VNU Journal of Science: Education Research

Journal homepage: https://js.vnu.edu.vn/ER

Original Article

Analysis of Education Data in Vietnam (2002–2023) and Student Enrollment Projections to 2030

Vu Thao Nguyen[1,*], Trinh Bui Quang Minh[2], Tran Gia Khanh[3]

[1]*Nanyang Technological University - 50 Nanyang Avenue, Singapore*
[2]*Hanoi - Amsterdam High School for the Gifted, 1 Hoang Minh Giam, Yen Hoa, Hanoi, Vietnam*
[3]*Hanoi National University of Education High School for the Gifted, 136 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Received 10th September 2026
Revised 03rd November 2026; Accepted 10th November 2026

Abstract: In response to the need for the digitalization of educational management processes, this paper addresses the analysis of education data. Based on a dataset covering student enrollment in primary, lower secondary, and upper secondary schools across provinces and cities in Vietnam from 2002 to 2023, the study employs descriptive, diagnostic, and predictive analytical approaches. For predictive analysis, the paper utilizes linear regression and the Prophet forecasting model - both integral components of a data science workflow. The analysis and projections are grounded in empirical data from the Vietnamese education sector, sourced from the General Statistics Office of Vietnam. Data processing and forecasting were conducted using Jupyter notebooks within the Google Colab environment, incorporating various machine learning models. The study provides enrollment projections for provinces and cities across Vietnam up to the year 2030.

Keywords: Education, Statistical analysis, Data science, Regression, Machine learning.

1. Introduction

Education data refers to information concerning students at levels below college or university. Statistical data is essential for conducting quantitative statistical analyses. Although statistics is often regarded as the science of the 21st century [1], the application of inferential statistics to education data in Vietnam remains limited. While descriptive statistics are utilized, they are not yet widely applied.

In the context of digital transformation and "smart education", the demand for statistical analysis in the educational sector has become increasingly critical for supporting data-driven decision-making. This is because statistical analysis helps evaluate the quality of teaching and learning, monitor student academic performance, and identify factors that influence learning outcomes. Furthermore, statistical methods are essential for supporting: i) The forecasting of educational development trends; ii) Resource allocation; and ii) The formulation of appropriate policies. Currently, data analysis and machine learning tools applied to educational data enable enhanced management efficiency, the personalization of the learning

* Corresponding author.
E-mail address: vuthaonguyen26052005@gmail.com

https://doi.org/10.25073/2588-1159/vnuer.5597

process, and improvements in the quality of education.

This paper addresses three main topics: educational datasets in Vietnam from 2002 to 2023; the steps involved in analyzing educational data; and a forecasting task based on the dataset - specifically, predicting the number of students in each province by 2030. These topics fall within the scope of data science. Data science encompasses multiple phases, ranging from data collection and preprocessing to model application. The processing of educational data should follow these phases of data science.

2. Research Context

2.1. Related Research

Numerous studies have addressed statistical analysis, highlighting its role, processes, and applications in relation to datasets. In a study on interactive exploratory data analysis systems using interpretable models, E. Kalita et al., [2] observed that complex tabular datasets with diverse characteristics often require specialized expertise to interpret, thereby placing a significant burden on educational administrators. They addressed the need to simplify the initial analysis and visualization of these datasets, providing insights into data quality and the relationships between features. In another publication [3], E. López-Meneses and colleagues demonstrated the role of statistical analysis and visualization - specifically through various types of charts - in exploratory data analysis. Correlation plots proved suitable for examining large datasets dominated by quantitative features, while also clarifying the relationships between categorical variables in smaller datasets.

While the studies in [2, 3] utilize the R programming language, the following section will employ Python within a machine learning framework. The procedures proposed by the previous authors will be revisited and illustrated using a Vietnamese educational dataset.Chuah et al., [4] evaluated the need for data mining tools over the past decade and discussed the applications and impact of information technology in education. It is undeniable that educational administrators need a clear understanding of students' learning experiences and the effectiveness of educational activities. In addition to the role of data - a topic to be addressed later - the authors also touched upon issues of privacy and security; however, these subjects fall outside the scope of the current study on educational data in Vietnam. Regarding the predictive tasks to be presented later, beyond traditional mathematical models such as regression, the application of educational data mining - facilitated by artificial intelligence - has become essential in the digital age. According to S. El-Kamand et al., [5] - who conducted a review of nearly 800 articles published between 2000 and 2024, indexed in the Scopus database - information technology can be leveraged to predict student academic outcomes, support personalized learning, and enable timely interventions through the analysis of student data. Notably, research findings indicate that artificial intelligence has the potential to significantly improve educational outcomes, while also optimizing student monitoring, resource allocation, and the overall operational efficiency of educational institutions.

Studies [4, 5] have highlighted the processes involved in handling educational data, as well as the benefits these processes offer to students, teachers, and educational institutions. The remainder of this article applies this research methodology: first, by analyzing the data, and subsequently, by utilizing artificial intelligence tools to generate forecasts.

2.2. Educational Dataset

This article employs an educational dataset available on the official Vietnamese government portal at https://www.nso.gov.vn/ [6]. The data includes the number of K-12 students as of September 30th of each year, disaggregated by locality. Currently, the data for the period from 2002 to 2023 is organized

according to the historical provincial and municipal boundaries, rather than the contemporary two-tier administrative structure. Regarding Vietnam's two-tier local government model (comprising the provincial and commune levels, with the district level removed): i) 2025 marks the point at which this concept is formally institutionalized within the Vietnamese legal system, following the National Assembly's adoption of amendments to the 2013 Constitution; and ii) The Law on the Organization of Local Government, enacted in 2025, stipulates that administrative units shall be organized into two tiers. The dataset excerpt is illustrated in Figure 1.

	Khu vực	Tổng 2002	Tiểu học 2002	Trung học cơ sở 2002	Trung học phổ thông 2002	Tổng 2003	Tiểu học 2003	Trung học cơ sở 2003	Trung học phổ thông 2003	Tổng 2004
0	Đồng bằng sông Hồng	3854343	1668289	1525283	660771	3770728	1585341	1509350	676037	3688941
1	Hà Nội	503132	219925	180226	102981	497298	217863	179351	100084	495941
2	Hà Tây	544383	231368	210165	102850	535666	220645	207855	107166	525427
3	Vinh Phúc	256114	112442	101372	42300	252047	106120	100267	45660	245742
4	Bắc Ninh	229896	98905	88251	42740	226978	93905	88676	44397	223535

Figure 1. Dataset from the General Statistics Office (viewed in Google Colab).

The dataset is provided as an Excel file (*.xlsx), where rows represent localities, and columns represent characteristics based on: i) Total number of students; ii) Primary school students; iii) Lower secondary school students; and iv) Upper secondary school students [6]. The dataset contains 70 local units, corresponding to the rows. Each row consists of 89 features, corresponding to the 89 columns in the spreadsheet. While processing the data using Excel functions is one possible approach, this article employs Exploratory Data Analysis (EDA) [7] via the Python programming language within the Google Colab Jupyter environment [8].

To focus specifically on individual provinces and cities, the dataset excludes broader regional or zonal units. A total of 64 provinces and cities are considered. The list of localities and dataset variables is shown in Figures 2 and 3, respectively.

To clarify the phases of educational data processing, this paper addresses the Exploratory Data Analysis (EDA) phase, which includes data cleaning. The modeling phase focuses on developing a predictive model. Based on an examination of the dataset and its underlying trends, this paper employs linear regression. However, as noted in [10], the Prophet regression model is also suitable for educational datasets.

The subsequent sections of this paper present the results of statistical analysis on the educational dataset, along with forecasts generated using both the linear regression model [9] and the Prophet regression model [10].

Structure of the paper: Section 2 covers the descriptive statistical parameters of the dataset and the regression models used for forecasting; Section 3 details the experiments, providing quantitative results and visual illustrations. The paper concludes with a summary and a list of references.

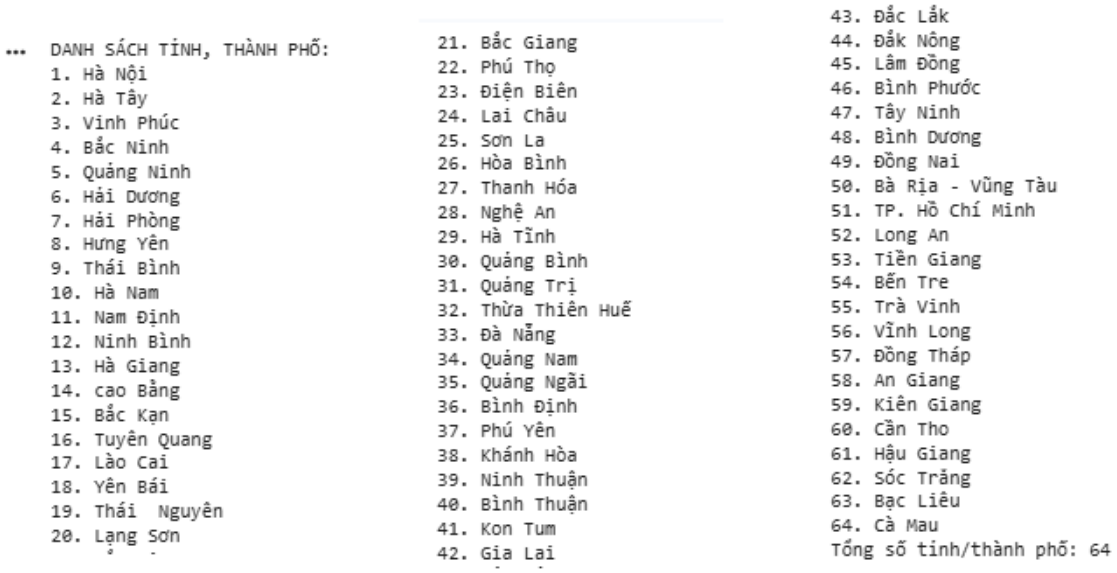

Figure 2. Names of provinces and cities.



Figure 3. Characteristics of provinces and citie.

2.3. Aspects of Exploratory Data Analysis (EDA)

For an educational dataset categorized by province/city and year, the EDA focuses on the following descriptive statistical components:

i) Data scale and structure: Determining the number of records, the attributes, and the data type of each variable;

ii) Measures of central tendency: Calculating the mean, median, and mode to identify typical values;

iii) Measures of dispersion: Calculating the standard deviation, variance, minimum and maximum values, and the range to assess the spread of the data;

iv) Data distribution: Utilizing histograms, boxplots, and metrics such as skewness and kurtosis to examine the shape of the distribution and the degree of asymmetry in the data;

v) Outlier detection: Use boxplots or the IQR (Interquartile Range) method to identify and examine anomalous values;

vi) Missing data analysis: Calculate the count and proportion of missing values for each attribute to inform decisions regarding data removal, imputation, or interpolation;

vii) Correlation analysis: Utilize Pearson and Spearman correlation coefficients, along with heatmaps, to assess relationships between variables and support feature selection;

viii) Time-trend analysis: Use line charts to track changes in educational indicators over the years, identifying upward, downward, or stable trends;

ix) Group comparisons: Compare data across provinces, regions, and educational levels using bar charts, boxplots, or statistical tables;

x) Data visualization: Employ histograms, boxplots, bar charts, line charts, scatter plots, and heatmaps to illustrate distributions, trends, and relationships between variables.

While the aspects mentioned above have not yet been detailed, Section 3 will explore these key areas in depth.

2.4. Regression and Forecasting

Regression is a statistical and machine-learning method used to model the relationship between a dependent variable and one or more independent variables, with the aim of explaining and predicting the value of the dependent variable. By constructing models based on observed data, regression helps estimate trends and evaluate the extent to which various factors influence the outcome under study. The need for and methods of forecasting in education are also addressed in [11, 12].

The role of regression in forecasting: i) Regression enables the prediction of a target variable's value based on the relationships identified within historical data, thereby supporting the forecasting of future trends; and ii) In the field of education, regression can be used to forecast the number of students, teachers, or educational institutions, as well as academic outcomes, thereby supporting planning and management decision-making.

The objective of linear regression is to construct a linear function that describes the relationship between input and output variables, while also predicting the value of a dependent variable [9]. A simple linear regression model takes the form: $y = aX + b$, where X is the input variable, a and b are coefficients, and y is the output variable. Here, a is referred to as the regression coefficient, and b is known as the intercept.

Linear regression, which has been in use since 1885, serves several purposes: i) Forecasting future values; ii) Assessing the impact of independent variables on a dependent variable; iii) Analyzing data trends; and iv) Supporting decision-making across fields such as economics, education, healthcare, finance, and data science. For instance, by using annual educational data - with the year (covering the 2002 - 2023 period) as the independent variable and the total number of schools as the dependent variable - a model can

forecast the number of schools for the 2024 - 2030 period and evaluate the growth rate of the educational system. Regarding its role in forecasting: i) This method helps predict the value of a target variable based on relationships identified within historical data, thereby facilitating the projection of future trends; and ii) In the field of education, this method can be used to forecast student enrollment, the number of teachers, the number of educational institutions, or academic outcomes, thereby supporting planning and management decision-making.

The Prophet regression model, introduced in 2017, is designed to forecast future time-series values based on trends and patterns derived from historical data [10]. Developed by Meta, Prophet is a time-series forecasting model built to handle data that contains trends accounting for trends over time, seasonality, and cyclical fluctuations. This model enables accurate forecasting even when data is missing or when there are points where the trend shifts.

The role of Prophet: i) It assists in forecasting future trends based on historical data, thereby supporting planning and decision-making through reliable projections; and ii) Using such datasets, Prophet can be applied to forecast annual figures - such as the number of students, teachers, schools, or specific educational targets - which facilitates effective planning and resource allocation

2.5. Preparing Educational Data for Machine Learning Tasks and Tools

The dataset collected from https://www.nso.gov.vn/ [6] has been adjusted to facilitate statistical analysis. For use with machine learning tools, files in *.xlsx format can be converted into *.csv format. Several preprocessing steps, in preparation for the application of machine learning models: such as information about the dataset, checking for missing data, data transformation, removing duplicate data, etc.

	count	mean	std	min	25%	50%	75%	max
Tổng 2002	65.0	302008.	194931.	74464.0	191343.0	256413.5	349369.0	1167217.0
Tiểu học 2002	65.0	151902.	99755.	35632.0	98805.0	129131.5	162834.0	670546.0
Trung học cơ sở 2002	65.0	108806.	71790.	28399.0	71034.0	88743.0	124148.0	376454.0
Trung học phổ thông 2002	65.0	41051.	29060.	8641.0	23687.0	33177.5	49031.0	143291.0
Tổng 2003	65.0	299158.	195030.	73381.0	187394.0	247467.0	341179.0	1201636.0
...	...			...	...	...	...	...
Trung học phổ thông 2022	65.0	49660.	48896.	8166.0	27479.0	38026.5	49538.0	294150.0
Tổng 2023	65.0	309255.	285031.	61072.0	174570.0	233761.5	298964.0	1700638.0
Tiểu học 2023	65.0	145509.	132132.	28874.0	82799.0	108172.5	139139.0	762387.0
Trung học cơ sở 2023	65.0	111462.	101930.	23704.0	65182.0	82627.0	109212.0	623767.0
Trung học phổ thông 2023	65.0	52232.	51877.	8494.0	28649.0	39663.0	52837.0	314484.0

Figure 4. Descriptive statistics for the dataset.

Python is the language used for these experiments, utilizing Jupyter Notebooks within the Google Colab environment [8]. Google Colab provides an online programming environment that allows users to create and execute Jupyter Notebooks directly in a web browser, eliminating the need for software installation. The platform supports Python along with numerous libraries for data science and machine learning, such as NumPy, Pandas, Matplotlib, and Scikit-learn. Furthermore, Google Colab integrates with Google Drive and provides access to GPUs and TPUs, facilitating the efficient processing of data analysis and artificial intelligence tasks. Due to its convenience, free accessibility, and ease of

sharing, Google Colab is widely used in education, research, and the development of data science applications. Descriptive statistics for the dataset are summarized in Figure 4.

Additional statistics include the mean number of students and the standard deviation over the years, as well as comparisons of student numbers across different localities, measured by: i) Absolute increase; and ii) Growth rate. These results are presented in the form of charts. The program also examines statistical values for each region in Vietnam. Annual means and standard deviations, changes in enrollment, and the ten localities with high growth are shown in Figures 5, 6 and 7, respectively.

Jupyter's graphical tools enable the visual representation of data. Graphical displays help users visualize the data, highlighting trends, patterns, and distributional characteristics that might be difficult to discern from a raw data table alone. Charts also assist in identifying outliers, correlations between variables, and changes in the data over time. Consequently, the processes of analysis, result interpretation, and decision-making become faster, more accurate, and easier to understand.

	Năm	Mean	Std
0	2002	302008	194931
1	2003	299158	195030
2	2004	282049	197856
3	2005	274647	194533
4	2006	268652	192777
5	2007	259657	189282
6	2008	253781	202910
7	2009	250082	201607

	Năm	Mean	Std
8	2010	248094	203247
9	2011	247938	204756
10	2012	247272	206049
11	2013	249812	210277
12	2014	252581	215310
13	2015	256820	220832
14	2016	259227	225848

	Năm	Mean	Std
15	2017	265819	233797
16	2018	274851	247186
17	2019	283915	255822
18	2020	292525	265624
19	2021	298403	272501
20	2022	306875	280919
21	2023	309255	285031

Figure 5. Mean number of students and standard deviation.

	Khu vực	2002	2023
0	Hà Nội	503132.0	1700638.0
1	Hà Tây	544383.0	233761.5
2	Vinh Phúc	256114.0	258190.0
3	Bắc Ninh	229896.0	252360.0
4	Quảng Ninh	229780.0	276308.0

	Khu vực	2002	2023	Increase
0	Hà Nội	503132.0	1700638.0	1197506.0
1	Hà Tây	544383.0	233761.5	-310621.5
2	Vinh Phúc	256114.0	258190.0	2076.0
3	Bắc Ninh	229896.0	252360.0	22464.0
4	Quảng Ninh	229780.0	276308.0	46528.0

Figure 6. Growth rate.

	Khu vực	2002	2023	Increase	Growth %
0	Hà Nội	503132.0	1700638.0	1197506.0	238.010303
48	Bình Dương	151654.0	407653.0	255999.0	168.804647
51	TP. Hồ Chí Minh	870228.0	1370911.0	500683.0	57.534692
32	Đà Nẵng	158695.0	225394.0	66699.0	42.029680
12	Hà Giang	151931.0	202320.0	50389.0	33.165713
41	Kon Tum	98320.0	128759.0	30439.0	30.959113
42	Gia Lai	264382.0	338238.0	73856.0	27.935336
49	Đồng Nai	487424.0	603386.0	115962.0	23.790786
24	Sơn La	231848.0	284887.0	53039.0	22.876626
50	Bà Rịa - Vũng Tàu	197200.0	239963.0	42763.0	21.685091

Figure 7. Ten localities with high growth in student numbers.

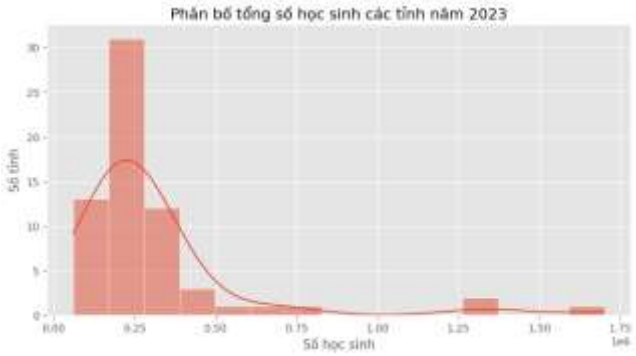

Figure 8. Student distribution.

The distribution shown in Figure 8 indicates the following:

i) Most provinces/cities have a low to moderate number of students. The highest concentration of provinces falls within the range of 100,000–400,000 students, with the 200,000–300,000 range showing the highest frequency. This suggests that the majority of localities have a student population that falls within a relatively specific, concentrated range;

ii) The distribution is right-skewed. This is a notable characteristic; most provinces have a relatively low number of students, while only a small number have a very large student population. Statistically, in a right-skewed distribution, the mean is typically greater than the median;

iii) There is a significant disparity between localities. The chart reveals that some provinces have a student population far exceeding that of most others - specifically in the 1.3–1.7 million range. These observations warrant separate consideration during data analysis;

iv) The data does not follow a normal distribution. This must be taken into account when selecting statistical methods and developing analytical models;

v) Practical implications for educational data: The chart suggests several observations: i) Most provinces have a moderate number of students, with the scale of education being relatively similar across many localities; ii) A few provinces have a very large number of students, implying a potentially higher demand for schools, classrooms, and teachers; and iii) The right-skewed distribution indicates an uneven distribution of student populations across different localities. The box plot and comparison across localities are presented in Figures 9 and 10, respectively.

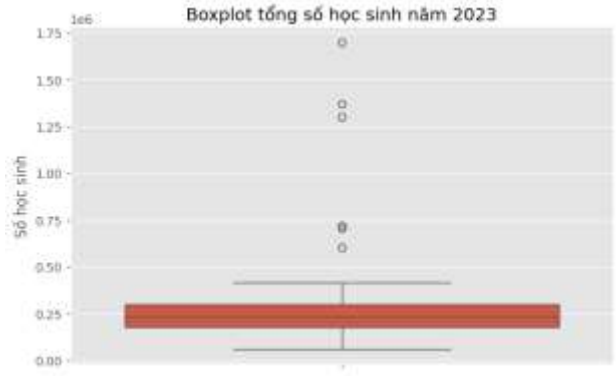

Figure 9. Box plot showing data anomalies.

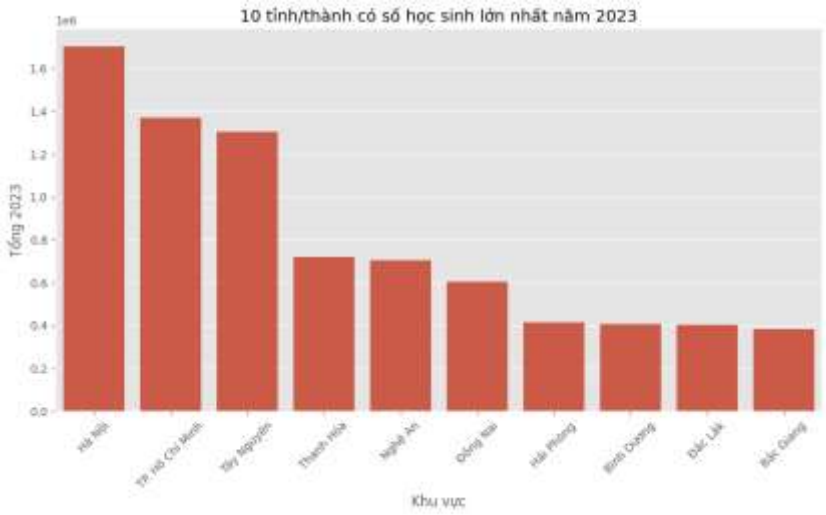

Figure 10. Comparison between localities.

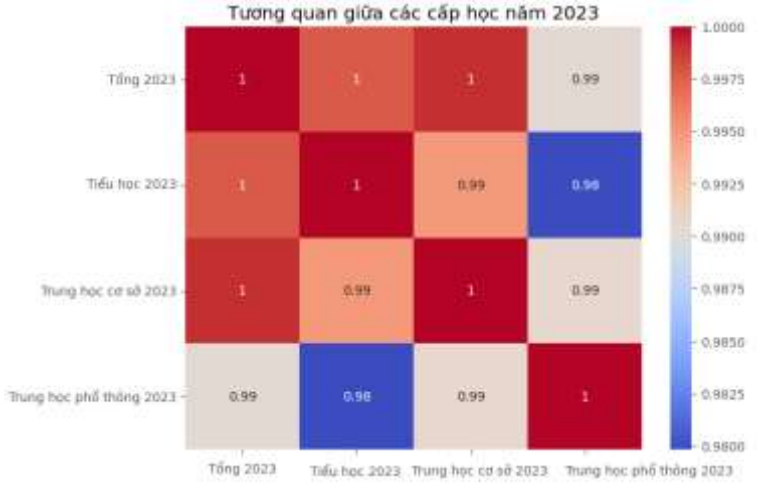

Figure 11. Correlation between educational levels.

The chart in Figure 11 shows that:

i) There is a very strong positive correlation between educational levels, with all correlation coefficients ranging from 0.98 to 1.00. This indicates that provinces with a large number of primary school students also tend to have a large number of students at the lower secondary and upper secondary levels;

ii) The correlation between primary education and upper secondary education is the lowest (0.98). While still very strong, this suggests some variation in the relative scale of these two levels across different localities;

iii) There is a very high correlation (0.99) between lower secondary and upper secondary education, indicating that student enrollment at these two levels tends to fluctuate in the same direction across provinces and cities;

iv) The total number of students shows an almost perfect correlation with each individual educational level. This is consistent with the nature of the data, as the total number of students is the sum of students across the constituent educational levels. The national enrollment trend over time is illustrated in Figure 12.

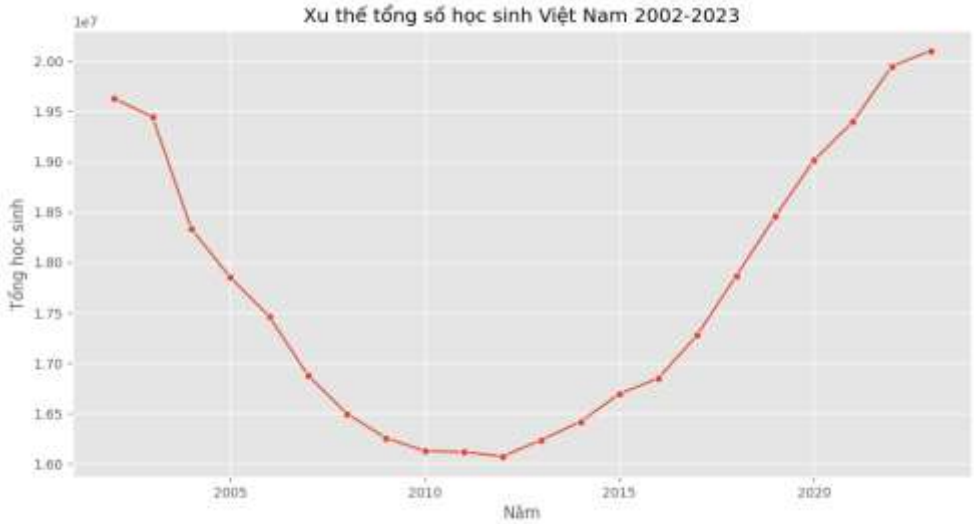

Figure 12. Trend in total number of students.

Practical implications for educational data: These results can be interpreted as follows: i) More populous localities generally have larger student populations across all three educational levels; ii) Student enrollment across different educational levels tends to vary in the same direction across localities; iii) Differences in population size and demographic distribution may help explain these correlations; and iv) These findings can be used to examine the relationship between the scale of different educational levels and the geographical distribution of students. Trends by educational level and the 2023 enrollment composition are presented in Figures 13 and 14, respectively.

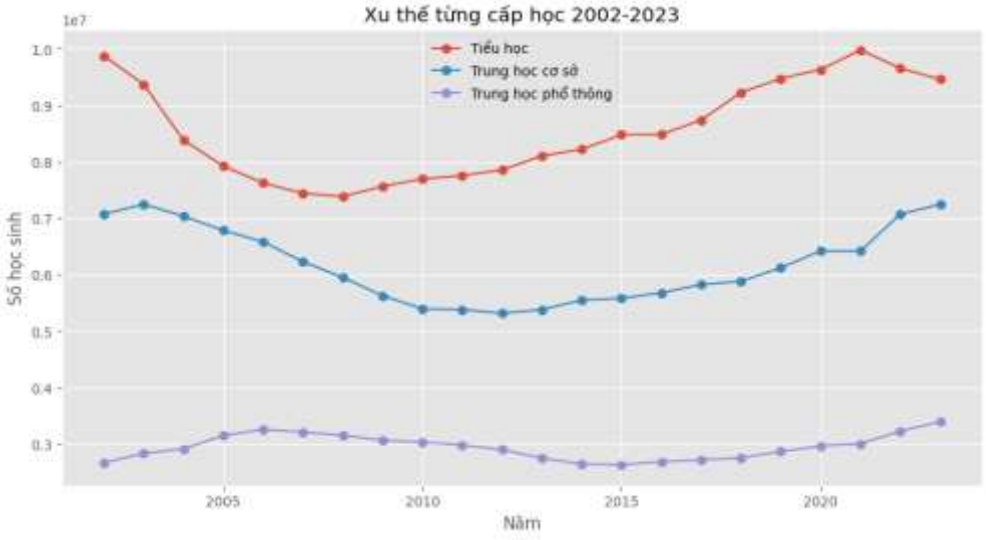

Figure 13. Trend by each educational level.

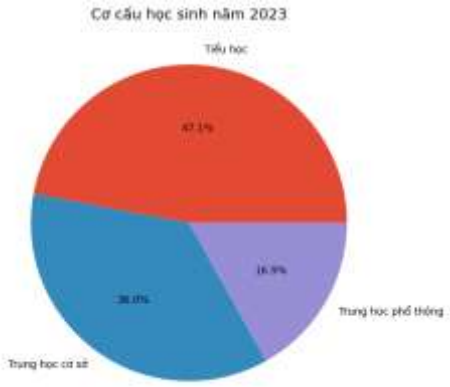

Figure 14. Composition of students by educational level.

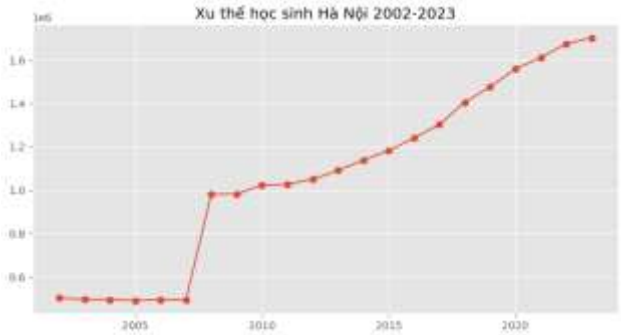

Figure 15. Trend for Hanoi specifically.

The trend line in Figure 15 shows: i) For the 2002–2007 period, the trend line remains nearly flat, stabilizing at approximately 500,000 students; ii) From 2007 to 2008, there is a sharp, sudden increase from about 500,000 to nearly 1,000,000 students (likely due to the expansion of Hanoi's administrative boundaries in 2008); and iii) For the 2008–2023 period, the trend line shows a consistent and strong year-on-year increase, rising from under 1 million students in 2008 to approximately 1.7 million students by 2023.

Data clustering results, obtained using the K-means algorithm [7], reveal the existence of three clusters. The number of clusters was determined using the "Elbow Method." Currently, the 64 provinces and cities are classified into clusters 0, 1, and 2. Figure 16 shows the Elbow Method results, while Figure 17 presents the resulting clusters.

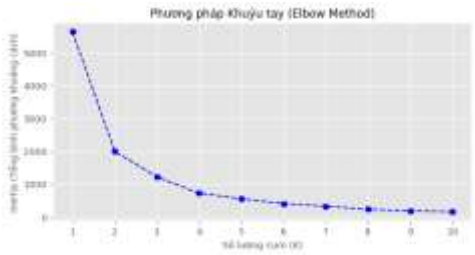

Figure 16. Results of the elbow algorithm.

```
=== THỐNG KÊ SỐ LƯỢNG TỈNH THEO CỤM ===
Cum_KMeans
0    58
1     2
2     4
Name: count, dtype: int64

📍 Cụm 0 (58 tỉnh/thành):
Hà Tây, Vĩnh Phúc, Bắc Ninh, Quảng Ninh, Hải Dương,
Hải Phòng, Hưng Yên, Thái Bình, Hà Nam, Nam Định, Ninh Bình, Hà Giang,
cao Bằng, Bắc Kạn, Tuyên Quang, Lào Cai, Yên Bái, Thái  Nguyên, Lạng Sơn,

📍 Cụm 1 (2 tỉnh/thành):
Hà Nội, TP. Hồ Chí Minh

📍 Cụm 2 (4 tỉnh/thành):
Thanh Hóa, Nghệ An, Đắc Lắk, Đồng Nai
```

Figure 17. Results of clustering provinces and cities based on their characteristics.

2.6. Student Population Forecasting

Forecasting for the year 2030 using linear regression: a linear regression model is applied to identify linear upward or downward trends based on data from 2002 to 2023, which is then used to project the student population for 2030. The province-level projections for 2030 are shown in Figure 18.

```
=== KẾT QUẢ DỰ BÁO NĂM 2030 ===
Tỉnh/Thành phố Dự báo năm 2030
Hà Nội                2,161,167
TP. Hồ Chí Minh       1,587,563
Đồng Nai                605,594
Nghệ An                 499,294
Thanh Hóa               493,827
Bình Dương              477,740
Hải Phòng               393,989
Gia Lai                 348,443
An Giang                340,104
Hải Dương               338,546
Bắc Giang               320,667
Đắc Lắk                 295,664
Sơn La                  292,361
Nam Định                289,334
Long An                 284,741
Thái Bình               277,245
Đồng Tháp               275,208
Quảng Ninh              267,756
Phú Thọ                 264,560
Kiên Giang              262,737
Lâm Đồng                257,501
Tiền Giang              254,575
Bắc Ninh                251,660
Bà Rịa - Vũng Tàu       237,619
Thái  Nguyên            234,140
Đà Nẵng                 227,335
Quảng Nam               226,428
Hưng Yên                226,119
Vĩnh Phúc               219,273
Bình Định               217,914
Bình Phước              216,192
Bình Thuận              208,505
Khánh Hòa               205,402
Hà Giang                198,159
Sóc Trăng               194,117
Cà Mau                  191,204
Tây Ninh                191,087
Hà Tĩnh                 177,303
Vĩnh Long               169,570
Lào Cai                 166,835
Thừa Thiên Huế          166,207
Quảng Ngãi              164,495
Bến Tre                 161,413
Trà Vinh                159,059
Yên Bái                 157,979
Ninh Bình               154,724
Hòa Bình                153,265
Cần Thơ                 151,593
Phú Yên                 146,193
Quảng Bình              144,341
Tuyên Quang             141,910
Hà Nam                  139,123
Bạc Liêu                137,131
Kon Tum                 131,802
Đắk Nông                118,647
Lạng Sơn                117,313
Điện Biên               116,092
Hậu Giang               108,752
Lai Châu                108,175
Ninh Thuận              107,992
Quảng Trị               107,858
cao Bằng                 75,597
Bắc Kạn                  46,322
Hà Tây                   10,202

Tổng số tỉnh/thành phố: 64
```

Figure 18. Forecast of the number of students by province/city for the year 2030.

To evaluate the accuracy of the Linear Regression model, one can use the following four metrics:

i) MAE: Mean Absolute Error;

ii) MSE: Mean Squared Error;

iii) RMSE: Root Mean Squared Error;

iv) R^2: Coefficient of Determination, which reflects the goodness of fit of the model.

Since actual data for 2030 is not yet available, the model will be evaluated by using data from the 2002–2023 period, split into a training set and a test set. The linear regression evaluation is presented in Figure 19. Prophet projections and their graphical comparison are shown in Figures 20 and 21, respectively.

```
=== ĐÁNH GIÁ MÔ HÌNH LINEAR REGRESSION ===
   Tỉnh/Thành phố        MAE           MSE        RMSE        R2
          Hà Nội  19992.039 4.590951e+08  21426.504     0.930
          Hà Tây 196499.140 4.018904e+10 200472.039 -3364.993
       Vĩnh Phúc  87138.828 7.862896e+09  88672.973   -62.263
        Bắc Ninh  60845.102 3.936883e+09  62744.582   -18.015
      Quảng Ninh  55273.311 3.156878e+09  56186.101   -32.971
       Hải Dương  96687.685 9.889190e+09  99444.407   -27.412
       Hải Phòng 100787.555 1.047888e+10 102366.384   -39.121
        Hưng Yên  71836.206 5.444826e+09  73789.065   -30.938
       Thái Bình  55491.196 3.230111e+09  56834.064   -74.920
          Hà Nam  40416.200 1.731624e+09  41612.787   -38.652
        Nam Định  79601.945 6.806119e+09  82499.206   -43.685
       Ninh Bình  63160.290 4.272213e+09  65362.170   -31.575
        Hà Giang  32041.156 1.133564e+09  33668.448    -7.836
       cao Bằng  29816.498 9.598122e+08  30980.837   -46.883
         Bắc Kạn  19510.561 4.014099e+08  20035.215   -70.168
     Tuyên Quang  51689.877 2.792724e+09  52846.226   -60.323
         Lào Cai  32297.139 1.096090e+09  33107.246   -21.466
         Yên Bái  38085.157 1.517901e+09  38960.245   -41.740
    Thái  Nguyên  60002.621 3.763140e+09  61344.435   -31.329
        Lạng Sơn  45593.739 2.215770e+09  47071.971   -56.862
       Bắc Giang 109863.011 1.292752e+10 113699.273   -25.488
         Phú Thọ  85353.404 7.613530e+09  87255.544   -41.773
       Điện Biên  55320.070 3.194679e+09  56521.495   -77.313
        Lai Châu  26592.826 7.152566e+08  26744.282  -146.668
          Sơn La  41295.290 1.719307e+09  41464.530   -68.979
        Hòa Bình  44681.696 2.057058e+09  45354.801   -82.405
      Thanh Hóa 262747.513 7.417302e+10 272347.247   -44.808
        Nghệ An 247504.086 6.660276e+10 258075.102   -30.168
         Hà Tĩnh  92220.636 9.323325e+09  96557.367   -41.909
      Quảng Bình  53140.883 3.051202e+09  55237.684   -41.496
       Quảng Trị  29406.164 9.700592e+08  31145.773   -25.645
Thừa Thiên Huế  41679.056 1.954949e+09  44214.806   -37.291
       Đà Nẵng  37706.844 1.494776e+09  38662.332   -12.884
     Quảng Nam  64291.923 4.583304e+09  67700.104   -33.680
     Quảng Ngãi  53159.378 3.050968e+09  55235.568  -108.586
      Bình Định  39813.475 1.722110e+09  41498.313  -250.723
        Phú Yên  15489.800 2.549695e+08  15967.765  -333.642
      Khánh Hòa  25194.584 6.720239e+08  25923.424   -67.917
     Ninh Thuận  16601.370 3.053574e+08  17474.479   -23.207
     Bình Thuận  40338.241 1.736335e+09  41669.352   -71.661
       Kon Tum  10787.955 1.245013e+08  11158.013    -6.701
       Gia Lai  13967.855 2.299161e+08  15162.985    -1.674
       Đắc Lắk  94365.975 9.652345e+09  98246.347   -73.762
       Đắk Nông  49182.947 2.553134e+09  50528.546   -53.317
       Lâm Đồng  35060.113 1.341909e+09  36632.075   -20.120
     Bình Phước  16818.725 2.946323e+08  17164.856   -12.816
       Tây Ninh  20699.259 4.318758e+08  20781.622  -177.428
     Bình Dương  70335.737 4.988273e+09  70627.709   -10.513
       Đồng Nai  98178.205 9.923447e+09  99616.500   -26.008
Bà Rịa - Vũng Tàu  29660.658 9.243703e+08  30403.458   -15.944
  TP. Hồ Chí Minh  46225.432 2.568315e+09  50678.544    -6.467
        Long An  20669.344 4.291568e+08  20716.099  -127.286
     Tiền Giang  36315.339 1.327267e+09  36431.680  -184.542
        Bến Tre  36143.910 1.332989e+09  36510.123  -583.832
       Trà Vinh  32864.816 1.107011e+09  33271.772  -151.111
       Vĩnh Long   8202.153 6.770234e+07   8228.143  -138.641
       Đồng Tháp   1732.783 3.879758e+06   1969.710     0.292
       An Giang  34958.784 1.236411e+09  35162.630   -25.508
      Kiên Giang  59095.967 3.525848e+09  59378.851  -619.122
        Cần Tho  72308.461 5.287932e+09  72718.171 -3104.520
      Hậu Giang  15127.623 2.326094e+08  15251.538   -39.590
      Sóc Trăng  39604.540 1.612336e+09  40153.901  -269.231
       Bạc Liêu   9724.928 9.646371e+07   9821.594   -48.703
        Cà Mau  20179.850 4.084188e+08  20209.375   -33.319
```

Figure 19. Evaluating the accuracy of a linear regression model.

```
=== KẾT QUẢ DỰ BÁO PROPHET NĂM 2030 =
   Tỉnh/Thành phố Dự báo năm 2030
           Hà Nội       2,161,656
    TP. Hồ Chí Minh       1,562,921
        Đồng Nai         734,759
       Bình Dương         549,202
         Nghệ An         499,514
       Thanh Hóa         494,156
        Nam Định         420,120
         Đắc Lắk         410,911
       Hải Phòng         394,161
         Gia Lai         377,934
       Thái Bình         358,029
       Quảng Nam         341,737
          Sơn La         341,469
         An Giang         340,204
       Hải Dương         338,834
        Lâm Đồng         321,530
       Bắc Giang         320,695
         Long An         302,046
       Bình Định         287,910
Bà Rịa - Vũng Tàu         282,844
        Đà Nẵng         282,522
      Bình Thuận         275,707
      Đồng Tháp         275,176
      Quảng Ninh         268,097
        Phú Thọ         264,562
      Kiên Giang         262,741
       Hà Giang         256,402
     Tiền Giang         254,552
       Bắc Ninh         251,943
Thừa Thiên Huế         251,583
      Quảng Ngãi         251,394
      Khánh Hòa         243,700
     Bình Phước         237,643
  Thái  Nguyên         234,215
       Hưng Yên         226,210
      Vĩnh Phúc         219,387
        Lào Cai         208,556
         Cà Mau         207,911
       Tây Ninh         206,333
        Hà Nam         200,530
       Trà Vinh         196,761
      Sóc Trăng         194,104
        Bến Tre         185,481
        Hà Tĩnh         177,415
       Vĩnh Long         169,583
        Phú Yên         168,251
        Yên Bái         157,964
      Ninh Bình         154,778
      Hòa Bình         153,290
        Cần Tho         151,634
        Kon Tum         148,099
      Quảng Bình         144,550
     Tuyên Quang         141,921
     Ninh Thuận         141,697
        Bạc Liêu         137,384
       Quảng Trị         135,839
       Đắk Nông         118,826
        Lạng Sơn         117,345
       Điện Biên         117,034
       Hậu Giang         109,695
        Lai Châu         108,180
       cao Bằng          75,776
         Bắc Kạn          46,330
         Hà Tây          10,454

Tổng số tỉnh/thành phố: 64
```

Figure 20. Forecasting results using the Prophet regression model.

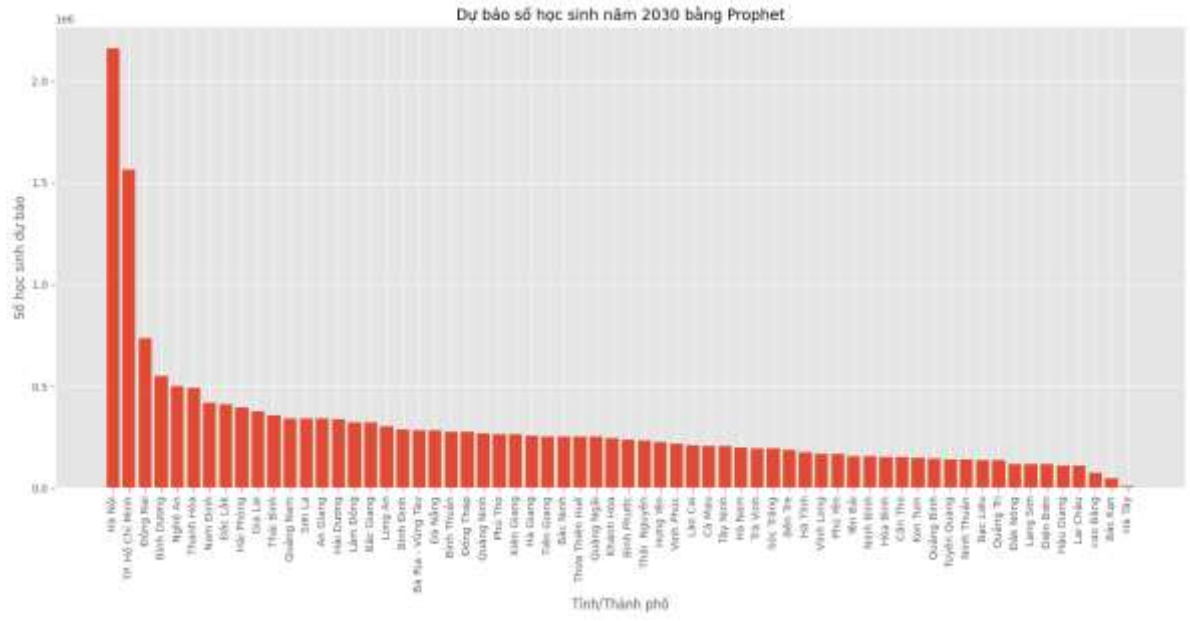

Figure 21. Forecast chart, corresponding to the results mentioned above.

Between these two methods, the Prophet model yields more accurate forecasts. This comparative assessment is based on the performance of both models. Consequently, this paper proposes using the Prophet model to forecast student enrollment across different educational levels in Vietnam. Charts generated by the program indicate that the Prophet model provides more plausible forecasts for the year 2030 when compared against actual data from 2002 to 2023 for specific localities. The evaluation metrics for the Prophet model are reported in Figure 22.

3. Conclusion

Based on a dataset concerning education in Vietnam, this study has carried out the following: i) Conducted exploratory data analysis (EDA), presenting the findings through statistical tables, descriptive statistics, and corresponding charts; and ii) Employed a linear regression model and the Prophet regression model to forecast student enrollment by educational level for the year 2030. The statistical analysis process outlined in this paper holds significant value for educational management: by leveraging the obtained results - particularly through data visualizations - administrators and policymakers can make quantitative, data-driven decisions. This research contributes to the data analysis workflow. The application of data science within the education sector is an urgent necessity. These analytical results empower educational administrators to make quantitative decisions, thereby supporting the digital transformation of management practices.

Three points for further discussion are proposed:

i) The role of various data science phases in processing educational datasets. The data science process serves as a systematic backbone, transforming raw, unwieldy educational datasets into actionable insights:

- Data collection and cleaning: This involves standardizing data structures from various sources (such as the General Statistics Office) and rigorously addressing missing or anomalous data, thereby ensuring the reliability of subsequent in-depth analysis.
- Exploratory Data Analysis (EDA): Through descriptive statistical measures,

analysis of right-skewed distributions, and correlation charts, EDA provides a comprehensive overview of student population scales, highlights significant disparities between localities, and identifies specific fluctuations (such as changes in administrative boundaries).

- Modeling and visualization: The application of analytical models translates numerical data into intuitive, accessible visualizations, establishing a robust, evidence-based foundation for strategic planning within the education sector.

Tỉnh/Thành phố	MAE	MSE	RMSE	R2
Hà Nội	20035.793	4.617929e+08	21489.367	0.930
Hà Tây	196043.996	4.000620e+10	200015.495	-3349.679
Vĩnh Phúc	87115.970	7.858797e+09	88649.857	-62.230
Bắc Ninh	13405.044	2.458000e+08	15678.010	-0.187
Quảng Ninh	55099.785	3.136903e+09	56008.059	-32.757
Hải Dương	96561.405	9.863520e+09	99315.254	-27.339
Hải Phòng	100405.572	1.040007e+10	101980.714	-38.820
Hưng Yên	13075.609	1.833927e+08	13542.257	-0.076
Thái Bình	8739.101	8.069110e+07	8982.823	-0.897
Hà Nam	8383.852	7.672981e+07	8759.556	-0.757
Nam Định	22400.707	5.706366e+08	23888.002	-2.746
Ninh Bình	63061.374	4.258978e+09	65260.844	-31.474
Hà Giang	10783.770	1.485038e+08	12186.213	-0.158
cao Bằng	6207.180	4.799595e+07	6927.911	-1.394
Bắc Kạn	19471.197	3.997918e+08	19994.795	-69.881
Tuyên Quang	51599.598	2.783184e+09	52755.893	-60.114
Lào Cai	7163.817	5.550859e+07	7450.409	-0.138
Yên Bái	37988.217	1.510254e+09	38861.986	-41.525
Thái Nguyên	59999.166	3.762384e+09	61338.279	-31.322
Lạng Sơn	45502.377	2.206977e+09	46978.472	-56.632
Bắc Giang	109598.471	1.286618e+10	113429.186	-25.362
Phú Thọ	85220.531	7.590088e+09	87121.114	-41.642
Điện Biên	55245.513	3.186253e+09	56446.908	-77.106
Lai Châu	26340.754	7.017926e+08	26491.368	-143.888
Sơn La	5242.125	3.954320e+07	6288.339	-0.609
Hòa Bình	44603.122	2.049911e+09	45275.942	-82.115
Thanh Hóa	263089.794	7.435566e+10	272682.340	-44.921
Nghệ An	246824.044	6.624688e+10	257384.684	-30.001
Hà Tĩnh	25086.169	7.864491e+08	28043.701	-2.620
Quảng Bình	20913.120	4.966356e+08	22285.322	-5.917
Quảng Trị	28103.036	8.895632e+08	29825.545	-23.434
Thừa Thiên Huế	24157.882	6.893823e+08	26256.091	-12.503
Đà Nẵng	8408.990	7.466987e+07	8641.173	0.306
Quảng Nam	34028.172	1.361658e+09	36900.653	-9.303
Quảng Ngãi	9163.532	1.042479e+08	10210.184	-2.744
Bình Định	11956.513	1.638048e+08	12798.624	-22.944
Phú Yên	12159.449	1.581852e+08	12577.169	-206.615
Khánh Hòa	8168.664	7.292928e+07	8539.864	-6.479
Ninh Thuận	11635.063	1.546148e+08	12434.418	-11.257
Bình Thuận	21351.957	4.928604e+08	22200.460	-19.625
Kon Tum	3947.859	1.759640e+07	4194.806	-0.088
Gia Lai	11354.339	1.578993e+08	12565.798	-0.836
Đắc Lắk	59833.270	3.985052e+09	63127.268	-29.866
Đắk Nông	49131.751	2.547939e+09	50477.115	-53.206
Lâm Đồng	14767.576	2.500481e+08	15812.908	-2.935
Bình Phước	4228.425	1.874942e+07	4330.060	0.121
Tây Ninh	4596.306	4.218702e+07	6495.153	-16.429
Bình Dương	70121.605	4.957833e+09	70411.882	-10.443
Đồng Nai	15993.532	2.729727e+08	16521.885	0.257
Bà Rịa - Vũng Tàu	10557.135	1.173290e+08	10831.853	-1.151
TP. Hồ Chí Minh	37666.608	2.595559e+09	50946.629	-6.546
Long An	12417.791	1.982906e+08	14081.570	-58.274
Tiền Giang	36305.903	1.326577e+09	36422.205	-184.445
Bến Tre	27001.594	7.401909e+08	27206.450	-323.749
Trà Vinh	32697.619	1.095755e+09	33102.193	-149.564
Vĩnh Long	8589.789	7.420555e+07	8614.264	-152.054
Đồng Tháp	1839.532	4.175105e+06	2043.307	0.238
An Giang	34945.181	1.235483e+09	35149.442	-25.488
Kiên Giang	59097.530	3.526022e+09	59380.313	-619.153
Cần Thơ	72180.828	5.269207e+09	72589.304	-3093.523
Hậu Giang	15077.573	2.310738e+08	15201.113	-39.322
Sóc Trăng	39435.411	1.598714e+09	39983.919	-266.948
Bạc Liêu	9741.075	9.678371e+07	9837.871	-48.868
Cà Mau	19915.730	3.978361e+08	19945.828	-32.430

Figure 22. Evaluation of the Prophet model's forecasting results.

ii) Significance for education administrators: The application of data analysis and machine learning offers profound strategic value for administrators at all levels:

- Evidence-based decision-making: Moving beyond intuition, administrators can gain an accurate understanding of student population size, growth rates, and demographic composition in each locality, enabling timely adjustments to macro-level policies.

- Optimized resource allocation: Identifying provinces and cities with exceptionally large student populations- such as Hanoi, which has seen a sharp increase to 1.7 million students - allows the education sector to proactively plan school networks, allocate budgets for infrastructure, and distribute teaching staff in a fair and rational manner.

- Long-term planning: Provides a reliable forecasting foundation for developing educational strategies that anticipate future demographic shifts.

iii) The significance of machine learning tools for data forecasting: Machine learning tools - such as linear regression models and, notably, Meta's Prophet model - play a pivotal role in educational data forecasting:

- Flexible time-series processing: Machine learning models are capable of capturing complex trends, seasonality, and sudden change points within long-term historical data (spanning from 2002 to 2023).

- High accuracy and strong adaptability: Through the comparison of evaluation metrics (such as MAE, MSE, RMSE, and R^2), this research demonstrates that the Prophet model yields more reasonable and realistic forecasts compared to traditional methods.

- Enhanced future forecasting capabilities: Machine learning tools enable accurate projections of student enrollment by province up to 2030, serving as a powerful instrument to help the education sector proactively address various development scenarios in the digital era.

References

[1] J. W. Creswell, J. D. Creswell, *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research* (6th ed.). Pearson, 2018.

[2] E. Kalita, S. S. Oyelere, S. Gaftandzhieva, K. N. V. P. S. Rajesh, S. Kumar Jagatheesaperumal, A. Mohamed, Y. M. Elbarawy, A. S. Desuky, S. Hussain, M. A. Cifci, P. Theodorou, S. Hilčenko, J. Hazarika, and T. Ali, "Educational Data Mining: A 10-year Review", *Discover Computing*, Vol. 28, No. 1, Art. No. 81, 2025, https://doi.org/10.1007/s10791-025-09589-z.

[3] E. L. Meneses, P. C. M. Moreno, C. G. Herrerías, N. Pelícano-Piris, "Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics", *Computers*, Vol. 14, No. 2, Art. No. 68, 2025, https://doi.org/10.3390/computers14020068.

[4] Chuah, T. C. Hewitt, S. A. Ali, M. May, T. Xu, D. Christiadi, P. Y. I. Choi, E. E. Gardiner, T. D. Andrews, "EDAmame: Interactive Exploratory Data Analyses with Explainable Models", *Bioinformatics*, Vol. 41, No. 6, Art. No. btaf340, Jun, 2025, https://doi.org/10.1093/bioinformatics/btaf340.

[5] S. E. Kamand, J. M. W. Quinn, M. J. Cowley, "ggEDA: Visualisations for Exploratory Data Analysis using Tiled One-dimensional Graphics and Parallel Coordinate Plots", *F1000Research*, Vol. 14, 2025, pp. 1248, https://doi.org/10.12688/f1000research.168305.1.

[6] https://www.nso.gov.vn/Statistics Office – Ministry of Finance, Dataset on Students in Vietnam's Education System, 2026, https://www.nso.gov.vn/2026 (accessed on: September 25th, 2026) (in Vietnamese).

[7] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*, Springer, 2023.

[8] D. C. Montgomery, E. A. Peck, G. G. Vining, *Introduction to Linear Regression Analysis* (6th ed.), Wiley, 2021.

[9] D. Paper, *TensorFlow 2.x in the Colaboratory Cloud: An Introduction to Deep Learning on Google's Cloud Service*, Apress, 2021.

[10] S. J. Taylor, B. Letham, Forecasting at Scale, *The American Statistician*, Vol. 72, No. 1, 2018, pp. 37-45, https://doi.org/10.1080/00031305.2017.1380080.

[11] T. T. T. Ngo, "Forecasting the Number of High School Graduates for the Period 2020–2030", Journal of Education, Vol. 509, No. 1, pp. 9–13, Sep, 2021 (in Vietnamese).

[12] T. An, "Digital Data in Education Management – An Inevitable Trend for a Breakthrough in Educational Development", Journal of Education, Oct. 8, 2025, Forecasting at Scale, *The American Statistician*, Vol. 72, No. 1, 2018, pp. 37-45 (in Vietnamese).