entrance aspirations, participation in class committees/Youth Union, and lecturer competence to examine their relationship with academic performance. Academic performance varied based on students' personal characteristics; female students generally performed better than male students. Students admitted through the second-choice option performed better than those admitted through the other choices. Additionally, class committee/Young Union members tended to have higher scores than other students. Nguyen Thi Nhu Quynh [18] argued that teachers are the most critical factor affecting students' academic performance, along with other factors such as subjectivity and school facilities.

Overall, most reports have analyzed predictive models based on factors influencing academic performance or short-term outcomes from a single course or specific period. However, integrating the analysis of factors affecting academic performance with comprehensive results over the entire university duration can yield more robust outcomes. Effective and detailed methodologies for constructing datasets can significantly improve predictive models. This approach is vital for advancing digital transformation in education.

3. Datasets, Experimental Design, Materials and Methods

In this paper, we introduce the data set collected from students and alumni of Hanoi Metropolitan University (HNMU). HNMU is an esteemed public university in Hanoi, Vietnam, offering educational programs in pedagogical fields and others.

The data on students' learning processes are collected through departments such as the

Training Management Department or the Office of Professional Faculty. These data come in various forms (numeric, text, functions, etc.) and are often stored in different formats (PDF, Excel, Word). On another hand, the dataset originated from survey research also conducted from June 2023 to March 2024 through a wide survey on the factors influencing students' learning outcomes.

The questionnaire was thoughtfully crafted, drawing on insights from prior research on factors influencing academic performance and incorporating feedback from experts. It was subsequently refined to suit the specific needs of pedagogy students at Hanoi Metropolitan University, resulting in a comprehensive and targeted survey. To ensure its effectiveness, the research team conducted a pilot test with a small group of students, allowing us to determine the most appropriate response formats for each question, such as short answers and Likert scales.

Following this, we launched an online survey via Google Forms, distributing it to all current students in the Faculty of Education. For graduates, we employed direct or online interviews through social media platforms. Academic records and graduation data were obtained directly from the university's training management office to ensure accuracy.

Survey responses and academic data were collected independently and later integrated into a unified dataset using student identification numbers. To maintain objectivity and protect privacy, we ensured that all data was used solely for research purposes and that personal information was anonymized during preprocessing.

The study utilizes a sample of 992 students from seven cohorts between 2014 and 2020, who have graduated, and three cohorts from 2021 to 2023, who are currently studying in the Mathematics Education and Physics Education programs. Retaining educational records and conducting surveys with these alumni posed significant challenges, requiring the research team to invest an entire year in completing this process. The sample encompasses 89 attributes, including personal information, family factors, environmental influences, and academic performance throughout their studies, ensuring comprehensive and representative coverage.

As traditional disciplines, Mathematics Education and Physics Education are a well-represented sample size, which is sufficient to reflect the scale of training at Hanoi Metropolitan University for these majors, where the average annual enrollment is around 100 students. This is also the largest sample in majors that the research team was able to gather. Despite the challenges faced in data collection, the current sample size is robust enough to ensure the representativeness and reliability of the analysis results. A total of 1000 responses were received; however, only 992 valid responses were retained for further analysis after excluding invalid submissions. Of these, 715 observations with valid graduation labels were accepted for further analysis of correlations with academic performance. The initial dataset was largely inconsistent and was processed using Excel. Additionally, we excluded data from students who did not wish to provide their pre-university academic performance. Consequently, the complete dataset comprised 992 responses and was subjected to analysis using IBM SPSS Version 27.

The questionnaire highly correlated with the results of exploratory factor analysis (EFA), indicating that the KMO index $= 0.848 > 0.5$ and the EFA analysis is suitable for the research data. The Chi-square value in the Bartlett test is large, with a significance level of $\text{sig} = 0.000 (< 0.05) = 0.000 (< 0.05)$. Therefore, the observed variables are correlated with each other in the overall scope. The Pearson correlation analysis is opted aligned with our study's primary objective of evaluating the relationships and differences between variables within the dataset. Pearson correlation, specifically, measures the

linear relationship between two variables, allowing us to identify both strong and weak correlations. This approach offers an initial insight into how various factors impact academic outcomes, providing a foundation for more advanced analyses.

Table 1. KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.848
Bartlett's Test of Sphericity	Approx. Chi-Square	18724.482
	Df	666
	Sig.	0.000

3.1. Values of the Data

Enhancing the current relatively limited educational science data repository by supplementing it with an additional dataset, including demographics, learning-influential factors, and progressing academic performance (with 992 samples across 89 fields and detailed information).

Streamlining the application of artificial intelligence and modern analytical tools like machine learning and deep learning, for educational analysis, thus aiding decision-making in educational activities (support a most valued training dataset to machine learning or deep learning models in Learning Analysis - LA).

Enabling researchers to delve into correlations among learner characteristics, school dynamics, societal influences, and academic accomplishments by analyzing the concrete dataset.

Assess students' study habits and potential academic advancement.

Support educational administrators in enhancing the quality of teaching and learning through data-driven decision-making.

Empowers students to self-assess their learning approaches and set personalized objectives.

Identifying and offering assistance to students who may be at risk of academic underperformance or dropout.

Table 2. Descriptive Specifications

Subject	Education, Higher Education.
Specific subject area	Learning analytics, Machine learning.
Type of data	Table/Figure/Excel file/Sav file
How data were acquired	The data were collected through online survey, direct inquiries to postgraduate students, and from the Training Department of Hanoi Metropolitan University. Then, the dataset was converted into .xlsx format for formal analysis in SPSS v.27.
Data format	Raw Analyzed
Parameters for data collection	The survey subjects were students majoring in Natural Sciences at Hanoi Metropolitan University (cohorts 14 - 23).
Description of data collection	The data collection sources comprise online survey data, direct inquiries to postgraduate students, and data from the Training Department of Hanoi Metropolitan University.
Data source location	The data was collected from Hanoi Metropolitan University.

4. Results

This model was constructed using a dataset comprising 992 students enrolled at Hanoi Metropolitan University from 2014 to 2023. The dataset encompasses three primary groups of variables: (A) Personalization factors of participating students, including gender and parents' educational levels, etc; (B) Factors influencing learning outcomes, including study time, social media usage time, scholarships, health status, and employment status, etc; and (C) Academic performance, including pre-university and university academic performance, etc.

A. PERSONALIZATION FACTORS

Table 3. Statistics of personalization factors gender, major, and cohort

	Gender		Major		
Total	Male	Female	Mathematics Teaching		Physics Teaching
992	161	830	744		248
Cohorts					
Total	14	15	16	17	18
	148	112	101	89	107
992	19	20	21	22	23
	66	92	120	107	50

B. FACTORS INFLUENCING LEARNING OUTCOMES

Table 4. Individual factors influencing learning outcomes

	Frequency	Percent	Valid Percent
Parents' educational level			
Primary school	4	0.3	0.4
Secondary school	82	6.9	8.3
High school	587	49.6	59.2
College	300	25.3	30.2
University	18	1.5	1.8
Others	1	0.1	0.1
Total	992	83.8	100.0
Part-time job			
No	315	26.6	31.8
Yes	677	57.2	68.2
Total	992	83.8	100.0
Funding for tuition fees			
Yourself	293	24.7	29.5
School	26	2.2	2.6
Family	673	56.8	67.8
Total	992	83.8	100.0

Study time			
Less than 1 h	7	0.6	0.7
From 1 to 3 h	162	13.7	16.3
From 3 to 6 h	554	46.8	55.8
From 6 to 8 h	256	21.6	25.8
Over 8 h	13	1.1	1.3
Total	992	83.8	100.0
Social media usage time			
Over 8 h	36	3.0	3.6
From 6 to 8 h	112	9.5	11.3
From 3 to 6 h	433	36.6	43.6
From 1 to 3 h	391	33.0	39.4
Less than 1 h	20	1.7	2.0
Total	992	83.8	100.0
The total number of social media platforms used			
1	84	7.1	8.5
2	369	31.2	37.2
3	287	24.2	28.9
4	197	16.6	19.9
5	55	4.6	5.5
Total	992	83.8	100.0
Health condition			
Sick and weak	69	5.8	7.0
Slightly weak	134	11.3	13.5
Strong	563	47.6	56.8
Totally strong	226	19.1	22.8
Total	992	83.8	100.0
Groups of admision subjects			
Social sciences	414	35.0	41.7
Natural sciences	578	48.8	58.3
Total	992	83.8	100.0
Methods of admission			
Academic records	437	36.9	44.1
High school examination scores	520	43.9	52.4
Direct admission	0	0	0
Others method	35	3.0	3.5
Total	992	83.8	100.0
Ranking choices			
1	45	3.8	4.5
2	71	6.0	7.2

3	112	9.5	11.3
4	192	16.2	19.4
5	289	24.4	29.1
6	283	23.9	28.5
Total	992	83.8	100.0
Scholarship			
No	847	71.5	85.4
Yes	145	12.2	14.6
Total	992	83.8	100.0

Table 5. Environmental factors

Environmental factors	N	Range Statistic	Min	Max	Mean		Std. Deviation	Variance
					Statistic	Std.Error		
Level of environmental adaptation	992	4	1	5	3.23	0.026	0.655	0.809
Learning methods	992	4	1	5	3.33	0.023	0.509	0.713
Level of school support	992	4	1	5	2.84	0.027	0.703	0.839
Level of instructor support	992	4	1	5	3.37	0.023	0.548	0.740
Facility conditions	992	4	1	5	2.54	0.034	1.128	1.062
Quality of instructors	992	4	1	5	3.27	0.025	0.603	0.776
Suitability of the training program	992	4	1	5	3.76	0.029	0.848	0.921
Competitiveness in studies	992	4	1	5	3.50	0.023	0.541	0.735
Influence of friends	992	4	1	5	3.04	0.038	1.408	1.187
Level of interest in the field of study	992	4	1	5	3.30	0.022	0.489	0.699

Environmental factors were assessed using a Likert scale. The Likert scale presents a series of statements or items to which respondents indicate their level of agreement or disagreement on a designated scale. In this article, then environment factors are scaled in 5 levels from "Strongly Disagree" to "Strongly Agree".

In table 5, "N" represents the total number of observations or data points in the dataset.

In statistics, the "range" is a measure of the dispersion or spread of a set of data points. The value is calculated as the difference between the maximum and minimum values in the dataset. The formula for calculating the range (R) is: $R = \max value - \min value$, where Min and Max denote the lowest level while Max refers to the highest level.

The mean statistic is a measure of central tendency that represents the average value of a set of data points. The formula for calculating the mean $(\bar{x})$ of a set of n data points $(x_1, x_2, ...x_n)$ is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

The standard deviation, also known as the standard deviation, is a quantity used to measure the degree of dispersion of a given dataset presented in a frequency table. It can be concluded that this method is widely used to measure the variability of a dataset. If the variability or dispersion of the data is greater, the standard deviation is greater than the mean value. The formula for calculating the standard

deviation is $S=\sqrt{\frac{\sum_i^n (x_i-\bar{x})^2}{n-1}}$, in which N is the number of observations in the sample; x_i represents each individual value in the sample and $\bar{x}$ is the mean (average) of the sample.

Variance is a measure of how much the values in a dataset differ from the mean (average) of the dataset. A high variance indicates that the values are widely spread from the mean, whereas a low variance indicates that the values are clustered closely around the mean. The formula for calculating the variance (σ^2) of a set of n data points $(x_1, x_2, ..., x_n)$ is $\sigma^2=\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n}$ $\sigma^2=\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n}$. In which x_i is individual data point, $\bar{x}$ is mean of the dataset, and n is the total number of data points.

Table 6. Correlation between individual factors influencing learning outcomes and university students' academic performance

Individual factors	Pearson Correlation	Sig. (2-tailed)	N
Parents' educational level	0.670**	0.000	715
Partime job	0.042	0.259	715
Funding for tuition fees	0.031	0.414	715
Study time	0.604**	0.000	715
Social media usage time	0.379**	0.000	715
Health condition	0.024	0.527	715
Groups of admision subjects	0.019	0.620	715
Methods of admission	0.254**	0.000	715
Ranking choices	-0.031	0.406	715
Scholarship	0.452**	0.000	715

The Pearson correlation coefficient (is denoted as r) is a measure of the linear relationship between two variables. The method quantifies the strength and direction of the association between variables. The Pearson correlation coefficient can take values between -1 and $+1$. First, $r=1$ indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally and $r=-1$ indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally. Finally, $r=0$ indicates no linear relationship between variables.

"Sig" typically stands for "significance" and is often used to denote the statistical significance level of the correlation coefficient. The Pearson correlation coefficient is highly significant when sig $>0.05<0.01$, significant when sig <0.05 and not significant when sig >0.05.

Table 7. Statistics of individual factors by cohort

Cohorts	N	Mean	Std. Deviation
Parents' educational level			
Cohort 14	148	3.32	0.682
Cohort 15	112	3.25	0.593
Cohort 16	101	3.06	0.544

Cohort 17	89	3.03	0.593
Cohort 18	107	3.03	0.590
Cohort 19	66	3.44	0.585
Cohort 20	92	3.33	0.576
Total	715	3.21	0.618
Study time			
Cohort 14	148	3.24	0.645
Cohort 15	112	3.24	0.633
Cohort 16	101	3.12	0.553
Cohort 17	89	3.09	0.596
Cohort 18	107	3.10	0.582
Cohort 19	66	3.35	0.511
Cohort 20	92	3.33	0.516
Total	715	3.21	0.593
Social media usage time			
Cohort 14	148	3.70	0.476
Cohort 15	112	3.74	0.440
Cohort 16	101	3.32	0.564
Cohort 17	89	3.01	0.612
Cohort 18	107	2.71	0.991
Cohort 19	66	2.76	0.842
Cohort 20	92	3.17	0.820
Total	715	3.26	0.786
Methods of admission			
Cohort 14	148	1.64	0.481
Cohort 15	112	1.71	0.514
Cohort 16	101	1.56	0.607
Cohort 17	89	1.38	0.489
Cohort 18	107	1.52	0.604
Cohort 19	66	1.76	0.681
Cohort 20	92	1.76	0.669
Total	715	1.62	0.581
Scholarship			
Cohort 14	148	1.15	0.357
Cohort 15	112	1.09	0.286
Cohort 16	101	1.07	0.255
Cohort 17	89	1.07	0.252
Cohort 18	107	1.13	0.339
Cohort 19	66	1.18	0.389
Cohort 20	92	1.23	0.422
Total	715	1.13	0.335

Approximately half of the personalized factors showed little to no correlation with GPA outcomes. These include part-time jobs, funding for tuition fees, ranking choices, groups of subjects for admission, and health conditions. On the other hand, there was a relatively strong correlation between parents' educational levels and study time, with Pearson's values surpassing 60%. The remaining three factors exhibit weaker correlations, with Pearson's coefficients ranging from 20% to 40%.

The factors with the highest correlations include parents' educational levels and study time, with correlation coefficients of 67% and 60.4%, respectively. According to Table 7, cohort 19 students exhibited the highest average parental educational level scores of 3.44. The factors with the highest correlations include parents' educational levels and study time, with correlation coefficients of 67% and 60.4%, respectively. Also, the Table 6 shows that the average study time across the seven cohorts ranged from 3 to 3.5, with cohort 17 having the lowest average value of 3.09.

In the Table 6, the remaining three factors with a lower correlation are social media usage, scholarships, and admission methods, with Pearson coefficients ranging from 20% to 40%. Both cohorts 18 and 19 spend the least amount of time on social media, with recording values of 2.71 and 2.76, respectively. The Table 7 also shows that most students use high school academic records and exam scores as their admission criteria. The percentage of Mathematics Education students receiving scholarships was 0.08 higher than that of Physics Education students.

Table 8. Statistical of individual factors by major

Majors	N	Mean	Std. Deviation
Parents' educational level			
Mathematics	551	3.21	0.631
Physics	164	3.20	0.574
Total	715	3.21	0.618
Study time			
Mathematics	551	3.22	0.594
Physics	164	3.16	0.589
Total	715	3.21	0.593
Social media usage time			
Mathematics	551	3.19	0.821
Physics	164	3.51	0.591
Total	715	3.26	0.786
Methods of admission			
Mathematics	551	1.63	0.576
Physics	164	1.57	0.597
Total	715	1.62	0.581
Scholarship			
Mathematics	551	1.15	0.354
Physics	164	1.07	0.251
Total	715	1.13	0.335

When major types were factored in, the disparity in Social media usage time between the two fields was highest, while the other factors were nearly equivalent.

Among the environmental factors, learning methods showed the strongest correlation, surpassing 67%. According to Table 10, the study methods of cohorts 14, 19, and 20 were the most effective, with an average value above 3.5. Next are competitiveness in studies and level of interest in the field of study, with correlation levels of 56% and 54%, respectively. The data show that the average value for the level of interest in the field of study varies slightly, ranging from approximately 3 to 3.4. Notably, competitiveness has significantly higher statistical values than the other factors, with the average values being above 3.6. Competitiveness greatly impacts academic performance because it can create pressure and motivate students to strive harder.

Table 9. Correlation between environmental factors influencing learning outcomes and university students' academic performance

Environmental factors	Pearson Correlation	Sig. (2-tailed)	N
Level of environmental adaptation	0.322**	0.000	715
Learning methods	0.677**	0.000	715
Level of school support	-0.112**	0.003	715
Level of instructor support	-0.019	0.605	715
Facility conditions	-0.055	0.142	715
Quality of instructors	-0.092*	0.014	715
Suitability of the training program	-0.358**	0.000	715
Competitiveness in studies	0.560**	0.000	715
Influence of friends	0.011	0.776	715
Level of interest in the field of study	0.540**	0.000	715

Additionally, the level of adaptation to the environment also correlated, but only at 32.2%. Cohorts 18 and 19 exhibited notably lower levels of adaptation to the environment compared to other groups. Concerning the Suitability of the training program, most students rate it as average, ranging from 3 to 4, with cohort 20 students being the exception, rating it poorly. Most students express high levels of competitiveness in their studies 9.

The table analyzing environmental factors by majors in Table 11 shows that students in Physics Education give considerably more positive evaluations than those in Mathematics Education. With other factors showing fewer notable differences.

Table 10. Statistical analysis of environmental factors by cohort

Cohorts	N	Mean	Std. Deviation
Level of environmental adaptation			
Cohort 14	148	3.62	0.768
Cohort 15	112	3.57	0.887
Cohort 16	101	3.29	0.792
Cohort 17	89	3.09	0.557
Cohort 18	107	2.98	0.879

Cohort 19	66	2.67	0.687
Cohort 20	92	3.16	0.745
Total	715	3.26	0.831
Learning methods			
Cohort 14	148	3.64	0.583
Cohort 15	112	3.46	0.670
Cohort 16	101	3.32	0.631
Cohort 17	89	3.20	0.625
Cohort 18	107	3.33	0.750
Cohort 19	66	3.62	0.627
Cohort 20	92	3.52	0.602
Total	715	3.45	0.658
Suitability of the training program			
Cohort 14	148	4.35	0.479
Cohort 15	112	4.26	0.440
Cohort 16	101	4.64	0.481
Cohort 17	89	4.24	0.853
Cohort 18	107	3.84	1.011
Cohort 19	66	3.33	0.709
Cohort 20	92	2.95	0.790
Total	715	4.01	0.875
Competitiveness in studies			
Cohort 14	148	3.75	0.558
Cohort 15	112	3.77	0.465
Cohort 16	101	3.65	0.591
Cohort 17	89	3.83	0.482
Cohort 18	107	3.40	0.725
Cohort 19	66	3.89	0.500
Cohort 20	92	3.71	0.545
Total	715	3.70	0.578
Level of interest in the field of study			
Cohort 14	148	3.32	0.765
Cohort 15	112	3.32	0.687
Cohort 16	101	3.00	0.663
Cohort 17	89	3.10	0.739
Cohort 18	107	3.09	0.680
Cohort 19	66	3.48	0.588
Cohort 20	92	3.33	0.631
Total	715	3.23	0.705

Table 11. Statistical analysis of environmental factors by field of study

Majors	N	Mean	Std. Deviation
Level of environmental adaptation			
Mathematics	551	3.30	0.738
Physics	164	3.12	1.076
Total	715	3.26	0.831
Learning methods			
Mathematics	551	3.43	0.662
Physics	164	3.50	0.641
Total	715	3.45	0.658
Suitability of the training program			
Mathematics	551	3.90	0.925
Physics	164	4.40	0.527
Total	715	4.01	0.875
Competitiveness in studies			
Mathematics	551	3.74	0.569
Physics	164	3.59	0.595
Total	715	3.70	0.578
Level of interest in the field of study			
Mathematics	551	3.23	0.721
Physics	164	3.24	0.647
Total	715	3.23	0.705

C. LEARNING OUTCOMES

Table 12. Correlation between pre-university academic performance and GPA

Pre-university academic performance	Pearson Correlation	Sig. (2-tailed)	N
Secondary school graduation exam scores	0.005	0.902	715
Mathemetics	0.582**	0.000	715
Literature	0.539**	0.000	715
English	0.472	0.000	715
History	-0.006	0.916	306
Geography	-0.149	0.009	306
Civic Education	0.051	0.377	306
Physics	0.450**	0.000	409
Chemistry	0.108	0.030	409
Biology	0.350	0.776	409
High school graduation exam scores	0.618**	0.000	715
Entrance English score	-0.026	0.482	715

Employing the Pearson Correlation analysis method, the results indicate notable correlations among subjects like Mathematics, Literature, English, Physics, high school graduation exam scores, and GPA. While Chemistry and Geography also show correlations, but not significant, at 10.8% and 14.9%, respectively.

Table 13. Statistical analysis of past academic performance by cohort 7

Cohorts	N	Mean	Std. Deviation
Mathemetics			
Cohort 14	148	6.831	1.1666
Cohort 15	112	6.582	0.9532
Cohort 16	101	5.881	1.1504
Cohort 17	89	5.202	1.2769
Cohort 18	107	5.680	0.9432
Cohort 19	66	6.373	0.7580
Cohort 20	92	7.887	0.5628
Total	715	6.377	1.2865
Literature			
Cohort 14	148	6.6926	1.12262
Cohort 15	112	6.3973	1.30684
Cohort 16	101	5.9604	0.61312
Cohort 17	89	6.0787	0.64026
Cohort 18	107	6.1355	0.87435
Cohort 19	66	6.3917	0.74470
Cohort 20	92	7.1685	0.77132
Total	715	6.4166	1.00609
English			
Cohort 14	148	4.905	1.1057
Cohort 15	112	5.195	0.7685
Cohort 16	101	4.715	0.9232
Cohort 17	89	4.742	1.0154
Cohort 18	107	4.548	0.9906
Cohort 19	66	4.812	1.1308
Cohort 20	92	6.289	0.7131
Total	715	5.019	1.0919
Physics			
Cohort 14	78	5.8526	0.90223
Cohort 15	58	5.4009	0.74330
Cohort 16	54	5.1194	0.83702

Cohort 17	45	4.7944	0.62679
Cohort 18	78	5.5160	1.06474
Cohort 19	42	8.0357	0.51375
Cohort 20	54	7.3935	0.99835
Total	409	5.9388	1.33300
High school graduation exam scores			
Cohort 14	148	18.7111	2.77542
Cohort 15	112	18.4598	2.40109
Cohort 16	101	16.8642	1.98837
Cohort 17	89	16.2831	2.09584
Cohort 18	107	16.8682	2.03752
Cohort 19	66	19.8931	2.69996
Cohort 20	92	21.9695	1.495358
Total	715	18.3612	2.87282

An analysis of the high school academic results indicates a positive trend in students' scores for National High School Examination subjects. This shows a notable enhancement in the number of incoming students at Hanoi Metropolitan University over time. However, English scores are notably lower than those of other subjects, averaging approximately 5.

Table 14. Statistical analysis of students' previous academic performance by major

Majors	N	Mean	Std. Deviation
Mathemetics			
Mathematics	551	6.436	1.3470
Physics	164	6.176	1.0369
Total	715	6.377	1.2865
Literature			
Mathematics	551	6.4698	0.99824
Physics	164	6.2378	1.01477
Total	715	6.4166	1.00609
English			
Mathematics	551	5.054	1.1419
Physics	164	4.902	0.8970
Total	715	5.019	1.0919
Physics			
Mathematics	320	6.0842	1.35172
Physics	89	5.4157	1.12307
Total	409	5.9388	1.33300
High school graduation exam scores			
Mathematics	551	18.542	2.98483
Physics	164	17.752	2.36754
Total	715	18.361	2.87282

According to the analysis table of high school academic results by major, mathematics education majors generally achieve higher average scores than Physics Education majors across all subjects, including junior high school examination scores. However, this variance is relatively minor.

Tables 15 and 16 illustrate that academic performance in most specialized courses is typically strongly correlated, with correlations typically exceeding 45%. The general subjects had a lower correlation compared with the specialized subjects, ranging from approximately 25% to 50%.

However, pedagogical internship courses show notably weak or even nonexistent correlations with GPA. This discrepancy arises because instructors and evaluators of these courses originate from general education institutions. Consequently, assessment scores are frequently inflated to help students, leading to artificially high evaluations.

Table 15. Correlation between the grades of courses for Mathematics Education students and GPA

Subjects of Mathematics Education	Pearson Correlation	Sig. (2-tailed)
Linear Algebra (3)	0.554**	0.000
Analysis 1(3)	0.450**	0.000
Analytic Geometry (2)	0.589**	0.000
Psychology (3)	0.500**	0.000
Philosophy of Marxism-Leninism (2)	0.453**	0.000
Informatics (2)	0.458**	0.000
Analysis 2(3)	0.521**	0.000
The basis of education (3)	0.404**	0.000
Political Economics of Marxism-Leninism (3)	0.392**	0.000
English (5)	0.318**	0.000
Electives (3)	0.267**	0.000
Pedagogical skill 1(2)	0.391**	0.000
History of Vietnamese Communist Party (2)	0.408**	0.000
Abstract Algebra (3)	0.695**	0.000
Analysis 3 (2)	0.548**	0.000
Methodology of Teaching Mathematics (2)	0.631**	0.000
Vietnamese Pratice (2)	0.213**	0.000
Electives (2)	0.224**	0.000
Affine Geometry and Euclid Geometry (2)	0.623**	0.000
Arithmetic (2)	0.495**	0.000
Pedagogical skill 2 (3)	0.508**	0.000
Ho Chi Minh Ideology (2)	0.462**	0.000
Complex function (3)	0.420**	0.000
Projective Geometry (2)	0.591**	0.000
Number Theory (2)	0.532**	0.000
Pedagogical skill 3(2)	0.447**	0.000

Practicing Pedagogy 1(2)	0.283**	0.000
General topology (2)	0.424**	0.000
Primary Algebra (3)	0.531**	0.000
Topology – Measeurement and integrals (2)	0.247**	0.000
Probability and Statistic (3)	0.483**	0.000
Differential Equations (3)	0.424**	0.000
Teaching Method of Mathematics (4)	0.454**	0.000
English for Mathemetics (2)	0.246**	0.000
Funtioncal Analysis (4)	0.567**	0.000
Laws (2)	0.274**	0.000
Partial Differential Equations (3)	0.521**	0.000
Public Administration and Sector Management (2)	0.229**	0.000
Linear Programing (2)	0.466**	0.000
Practicing Pedagogy 2 (3)	0.308**	0.000
Numerical Analysis (2)	0.492**	0.000
Primary Geometry (2)	0.607**	0.000
Teaching mathematics in English (2)	0.297**	0.000
Research Methodology (2)	0.605**	0.000
Electives 7 (2)	0.490**	0.000
Electives 8 (2)	0.39**	0.000
Electives 9 (2)	0.600**	0.000
Practicing Pedagogy 3	0.131**	0.002
Thesis of graduation(8)	0.610**	0.000
Teaching Algebraic divison(3)	0.550**	0.000
Teaching geometry division(3)	0.614**	0.000
Some advanced themes on numerical sequences and functions (2)	0.503**	0.000

Table 15. Correlation between the grades of courses for Physics Education students and GPA

Subjects of Physics Education	Pearson Correlation	Sig. (2-tailed)
General mechanics (3)	0.533**	0.000
Introduction to Earth Science (2)	0.626**	0.000
Psychology (3)	0.536**	0.000
Advanced Mathematics (3)	0.465**	0.000
Philosophy of Marxism-Leninism (3)	0.563**	0.000
Molecular Physics and Thermodynamics (3)	0.221**	0.000
Inform (2)	0.337**	0.000
Electricity and Magnetism (3)	0.578**	0.000

The basis of education (3)	0.625**	0.000
Political Economics of Marxism-Leninism (2)	0.312**	0.000
Pedagogical skill 1 (2)	0.395**	0.000
Electives 2 (2)	0.508**	0.000
English for Physics (2)	0.318**	0.000
Scientific Socialism (2)	0.418**	0.000
Oscillations and Waves (2)	0.422**	0.000
Pedagogical skill 2 (3)	0.433**	0.000
Optics (3)	0.542**	0.000
Mathematics for Physics (3)	0.484**	0.000
Atomic and Nuclear Physics (2)	0.504**	0.000
Electives 3 (2)	0.432**	0.000
Vietnamese Pratice (2)	0.527**	0.000
Electrical Engineering (3)	0.482**	0.000
Physics Pedagogy (3)	0.600**	0.000
Pedagogical skill 3 (2)	0.492**	0.000
General Physics Laboratory(2)	0.555**	0.000
Practicing Pedagogy 1 (2)	0.254**	0.000
Computer Science for Physics (2)	0.436**	0.000
Ho Chi Minh Ideology (2)	0.506**	0.000
History of Vietnamese Communist Party (2)	0.649**	0.000
High School Physics Experiments (3)	0.624**	0.000
Design and organization of students' project activities in physics lessons (4)	0.658**	0.000
Solid State Physics (2)	0.382**	0.000
Electives 5 (2)	0.271**	0.000
Electives 6 (2)	0.557**	0.000
Teaching Physics in Enligsh 2	0.499**	0.000
Guidance for organizing Experimental Activities (2)	0.681**	0.000
Laws (2)	0.209**	0.000
Methods for Solving High School Physics Problems (3)	0.601**	0.000
Practicing Pedagogy 2 (3)	0.322**	0.000
Public Administration and Sector Management (2)	0.253**	0.001
Electives 7 (2)	0.098	0.211
Quantum Mechanics (2)	0.596**	0.000
Electronic Engineering (3)	0.602**	0.000
Assessment and Evaluation in Physics Education (2)	0.581**	0.000
History of Physics (2)	0.662**	0.000
Development of the High School Physics Curriculum (2)	0.492**	0.000
Research Methodology (2)	0.529**	0.000

Astronomy (2)	0.252**	0.001
Electives 8 (2)	0.668**	0.000
Thesis of graduation (8)	0.596**	0.000
Practicing Pedagogy 3 (4)	-0.017	0.830
Modern Teaching Methods in Physics Education (3)	0.608**	0.000
Guidance on Integrated Topic-based Teaching in Natural Science (3)	0.514**	0.000
Theoretical Physics (2)	0.556**	0.000

The statistical discrepancy is clearly evident in elective courses. This variation arises because some of these courses are closely related to specialized subjects, whereas others cover diverse fields such as music, the arts, economics, data analysis, and history.

5. Discussion

In this study, we compile data from nearly a thousand students, incorporating their academic performance, demographics, and influential factors in student outcomes to construct an educational dataset comprising 89 key factors. Drawing insights from this dataset, several conclusions can be drawn. Notably, learning methods emerged as the most significant determinant of student academic achievement, while the level of school support and program suitability exhibited an inverse correlation, reflecting students' learning trajectories. Emphasizing infrastructure improvements and innovative training programs aligned with the advancements of the 4.0 technology era are highlighted, with a particular focus on enhancing students' English language proficiency to meet industry demands and foster educational sector integration and development.

Moreover, this study emphasizes the importance of support and collaboration among educational stakeholders to enhance teaching and learning quality. Leveraging digital transformation, the integration of technology for decision support is advocated to assist managers and education professionals in planning and directing learning initiatives. Using data mining tools and analyzing training datasets to inform enrollment decision-making is deemed essential in the current educational landscape.

The Pearson correlation analysis will quantify the impact of various factors on academic outcomes, enabling the identification of those with strong correlations for targeted improvement and development. Additionally, this analysis will address and help to mitigate the influence of factors that negatively affect students' learning processes.

The primary goal of this paper is to advance educational methodologies, assist students in making informed decisions, and improve overall learning effectiveness. Comprising 992 samples across 89 fields, the dataset was collected through direct methods such as questionnaires and indirect methods via training management units. It is organized into categories including Personalized Factors (Group A), Factors Affecting Learning Outcomes (Group B), and Learning Outcomes (Group C), covering both general academic performance and specific university module achievements. The study spans a wide array of factors, from personal, familial, and social contexts to environmental influences and academic history. Its aim is to compile a comprehensive, multidimensional dataset that captures the diverse elements impacting university students' academic outcomes. Such a dataset will serve as a valuable resource for advanced research, enabling deeper analysis into the learning process and pinpointing key factors that influence student success. This research also has practical applications, particularly for machine learning, deep learning, and statistical algorithms. These tools

can analyze the data to predict academic outcomes, issue early warnings for dropouts, and identify high-performing students for future opportunities. Furthermore, this study can assist in shaping policies for better student management, improving teaching strategies, and developing more effective higher education systems. The insights gained can also support students, educators, administrators, and employers in making informed decisions for personal development, educational adjustments, and workforce planning.

After collection, the dataset underwent rigorous processing, cleaning, and statistical analysis using techniques like Pearson correlation analysis, analysis of variance, standard error, standard deviation, and tests of homogeneity of variance.

By applying preliminary statistical analysis methods, the study provides an overview of the general context, advantages, and challenges in the teaching and learning processes of the Mathematics Education and Physics Education programs in the Faculty of Education, HNMU.

The variables in group A include personal information such as gender fields of study. We further investigated the influence of parental education levels on student's academic performances. Based on the Pearson Correlation analysis, parents' education level was found to be fairly strongly correlated with students' academic performances.

The variables in Group B investigate the factors influencing students' academic performance, categorized into two main groups: personal and environmental. For personal factors, study time, social media usage, admission methods, and scholarships significantly impact academic outcomes, with a notable correlation of 0.604 between study time and GPA. Among environmental factors, university learning methods had the most substantial effect on students' graduation results, boasting a Pearson correlation of 0.677. This underscores the importance of developing an effective and personalized study strategy, complemented by a well-structured, balanced schedule. Prioritizing tasks, adhering to a consistent plan, and regularly evaluating progress can greatly enhance academic performance.

However, the suitability of the training program exhibited a negative correlation. Statistical data reveals that while the first four cohorts rated the program highly, with an average Likert score above 4, the subsequent three cohorts expressed growing dissatisfaction, with Cohort K20 rating it at just 2.95. This indicates that the program has struggled to keep pace with the evolving demands of the current and future job market. To address the ongoing needs of modern society, the university must undertake substantial curriculum reforms. We recommend integrating new skills and technological trends into the curriculum. Such updates will equip students with essential professional skills like creative thinking, effective communication, and teamwork-key components for success in today's educational landscape.

Lastly, the variables in Group C focus on academic outcomes, including pre-university performance and university course grades. Analysis reveals that the English proficiency of students in the Mathematics and Physics Education programs has remained stagnant, with average scores hovering around 5 across seven cohorts. Therefore, the university must prioritize enhancing students' language skills to meet employment demands and support the education sector's integration and development.

The research faced several key challenges that impacted the depth and accuracy of its findings. Retrieving accurate information from graduates, especially those who completed their studies years ago, was particularly difficult. Many struggled to recall specific details such as high school exam scores, and some were reluctant to disclose their academic records, resulting in gaps and potential inaccuracies. The quality of survey responses also varied, with inconsistencies and incomplete answers requiring substantial data cleaning efforts. Sparse data, especially regarding pre-university performance, necessitated the use of estimated values, introducing the risk of bias. Moreover, aligning survey data with official academic

records was a complex task, as discrepancies between these sources required meticulous cross-referencing through ID students to ensure accurate correlations. Ensuring consistency across different student cohorts, each with unique academic experiences and external factors, further complicated the analysis. For example, the grading scale for middle school graduation exams varies by year, so we converted the scores from all years to a 10-point scale. Despite these obstacles, the research team adhered to a rigorous methodology, recognizing that openly addressing these limitations is essential for producing findings that are both credible and relevant to real-world educational practices.

6. Conclusion

In conclusion, this paper introduces a comprehensive educational dataset that integrates various facets of educational science to better understand and enhance students' learning experiences both before and during their university studies. Designed to facilitate research in educational science, the dataset supports the application of machine learning and deep learning models for predicting student outcomes.

The dataset established in this study serves as a crucial and reliable resource for relatively limited data repositories in educational sciences. However, basic statistical methods alone may not be sufficient to uncover the deeper, latent insights within the data. In the next phase of our research, we plan to incorporate machine learning and deep learning models to enhance predictive accuracy and detect more complex patterns. These advanced techniques not only improve the precision of our predictions but also allow us to explore nonlinear and multidimensional relationships that traditional statistical methods might overlook. The overarching objective is to develop models capable of analyzing, predicting, and providing insights into the relationship between influencing factors and student learning outcomes. This signifies an initial step toward a growing research trend-the application of artificial intelligence (AI) in education.

Acknowledgements

We are deeply grateful to all the students who participated in this study, as well as the teachers and staff of Hanoi Metropolitan University who supported the distribution of the questionnaires.

Conflict of Interest

The authors state that none of their known financial conflicts or interpersonal connections could have influenced the work that was published in this paper.

References

[1] A. B. Zorić, Benefits of Educational Data Mining, Journal of International Business Research and Marketing, Polytechnic Baltazar Zaprešić, Zaprešić, Croatia, Vol. 6, Issue 1, 2020, https://doi.org/10.18775/jibrm.1849-.2015.61.3002 (accessed on: February 28th, 2023).

[2] B. T. Dien, N. T. Thuy, D. T. T. Hue, P. T. T. Trang, Online Learning Experiences of Secondary School Students During COVID-19 - Dataset from Vietnam, Vol. 45, 2022, https://doi.org/10.1016/j.dib.2022.108662 (accessed on: February 28th, 2023).

[3] C. Marcus, R. K. Nathan, Study Habits, Skills, and Attitudes: The Third Pillar Supporting Collegiate Academic Performance, Vol. 3, Issue 6, 2008, https://doi.org/10.1111/j.1745-6924.2008.00089.x (accessed on: August 5th, 2022).

[4] E. G. Sarmiento, J. R Perez, L. O. Ospino, Big Data and Artificial Intelligence in the Development of Industry 4.0; A Bibliometric Analysis, IOP Conference Series Materials Science and Engineering, 2021, https://doi.org/10.1088/1757-899X/1154/1/012008 (accessed on: February 28th, 2023).

[5] J. F. Superby, J. P. Vandamme, N. Meskens, Determination of Factors Influencing the Achievement of the First-year University Students using Data Mining Methods, Workshop on Education, 2006 (accessed on: August 7th, 2022).

[6] K. Misiejuk, M. Khalil, B. Wasson, Tackling the Challenges with Data Access in Learning Analytics Research: A Case Study of Virtual Labs, Proceedings of the Technology-Enhanced Learning in Laboratories Workshop (TELL 2023), April 27, 2023, https://ceur-ws.org/Vol-

3393/TELL23_paper_9675_6.pdf, 2023 (accessed on: March 3th, 2023).

[7] K. Taylor, Machine Learning vs. Artificial Intelligence: Which is the Future of Data Science?, May 5, 2021 in Machine Learning, Artificial Intelligence, Contributors, Data Science, Data Conomy, 2021.

[8] M. Irfan, S. N. Khan, Factors Affecting Student Academics's Performance, 2012.

[9] M. S. Farooq, A. H. Chaudhry, M. Shafiq, G. Berhanu Factors Affecting Students' Quality of Academic Performance: A Case of Secondary School Level, Journal of Quality and Technology Management, Vol. 7, 2011, pp. 1-14.

[10] N. O. Omarova, A. A. Echilova, Big Data Technologies in the Education System, Computational and Strategic Business Modelling, 2024, pp. 567-577, https://doi.org/10.1007/978-3-031-41371-1_47 (accessed on: February 28th, 2023).

[11] S. Ali, H. Zubair, M. Fahad et al., Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus, American Journal of Educational Research, Vol. 1, 2013, pp. 283-289.

[12] S. P. Singh, M. Savita, S. Priya, Factors Affecting Academic Performance of Students, Indian Journal of Research, ISSN - 2250-1991, Vol. 5, Issue 4, 2016.

[13] T. Hartati, N. Fitria, M. A. A. Harahap, D. Dasari, 2023, Data-Driven Education: Data Processing as a Key to Improving the Quality of Mathematics Education, ALSYSTECH Journal of Education Technology, Vol. 2, No. 1, 2023, https://doi.org/10.58578/alsystech.v2i1.2361 (accessed on: March 3th, 2023).

[14] T. Trung, H. D. Anh, N. T. Trung, D. V. Hung, N. Y. Chi, P. H. Hung, 2020, Dataset of Vietnamese Student's Learning Habits During COVID-19, Vol. 30, 2020, https://doi.org/10.1016/j.dib.2020.105682 (accessed on: March 3th, 2023).

[15] W. Lan, R. Lanthier, Changes in Students' Academic Performance and Perceptions of School and Self Before Dropping Out of Schools, Journal of Education, 2003.

[16] N. T. T. An, N. T. N. Thu, D. T. K. Oanh, N. V. Thanh, Factors Affecting the Academic Performance of First- and Second-year Students at Can Tho University of Technology, Journal of Science, Can Tho University, Part C: Social Sciences, Humanities and Education, Vol. 46, 2016, pp. 82-89 (in Vietnamese).

[17] V. V. Kiet, D. T. T. Phuong, Research on the Key Factors Affecting Students' Academic Performance, Journal of Science, VNU: Educational Research, Vol. 3, No. 3, 2017, pp. 7 (in Vietnamese).

[18] N. T. N. Quynh, Research on Factors Affecting the Active Learning Method Results of 16DDS Pharmacy Students, Nguyen Tat Thanh University, Journal of Science and Technology, Nguyen Tat Thanh University, 2020 (in Vietnamese).