www.ijarp.com ISSN 2456-9992
Page 1 to 28
International Journal Advanced Research Publication

Research-Article
Volume: 2
Issue: 10

EDUCATIONAL AND SOCIOECONOMIC FACTORS ASSOCIATED WITH STUDENTS' ACADEMIC PERFORMANCE: A STATISTICAL ANALYSIS OF THE SPD24 DATASET

Sabera Islam, Mst Khadizatul Kobra*

Department of Business Studies, Human resource Management North Bengal International University, Rajshahi, Dhaka, Bangladesh.

Article Received	Article Revised	Published
11-09-2026	01-10-2026	21-10-2026

***Corresponding Author:** Mst Khadizatul Kobra
Department of Business Studies, Human resource Management North Bengal International University, Rajshahi, Dhaka, Bangladesh.
DOI: https://doi-doi.org/101555/ijarp.5423

ABSTRACT

The study investigated the association between educational factors, socioeconomic characteristics, and students' academic performance, based on the SPD24 Student Performance Dataset. A quantitative, observational design was utilized to analyze 98,000 student records across 41 variables . The analysis comprised descriptive statistics, Spearman rank correlations, chi-square tests of independence, Cramér's V effect-size estimation, ordinal logistic regression, and further comparisons of multinomial logistic regression and XGBoost. The dataset did not contain any missing values, duplicate records, or duplicate student IDs. Academic performance was classified into three categories; Low (24.68%), Medium (37.66%) and High (37.66%). Spearman correlation ($\rho = 0.9371$, $p < 0.001$) identified a very strong association between prior academic performance and the outcome, so this variable was left out of the primary regression model to avoid potential target leakage or outcome-construction overlap. Categorical association tests showed a statistically significant association between parental education level and academic performance, but the effect size was negligible (Cramér's V = 0.0100). Socioeconomic status was the only predictor that was statistically significant at the 0.05 level in the ordinal logistic regression model ($\beta = -0.0290$, $p = 0.0446$). However, the additional classification results showed limited predictive discrimination, with accuracy of 38.08%, balanced accuracy of 33.70%, macro-F1 score of 0.290, and multiclass ROC-AUC of 0.501. In general, the findings indicate that the associations found were mostly weak, and the observational design does not permit causal conclusions.

KEYWORDS: Academic performance; Educational factors; Socioeconomic status; SPD24 dataset; Ordinal logistic regression; Machine learning.

1. INTRODUCTION

Academic performance is a key measure of student progress in education. It is commonly used to evaluate learning outcomes, spot students who might require extra help, and guide educational planning. But academic achievement is determined by many interacting dimensions including students' learning behaviors, prior achievement, family background, socio-economic circumstances and access to educational resources. Therefore, it is important to understand the relationship between these factors and academic performance to be able to develop evidence-informed educational policies and support strategies. Simultaneously, establishing statistically significant relationships is not a guaranty that a specific factor has a meaningful practical effect or can be used to reliably predict an individual student's outcome.

Educational data mining (EDM) and learning analytics have improved the chances of analyzing student data through statistical and computational approaches. Earlier reviews [1]–[5] have documented the use of classification, regression, and other machine-learning techniques to study academic outcomes and potentially informative student characteristics. These methods can help in the analysis of large educational datasets, but the results of these analyzes are sensitive to the quality of the data, the definition of the outcome, the choice of predictors, and the evaluation procedures. Research has also emphasized the need for validation and careful interpretation when predictive models are designed for educational contexts [6], [7]. Hence, statistical association analysis and predictive modeling should be viewed as related but different analytical tasks.

Students come from different families with different resources, parents with different levels of education, different opportunities to learn, and different access to technology, so educational and socioeconomic conditions matter a lot for academic achievement.
The OECD's analysis of PISA 2022 reported large disparities in educational outcomes associated with socioeconomic background across participating education systems [8]. However, a systematic review of the relationship between socioeconomic status and academic performance in higher education found that this relationship can be relatively weak and that previous academic achievement and other educational experiences may have stronger associations [9]. The results indicate that socioeconomic indicators need not be assumed to have large and uniform effects across datasets and educational environments. Rather, their linkages should be empirically tested, considering both statistical significance and effect size.

Parental education, parental involvement, and socio-economic status of the household are the aspects frequently studied as predictors of the educational and family environment of the students. However, their associations with academic achievement might vary depending on the population studied, the definition of the outcome, measurement method, and other factors that are incorporated into the study. Academic achievement studies have thus turned to incorporating several dimensions of student and family traits rather than a single predictor [8], [9]. On a parallel note, the increased availability of educational information systems has made it possible to analyze demographic, behavioral, and institutional factors simultaneously. However, a high number of observations does not mean that there will be strong relationships or that the model will yield any predictions.

Learning behaviors and educational involvement are also common topics of investigation in academic performance studies. Attendance, studying hours, homework compliance, class participation, and availability of learning materials could be reasonable measures of students' involvement into the educational process. In review papers on predictive studies on student performance a large variety of behavioral, demographic and academic characteristics, along with differences between predictors and modeling approaches employed by various authors, has been considered [1], [2], [4], [5]. Learning analytics studies have likewise focused on the importance of taking into account the specifics of student data collection and interpretation of analytical results within an educational context [6], [7]. At the same time, the theoretical significance of a characteristic does not guarantee its empirical correlation with academic performance in all datasets. Further empirical examination is necessary for finding out whether such measures differentiate students between performance groups in a specific sample.

Machine learning algorithms have been widely employed in educational studies due to their ability to express complex patterns and classify students into outcome groups. The literature review has included many different supervised learning methods and their application to predicting academic performance [1]–[5]. For instance, XGBoost offers a scalable solution for gradient boosting in structured data [10]. Nonetheless, an algorithm cannot be appraised solely according to its accuracy as long as the performance is measured on different outcome groups. Class-specific recall, balanced accuracy, macro-averaged F1-score, and multiclass ROC-AUC serve as complementary measures for assessing classification abilities of a model. An algorithm may yield decent performance overall but perform poorly on a certain class. This issue is crucial when the educational application aims to identify students who are at risk of poor performance.

However, despite the progress made in educational data mining, a number of methodological concerns remain relevant. First, works focused on prediction do not necessarily distinguish between association detection and usefulness of prediction. Second, the effect size of statistically significant results can be negligibly small, especially in case of large datasets, and both statistical significance and practical importance should be taken into account. Third, features which are associated with the target variable are likely to be correlated with the process of target definition, which is known as target leakage or outcome construction overlap. The presence of such features in the dataset makes results of the modeling look more valuable than they would in practice. Finally, classification metrics as a whole may mask poor identification of a minority or hard-to-classify class. The recent review of educational data mining and learning analytics methods mentioned methodological diversity, validation problems, and the need to interpret predictive results more carefully [1], [3]–[7]. These considerations justify an approach based on analyzing associations first and assessing prediction secondarily.

To solve these problems, the current study focuses on the SPD24 Student Performance Dataset that comprises 98,000 rows and 41 columns. Outcome is represented by the variable Performance Score with three classes: Low, Medium, and High. The database encompasses educational variables like attendance percentage, studying hours, homework, participation in class, and availability of learning materials, along with socioeconomic and family-related features like parent's education level, socioeconomic status, parent’s job, and family size. Descriptive statistics, Spearman rank correlation, chi-square test of independence, Cramér’s V estimation of effect size, and ordinal logistic regression are used in the analysis. Multinomial logistic regression and XGBoost serve as additional classifiers rather than the main basis of the research.

One of the key aspects of the analysis is the diagnostics of the previous academic performance of students. This variable has an extremely high correlation with the dependent variable ($\rho = 0.9371$,

$p < 0.001$). Since this trend might be an effect of target leakage or some overlap between this variable and the process of outcome construction, previous academic performance will not be included in the main regression model and only considered during diagnostics and descriptive statistics. Such an approach is aimed at decreasing the chances of overestimating the contribution of other predictors into the result. The class-wise classification performance will also be taken into account to avoid masking the deficiencies.

The study addresses three research questions:

1. **RQ1:** What associations exist between selected educational and socioeconomic characteristics and students' academic performance in the SPD24 dataset?
2. **RQ2:** Which of the examined educational and socioeconomic variables show statistically significant associations with academic performance, and how large are the observed effects?
3. **RQ3:** How well do supplementary multinomial logistic regression and XGBoost models distinguish the three academic-performance categories when previous academic performance is excluded?

Three main contributions are made in the study. First, a statistical analysis is conducted on the basis of a large sample size on the influence of the educational and socioeconomic characteristics on a three-category variable representing academic performance. Second, effect sizes are considered together with statistical significance of the results, which allows identifying significant associations and distinguishing them from relationships that can be detected by statistical methods but do not have practical value. Third, additional classification methods are considered using several performance measures and analyzed for specific class-related limitations. According to the results, the observed associations are relatively weak: while the level of parental education statistically influences the performance in the categorical analysis, its effect size is zero, and the only statistically significant predictor of academic performance at 0.05 level in the ordinal logistic regression is low socioeconomic status. Moreover, the additional classification analysis reveals the weakness of discrimination: it does not allow detecting Low-performance students in the multinomial logistic regression results. That is why no assumption is made about the strong association of the selected variables with academic performance.

2. Literature Review

2.1. Educational Data Mining and Student Academic Performance

Educational data mining (EDM) and learning analytics have been gaining significance in analyzing the achievement of students, detecting trends within educational data, and informing academically supported strategies. In recent systematic reviews, it was established that algorithms such as classification, regression, decision trees, ensemble methods, and neural networks have commonly been utilized in investigating academic achievement and at-risk students [21], [22]. However, the high frequency of usage of an algorithm cannot determine its successful work in every educational context. The algorithm effectiveness depends on several factors such as the data set, the definition of an outcome variable, the predictors' relevancy and quality, and the validation process.

One of the major limitations of the existing literature on the topic is the inconsistency in terms of methodology and performance criteria. Researchers can predict test scores, class marks, academic advancement, or dropouts; however, these criteria measure different aspects of

educational achievement. A systematic review of the machine learning literature revealed significant variation in populations, predictor variables, and methods employed in predicting student outcome [21]. Likewise, a review of 162 articles by Alalawi et al. showed greater frequency of use of classification versus regression methods and a lack of evaluation of intervention efforts after identifying at-risk students [22]. This means that rapid progress has been made in predicting models, yet converting prediction results into valuable educational evidence poses a challenge.

Another issue that arises is the focus on accuracy of classification, without due attention being paid to performance in each class. In case the output variable comprises several ordered classes, one can obtain seemingly decent accuracy by simply predicting the most easily predictable class, whereas the other students would not get classified at all. Thus, besides traditional accuracy, one needs to take into consideration balanced accuracy, F1 score, class recall, as well as multiclass discrimination. This is especially true for dividing academic performance into Low, Medium, and High classes.

2.2. Socioeconomic Status and Parental Education

SES is one of the essential context dimensions when studying educational inequities. The dissimilarities in family resource levels, parental education, educational expectations, and learning opportunities might have an impact on the educational experience of students. Recently, Rakesh et al. conducted a systematic literature review on the connections between socioeconomic status and cognitive, language, and academic skills of youth [23]. The authors emphasize the necessity to consider the mechanisms and protective factors mediating the social situation influence on educational development. However, SES cannot be viewed as a deterministic factor that influences the student's performance.

Moreover, studies examining the relationship between family socioeconomic situation and achievement suggest that the SES-performance connection can take place through several pathways, instead of just being one direct link. For instance, Zhang et al. explored the connection between SES, parental academic participation, and children's academic performance in China [24]. The study shows that it is crucial to make a distinction between the family background and education that could take place within it. The connection between SES and performance depends on the context of education, measurement of SES, and family/school characteristics.

The level of parental education is one more socioeconomic indicator widely studied by researchers. Parental education could mean that parents are familiar with educational process, its demands or their ability to help with learning. Nevertheless, parental education does not mean household income, availability of education or parental involvement. These factors could coincide in some way, however, they have different aspects of the family context and must be considered as independent variables.

One more caution that researchers could get from recently conducted big studies is the idea that it cannot be taken for granted that family related variables have big impact on achievement. One 2025 meta-analysis of the influence of parental involvement of Chinese students based on 363 regional samples and 399,375 students found positive correlation but relatively small correlation of parental involvement with achievement depending on types of involvement and students' characteristics [25]. In addition, meta-analysis of parental involvement and math achievement showed positive correlation but variability depending on moderators [26]. It means that statistically significant connection between family context and achievement could be small.

In the context of this research, this body of literature informs that parental education, socio-economic status, parental occupation, and parental involvement should be studied separately. At the same time, this literature calls for effect sizes to be measured along with statistical significance. The results obtained by means of SPD24 show that there was statistical significance between parental education level and performance in the chi-square test; however, the value of Cramér's coefficient (V=0.0100) showed negligible significance. Hence, the result cannot be seen as an indication that parental education is a significant factor in determining academic performance.

2.3. Educational Engagement, Learning Behaviours, and Academic Outcomes

The connection between students' academic success and learning behaviors and education engagement is another popular research topic. Student attendance, studying time, completing homework, participating in classes, tutoring, and accessing learning resources are common elements of the analysis of performance due to their possible relevance for students' learning environment. Nevertheless, the fact that some variable is included in the model of prediction of the outcome does not mean that there is a significant or an independent relationship between this variable and the outcome.

A good example of this problem can be demonstrated by the literature on parental involvement. According to Kim's second-order meta-analysis based on 23 meta-analyses and 1,177 primary studies, positive relationships were found between parental involvement and student achievement, where parental expectations and aspirations are more effective than homework help [27]. This result shows that general indicators of parental involvement hide certain nuances that make it necessary to use more complex methods of evaluating this type of interaction between parents and students.

Further evidence from different educational contexts points to the interpretation of associations in the light of measurement and contextual limitations. Otani studied the impact of parental involvement on academic achievement in elementary and middle schools [28]; other researchers considered the influence of parental expectations, support, and educational engagement on performance. Thus, the necessity of distinction between the relevance of the variable and significance of its association in a particular dataset is evident.

In the current research, the observed associations with attendance rate, study time, homework, and some other behaviour characteristics were very small and statistically insignificant in the performed Spearman correlation tests. These findings contradict the expected finding that educational engagement measures always discriminate high-performing from low-performing students. Nevertheless, they do not prove the absence of associations of attendance or study time with achievement. Weak associations could be explained by the peculiarities of the operationalization of the variables, variability of the measures, differences in the sample, or the usage of the categorized measure of performance. The weak associations may also show the lack of informative content of the variables in terms of SPD24.

The differentiation is key to performing the critical review, as one must compare the new results with earlier studies without the implicit assumption that either the new results or those from earlier studies hold true across the board. This means that rather than accepting all theoretically relevant variables as predictors, one must verify the relationship based on the data.

2.4. Machine Learning, Model Evaluation, and Methodological Limitations

Machine learning models are widely used for the analysis of academic performance due to their capability to identify non-linear dependencies between variables. The literature on the subject proves the extensive use of supervised classification algorithms and features selection in educational prediction research [21], [22]. Predictive performance demonstrated by one study cannot be readily compared to the performance of another study due to differences in datasets, target definition, class distribution, and validation methodology.

The second methodological issue is related to the use of past academic performance as a predictor of current performance category. Past performance might be a good predictor of future performance and is widely used as a predictor variable in educational prediction research. Yet, the predictor and target variables should be carefully evaluated in the context of their independence as well as the feasibility of the predictor at the moment of prediction. When the target category is constructed using the same information as the predictor, the association represents the overlap of outcomes rather than prediction. This is also connected with the problem of data leakage leading to an overly optimistic predictive performance because of including information unavailable at the moment of prediction in model training [29].

The previous academic performance in the SPD24 analysis correlates with the outcome in an extremely high Spearman correlation ((ho=0.9371, p<0.001)). Such high correlation should be inspected because it is substantially higher than the correlations found for most other analyzed variables. The proper construction of the outcome should be confirmed before claiming any leakage. Therefore, the current study treats such a correlation as a target overlap factor and leaves the previous academic performance out of the main regression and machine learning analysis. In this way, the predictive task becomes more informative regarding the rest of the available characteristics.

It is important to take into account such aspects as class imbalance and the difference between classification performance and explanatory correlation when analyzing the model results. Having a large sample does not imply that there are sufficient predictors to classify the outcome classes. In particular, the overall accuracy measure may hide the fact that there is poor classification performance for a certain class. In educational data mining, fairness literature also highlights the problematic aspects of using demographic predictors that relate to the possible influence of the output of the model on different groups of students [30], [31].
In the current research, the researcher has compared unweighted and class weighted XGBoost classifiers and presented recall per class together with accuracy, balanced accuracy, macro-F1, and multiclass ROC-AUC scores. As can be seen from the results of testing the classifier on the test set, the unweighted classifier scored an accuracy of around 37.89% and a macro-F1 of about 0.289. Its recall on the Low category was around 0.0014. Class weighting has increased recall on the Low category to about 0.3271 and macro-F1 to 0.3296, at the expense of reducing accuracy to about 33.30%. The multinomial logistic regression classifier could not identify any examples of the Low category. It can be concluded that the problem is still quite difficult even without taking into account academic results.

2.5. Theoretical Framework

This study takes into consideration a contextual theory of educational achievement, mainly grounded in the theory of socioeconomic stratification, as well as the general theory of the connection between academic achievement and the characteristics of students, their family

background, and educational environment. According to this approach, such factors as socioeconomic status and parents' level of education are examples of contextual variables, while variables like attendance, study time, doing homework, parental involvement and availability of learning materials are educational/behavioural variables. While being analytically distinct, these categories of variables can be connected in reality.

The framework of this study can be aligned with the learning-analytics approach to educational research, which involves analysis of educational data to find out patterns potentially beneficial for understanding and decision-making processes [21], [22]. The idea is not to make assumptions about causal relationships between all of the selected variables and academic achievement but to check whether they are observable.

The theoretical framework underlies the basis of three dichotomies in analyses. The first is a separation between association and causation since the observational nature of the data set lacks evidence regarding temporal precedence as well as alternative explanations for the association observed. The second is a separation between statistical significance and substantive significance through effect size consideration. The last is a separation between explanatory analysis and prediction in which a variable can be significantly associated with the dependent variable but fail to improve the classification of the model.

This theoretical framework is relevant in the SPD24 research since it enables the study to assess education and socioeconomic status without assuming that the variables will make good predictors. Additionally, the framework makes a

conservative evaluation of the significance of low socioeconomic status

$$((\beta=-0.0290,\ p=0.0446)\backslash\backslash$$

$$((\beta=-0.0290,\ p=0.0446)\backslash\backslash$$

using the ordinal logistic regression approach. Although the variable was statistically significant at the 0.05 level in the fitted model, the coefficient was small while the proportional-odds assumption should be considered.

2.6. Research Gaps and Relevance to the Present Study

A number of gaps have been observed in the reviewed literature in relation to the current study. Firstly, there is diversity in outcomes, predictors, and evaluation measures being used in student performance prediction literature, thereby precluding any direct comparisons across the literature [21], [22]. Secondly, past academic performance has often been considered a predictor variable in the previous literature; however, there is need to take into account the connection between the variable and the definition of the target outcome in the case where the objective is to measure the impact of other educational and socioeconomic characteristics [29]. Thirdly, emphasis on classification accuracy understates the occurrence of poor performance within certain outcome categories.

This paper attempts to address these problems by conducting an analysis of 98,000 student records from SPD24 using descriptive, association-based, regression, and machine learning techniques. Three performance categories—Low, Medium, and High—are considered, while socio-

economic and educational factors are used in the analysis without prior academic performance. Effect sizes are calculated for categorical associations, while multiple performance measures are used to measure the classification quality.

In summary, the results can be considered as a partial contribution rather than an example of a very efficient predictive system. Parental education proved to be statistically associated with performance in the categorical analysis, but its effect size was very low. Only low SES was found to be statistically significant in the reported ordinal regression at the 0.05 level, while the coefficient estimate was very small. As for the additional classifiers, their discrimination power was low and they had difficulties with identifying the Low performance group. These results indicate that the variables and outcome variable used in SPD24 do not have enough separation power among performance categories if prior academic performance is disregarded.

This paper thus provides an analysis that clarifies both the statistical relationships as well as the predictive limits of a large dataset of students. The primary significance of this paper is methodological in nature; specifically, that it shows why researchers in education should separate the concepts of statistical significance and substantive significance, investigate target overlap, and analyze the predictive power for each class separately. Further studies can validate the process of generation of the outcome variable in the dataset, test the proportional-odds assumption, explore other definitions of the outcome variable, and investigate whether other contextual or longitudinal variables would improve the generalizability of the predictions made.

3. METHODOLOGY

3.1. Research Design

In this research, a quantitative and observational research approach was used to investigate the relationships between educational factors, socio-economic factors, and academic performance of the students. Descriptive statistics, non-parametric correlation analysis, Chi-Square test of independence, Cramér's V for effect size estimation, regression modeling, and comparisons with machine learning models were used for the analysis. The primary goal of the research was to establish the strength and the statistical proof of the relationship between selected educational and socio-economic factors on the one hand and academic performance on the other hand. The predictive ability of the performance was considered as an additional aspect of the research.

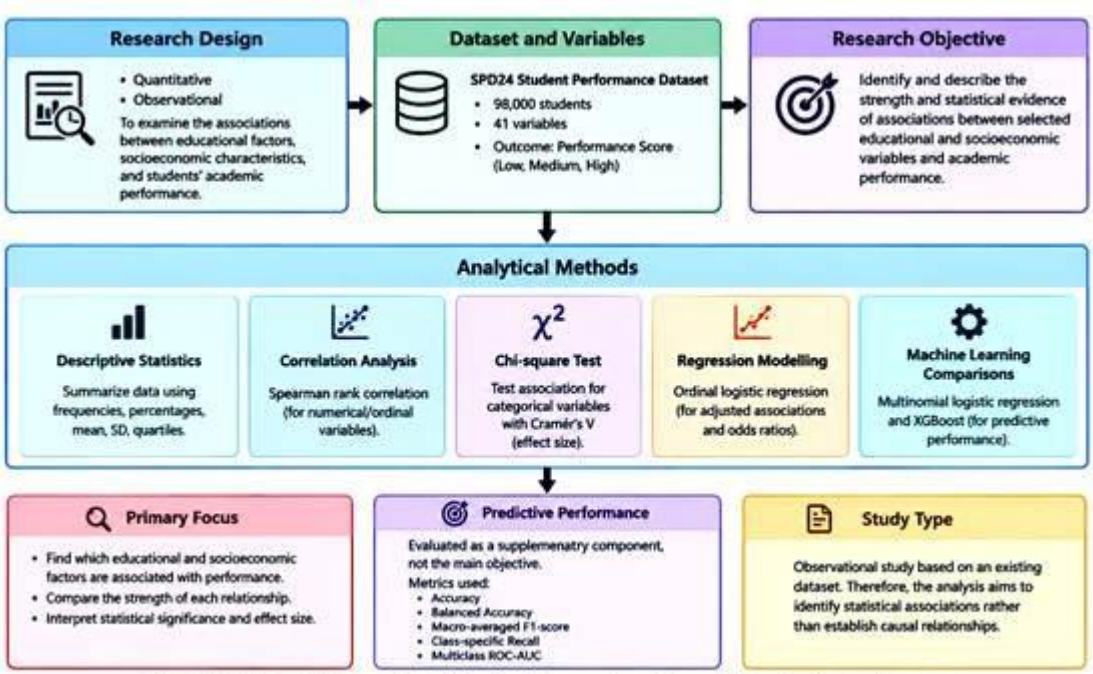


Figure 5. Analytical workflow for examining educational and socioeconomic associations with students' academic performance in the SPD24 dataset.

Figure 1. Analytical Workflow of the Study.

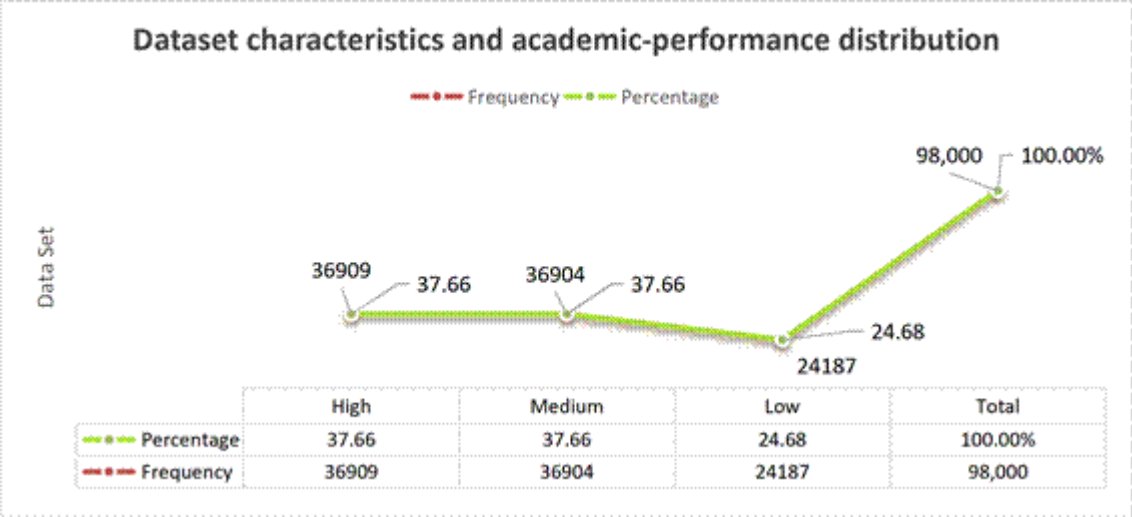
The following Figure 1 describes the quantitative, observational research approach employed in analyzing the connections among educational and socio-economic variables and students' academic achievement in the SPD24 dataset. The illustration encompasses the research approach, data and variables, descriptive statistics, Spearman correlation, chi-square test with Cramér's V measure, regression modeling, and additional machine learning approaches for comparison. Accuracy, balanced accuracy, macro-averaged F1-score, class-level recall, and multiclass ROC-AUC were used to evaluate predictive accuracy. The core purpose was the identification of statistical connections rather than causation.

The research study was observational in nature and employed an existing data set. Thus, the analysis was performed with the aim to establish statistical connections and not causation.

3.2. Dataset and Sample

The analysis involved the use of SPD24 Student Performance Dataset downloaded from Kaggle. The SPD24 Student Performance Dataset had 98,000 observations and 41 variables that included a student identifier, demographic variables, educational variables, socio-economic variables, proficiency levels and academic performance.

The response variable, Performance Score, was an ordinal categorical variable with three categories; Low, Medium and High. The original dataset was used for the analysis and there was no deletion of any observation due to missingness or duplication as there was no missing value, duplication and duplicate students in the original dataset.



	High	Medium	Low	Total
Percentage	37.66	37.66	24.68	100.00%
Frequency	36909	36904	24187	98,000

Figure 2. Dataset characteristics and academic-performance distribution.

The distribution of the response variable involved 24,187 students of low-performance category (24.68%), 36,904 students of medium-performance category (37.66%) and 36,909 students of high-performance category (37.66%).

3.3. Research Variables

Dependent Variable

Performance Score

Explanatory Variables

Educational Factors: Attendance Rate, Study Hours, Parental Involvement, Class Participation, Learning Resources Access, Internet Access, Homework Completion Rate, and Attendance of Tutoring Sessions

Socioeconomic Factors: Parental Education Level, Socioeconomic Status, Parental Employment Status, and Household Size

Control Variables: Age, Gender, Grade Level, School Type, and School Location

Previous Academic Performance was analyzed both in descriptive statistics and correlation analysis

Previous Academic Performance was left out of the primary regression analysis due to the extraordinarily strong relation with the dependent variable (Performance Score). It is possible that previous academic performance and the dependent variable are derived from the same source, or there is some target leakage, and hence, inclusion may give misleading results. This choice was made to report an accurate finding on independent explanatory variables rather than a misleadingly strong association. However, there is no evidence of how the dataset's target was generated.

Table 1. Main variables and their analytical roles

Variable group	Variables	Analysis
Outcome	Performance Score	Descriptive distribution, association tests, regression
Educational	Attendance, study hours, parental involvement, class participation, learning resources, internet access, homework completion, tutoring attendance	Spearman where numerical; chi-square where categorical; regression
Socioeconomic	Parental education, socioeconomic status, parental employment, household size	Chi-square/Cramér's V, regression
Controls	Age, gender, grade level, school type, school location	Multivariable regression
Diagnostic variable	Previous Academic Performance	Descriptive statistics, Spearman correlation, target-overlap diagnostic; excluded from primary regression and main XGBoost comparison

3.4. Data Quality Assessment and Preprocessing

The dataset was imported into Python using pandas from the Excel file. The first step involved checking the dimensions of the dataset, the column names, the data types of variables, missing values, duplicates, and duplicate students. Since there were neither missing values nor duplicate records, there was no need for any imputation process or deletion of duplicates. The original categorical variable was used as such, while for the ordinal analysis, the categories were coded numerically as follows:

$$Y_i = \begin{cases} 0, & \text{if the performance level is Low} \\ 1, & \text{if the performance level is Medium} \\ 2, & \text{if the performance level is High} \end{cases} \quad (1)$$

This coding indicated the ordering of categories, but it did not imply the spacing between the neighboring categories is numerical equal.

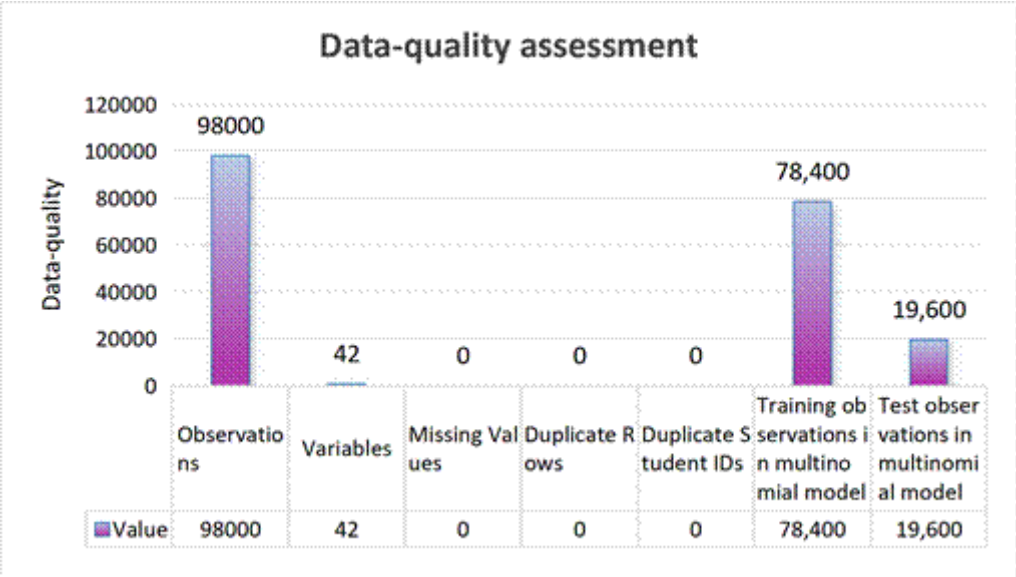


Figure 3. Data-quality Assessment.

Categorical predictors in regression models and machine-learning models were encoded by one-hot encoding. The omitted reference categories of these encoded variables were ignored to prevent the creation of dummy variables. Constant predictors were eliminated from the regression design matrix. The student ID number was not considered an explanatory variable.

3.5. Descriptive Statistical Analysis

Descriptive statistics have been done for the purpose of describing the sample and giving the distribution of the research variables. For numerical variables, the statistics were the number of observations, mean, standard deviation, minimum, first quartile, median, third quartile, and maximum. For categorical variables, frequency and percentages were done.

For numerical variable X, the sample mean is given by:

X	$=$	$(1/n)\sum_{i=1}^{n} X_i$	(2)

Where n denotes the number of observations. The sample standard deviation was calculated as

s_X	$=$	$\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$	(3)

For a category j, its percentage was calculated as:

$$njN\times100,$$

Pj	$=$	$\frac{n_j}{N}\times100$	(4)

Where Nj represents the number of observations in category, J, and N represents the total number of observations.

Descriptive analysis was used to summarize the sample, examine the balance of outcome categories, and provide context for the subsequent association tests.

3.6. Spearman Rank-Correlation Analysis

The method of Spearman's rank correlation was employed to find out the presence of any monotonous relationship between the ordinal outcome measure and certain numeric variables. Variables considered were attendance, study time, past academic performance, homework completion, reading proficiency, mathematics proficiency, science proficiency, language proficiency, sleeping time, use of screens, bullying, and family size.

Spearman's rank correlation is defined by the formula:

ρs	$=$	$Corr(R(X),R(Y))$	(5)

Where $R(X)$ and $R(Y)$ denote the ranks of the corresponding variables. When there are no tied ranks, the coefficient can also be expressed as

$$\rho s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (6)$$

where d_i is the difference between the ranks of the two variables in observation i. The use of the rank-correlation method was because the outcome categories were ranked while there was no need for a linear relationship between the numeric values.

The coefficient was then evaluated based on its sign and size. The p-value associated with the correlation was used to evaluate the evidence against the hypothesis that there was no correlation between the ranks. A cutoff point of $p<0.05$ $p<0.05$ was used in the analyses presented here. Because of the large sample size, the size of the coefficient was considered alongside its statistical significance.

3.7. Chi-Square Tests and Cramér's V

Chi-square test for independence was conducted to test whether there is an association between the distribution of performance score in different levels of some selected categorical independent variables. The variables that underwent testing include; gender, parental education, parental involvement, class participation, availability of learning material, access to internet, attending tutoring classes, SES, parental occupation status, school type, and school location.

A contingency table was constructed for each variable. The chi-square value was computed as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (7)$$

where O_{ij} denotes the observed frequency and E_ij denotes the expected frequency under the null hypothesis of independence. The expected frequency was calculated as

$$Eij = RiCjN, E_{ij} = \frac{R_i C_j}{N} \quad (8)$$

Where R_i and C_j denote the corresponding row and column totals.

Cramér's V was calculated to describe the strength of association between each categorical predictor and academic performance:

V	=	$\sqrt{\frac{\chi^2}{N\min(r-1,c-1)}}$	(9)

where N is the total number of observations, r is the number of rows, and c is the number of columns in the contingency table.

Chi-square test revealed whether there was any statistically significant association while Cramér's V revealed its strength. Both these measures were used for the analysis of the results. Cross tabulation with row percentage was conducted for some of the variables with respect to performance levels of Low, Medium, and High.

3.8. Primary Multivariable Regression Analysis

An initial multivariable ordinal logistic regression model was used to investigate the association between education and socioeconomic covariate's and the ordered variable of academic performance taking into account the effects of other variables like demographic and school-related factors. These variables include attendance, study hours, parent's involvement, class participation, access to learning materials, internet access, homework completion rate, tutor attendance, education of parents, socioeconomic status, occupation status of parents, family size, age, sex, grade level, school type, and school location. Prior academic performance was not included in this initial model due to the reason explained in section 3.3.

The general proportional-odds formulation is

$\log\left[\frac{\Pr(Y_i \le k)}{\Pr(Y_i > k)}\right]$	=	$\alpha_k - \boldsymbol{\beta}^T \mathbf{X}_i, \quad k = 0,1$	(10)

Where Y_i is the ordered outcome of student **i,** α_k is the threshold for outcome category **k,** X_i is the set of explanatory variables, and β is the set of coefficients. The relationship between predictor variables and the cumulative odds of falling into or below a particular performance category would be defined through this formulation under the assumption of the model.

Categorical variables were handled through the use of indicator variables, leaving out one reference category per variable. Coefficients, their standard error, the z-statistic, the p-value, and the 95% confidence interval were analyzed. The odds ratio was calculated by exponentiating the regression coefficient estimate:

ORj	=	$exp(\hat{\beta}_j)$	(11)

An odds ratio above one indicates higher cumulative odds in the direction defined by the model parameterization, whereas an odds ratio below one indicates lower cumulative odds. Interpretation must account for the chosen reference category and the sign convention of the fitted model.

Model convergence was checked, and variance inflation factors were calculated as a diagnostic of multicollinearity among the encoded predictors. Statistical significance was assessed using $p<0.05p<0.05$ Coefficient magnitude, uncertainty, and substantive relevance were considered alongside p-values.

3.9. Supplementary Multinomial Logistic Regression

Apart from those models, the multinomial logistic regression model was additionally used for prediction comparison purposes. The dataset was divided into training and test datasets using the 80:20 stratified split method with a fixed random state value of 42. Thus, the 78,400 data points were used for the training process, while 19,600 were used for testing.

Numerical features were standardized using the standardization pipeline, and a multinomial logistic regression model was fitted based on the training data. Moreover, the fitted model was tested using the test dataset.

The performance evaluation measures consisted of accuracy, balanced accuracy, macro-averaged F1-score, and multiclass ROC-AUC. Confusion matrix and precision, recall, and F1-score per class were calculated for error distribution analysis by performance classes.

Accuracy was measured as

$$Accuracy = \frac{Numberofcorrectpredictions}{N_{test}} \quad (12)$$

Balanced accuracy was calculated as the mean recall across the K classes:

$$Balanced\ Accuracy = \frac{1}{K}\sum_{k=1}^{K} Recall_k \quad (13)$$

The F1-score for class k was calculated as

$$F1k = 2\frac{Precision_k \cdot Recall_k}{Precision_k + Recall_k} \quad (14)$$

and the macro F1-score was calculated as

$$Macro-F1 = \frac{\frac{1}{K}\sum_{k=1}^{K} F1_k}{\frac{1}{K}\sum_{k=1}^{K} F1_k} \quad (15)$$

.

Macro averaging considers all classes equally. This method becomes very helpful when determining how well the classifier can identify all performance classes rather than just the more common ones.

3.10. Supplementary XGBoost Model Comparison

A separate XGBoost comparison evaluated three configurations using the same stratified five-fold cross-validation framework:

1. **Model A:** SelectKBest feature selection with k=10k=10, without class weighting.
2. **Model B:** All encoded features, without class weighting.
3. **Model C:** All encoded features, with class weighting.

The comparison did not consider the student id or previous performance of the student in the predictor set since the later had a high association with the outcome and there was uncertainty regarding the target construction overlapping.

For each configuration, the cross-validation results were described by the mean and standard deviation of the accuracy, balanced accuracy, macro F1-score, macro one-versus-rest ROC-AUC and class-specific recall measures. The macro F1-score was used to identify the configuration of choice during the cross-validation process. The out-of-fold test set was used after the comparison was completed and it did not influence the model selection.

The intention of this comparison was to investigate how feature selection and class-weighting influenced the distribution of predictive performance among the different outcomes. It was not aimed at identifying causal relationships nor substitute the analysis of the associations. Class-weighting may influence the recall and the prediction distribution without necessarily improving the model's discrimination capacity.

3.11. Statistical Software and Reproducibility

The workflow was developed in Python making use of libraries such as pandas, NumPy for data manipulation, SciPy for statistical tests, Matplotlib for graphical representation, statsmodels for regression, and scikit-learn for preprocessing, train/test splitting, and predictive model evaluation purposes. XGBoost has been employed for machine learning comparison.

The train/test split of the data according to the multinomial method has been conducted by means of a fixed random seed of 42, whereas for the XGBoost comparison a stratified five-fold cross-validation has been utilized within all three setups. The results have been exported in Excel and CSV files, while the figures were saved in PNG format.

3.12. Methodological Limitations

The data is observational and, hence, the analysis cannot prove any causation between the investigated educational or socioeconomic variables and differences in academic achievement. There is high correlation between previous academic achievement and the outcome in question, which is an indication of an overlap between the outcomes issue. However, there was no information on the data generation process by the official organization. Moreover, very large sample sizes may lead to statistical significance of even small correlations; hence, the magnitude and confidence interval should be considered.

The supplementary models revealed weak discrimination power during the evaluation. These findings were taken to indicate the performance of the models tested.

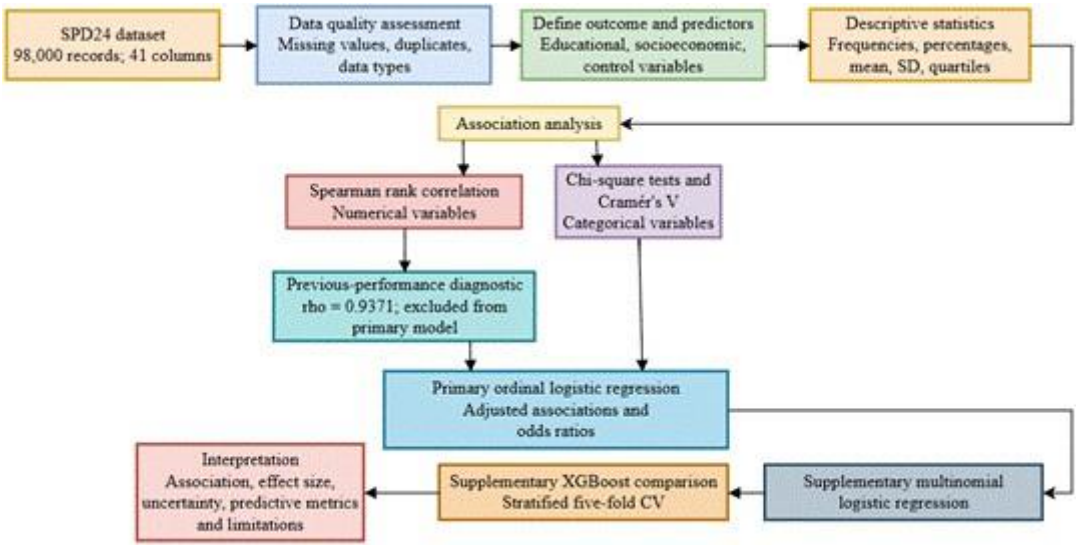


Figure 4. Data Analysis Workflow for Examining Educational and Socioeconomic Factors Associated with Academic Performance.

Here is the figure 4 demonstrating the analytic procedure used on the SPD24 dataset, containing 98,000 observations and 41 variables. The analysis starts with assessing data quality and establishing variables – outcome, predictor, and controls – followed by descriptive statistics and association analysis. Spearman rank correlation is used for numeric variables, and chi-square tests along with Cramer's V for categorical variables. Previous performance is evaluated as a diagnostic variable and excluded from the main regression model. Adjusted associations and odds ratios are calculated using ordinal logistic regression, while multinomial logistic regression and XGBoost are used as secondary approaches for comparison purposes.

4. RESULTS

4.1. Dataset Characteristics and Distribution of Academic Performance

The analysis was performed using the dataset named "SPD24 – Student Performance Dataset," that included 98,000 observations for students and 41 attributes. During the initial assessment of the data quality, the absence of any missing values, duplicates or duplicate Student ID's was found. The variable to be predicted was Performance Score and it comprised three classes – Low, Medium, and High.

As can be seen from Table 1 below, the share of the High class and Medium class was equal and comprised around 37.66%. The Low class was represented by 24.68% of observations. Even though the classes are slightly unbalanced, the difference between their shares was moderate. It is necessary to know this information for correct interpretation of the classification results since there is a possibility that the model will predict well in general but badly in a certain class.

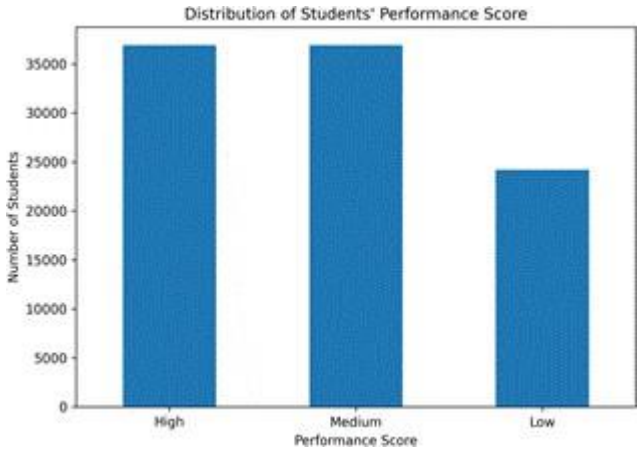


Figure 5. Distribution of academic performance categories in the SPD24 dataset.

4.2. Association Between Educational Factors and Academic Performance

Monotonic relations between numerical or ordinal variables and academic performance were analyzed using Spearman's rank-correlation test. In the case of the Performance Score, ordinal values were used: Low = 0, Medium = 1, and High = 2.

The relationship between the Previous Academic Performance and the Performance Score proved to be highly positive $(\rho=0.9371, p<0.001)$. The mean Previous Academic Performance values were 92.499 for the high-performing category, 77.475 for the medium category, and 65.003 for the low-performing category. The obtained data show that students who belonged to a better performance category had relatively high values for the previous academic performance.

The magnitude of the correlation requires special attention because it raises concerns about the possible target-overlap problem. Specifically, the high correlation and similar patterns of the group distribution make us think that Previous Academic Performance may contain significant part of information contained in the target variable. The provided data do not allow drawing conclusions about whether the target variable is based on this one. Among the other examined variables, the observed Spearman correlations were close to zero. Reading Proficiency showed a correlation of ($\rho=-0.0053$, $p=0.100184$), Math Proficiency showed ($\rho=0.0035$, $p=0.276049$), and Science Proficiency showed ($\rho=-0.0045$, $p=0.155550$). Language Proficiency was also close to zero ($\rho=-0.0002$, $p=0.9551$). Attendance Rate, Study Hours, and Homework Completion Rate similarly showed negligible correlations with the ordinal performance outcome.

Table 2. Spearman rank correlations with academic performance.

Variable	Spearman's ρ ho	p-value	Interpretation
Previous Academic Performance	0.9371	<0.001	Very strong positive association; investigate possible target overlap

Note: Correlations are reproduced from the supplied analysis output. A coefficient close to zero indicates little evidence of a monotonic relationship in this analysis; it does not rule out every possible nonlinear or subgroup relationship.

However, based on the correlations analysis, it can be seen that only one of these variables had a significant monotonic relationship to the dependent variable. All the other numerical or ordinal variables had a very small correlation coefficient, which was not statistically significant at the 0.05 level. Therefore, no significant monotonic relationship was found between the individual variables and the Performance Score variable.

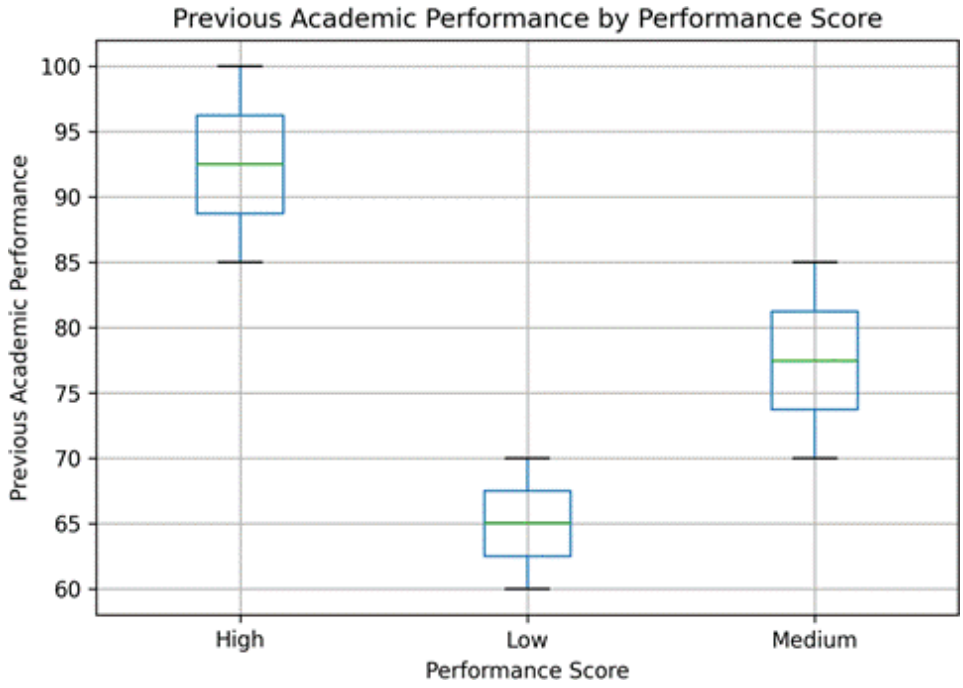


Figure 6. Distribution of Previous Academic Performance across current performance categories. This figure illustrates the strong association and helps assess possible target overlap.

4.3. Associations Between Categorical Factors and Academic Performance

Test for independence using chi-square tests was applied to evaluate whether there is a variation in the distribution of Performance Score across categories of the selected demographic, educational, and socioeconomic variables. Cramér's V measure was used to determine the extent of each categorical association. From the analysis, a statistically significant association was found between Parental Education Level and Performance Score ($\chi 2 = 19.4585$, $V = 0.0100$, $p<0.05$). But the Cramér's V measure was very low. This means that the association had a weak effect size despite being statistically significant. None of the other categorical variables provided in the given output revealed statistically significant association at the 0.05 level of significance. They include Gender, Parental Involvement, Class Participation, Availability of Learning Materials, Access to the Internet, Attended Tutoring Class, Socioeconomic Status, Parental Employment Status, School

type, and School Location. Their Cramér's V measures are also very low.

Table 3. Chi-square tests and Cramér's V for categorical factors.

Categorical factor	Chi-square statistic	Cramér's V	Significant at 0.05?
Gender	0.5954	0.0025	No
Parental Education Level	19.4585	0.0100	Yes
Parental Involvement	0.7853	0.0020	No
Class Participation	4.5958	0.0048	No
Access to Learning Resources	0.6264	0.0025	No
Internet Access	2.7236	0.0053	No
Attendance of Tutoring Sessions	0.0764	0.0009	No
Socioeconomic Status	6.0057	0.0055	No
Parental Employment Status	4.2422	0.0066	No
School Type	3.2299	0.0057	No
School Location	2.3970	0.0035	No

Note: The supplied results identify Parental Education Level as significant at $p<0.05$. Exact p-values and degrees of freedom for all categorical tests were not included in the results summarized here and should be copied from the original statistical output before final submission.

Cramér's V was calculated as:

$$V = \sqrt{\frac{\chi^2}{n_{\min}(r-1,c-1)}} \quad (19)$$

Where χ2 is the chi-square value, n is the sample size, and r and c represent the number of categories in each variable.

On the whole, Parental Education Level appears to be the only categorical variable with statistically significant chi-square values among the provided variables. However, due to the very low effect size, it can be said that the connection observed was rather weak.

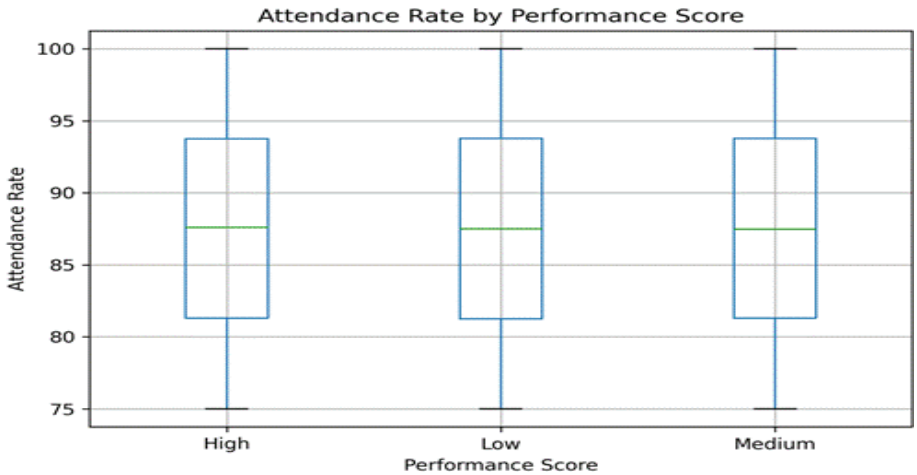


Figure 7. Distribution of Attendance Rate Across Academic Performance Categories.

Fig 7 shows the attendance rate distribution for students who have been categorized into High, Medium, and Low academic performance classes. The medians and interquartile ranges being similar indicate that there is little difference in attendance rates between the three classes.

4.4. Multivariable Ordinal Regression Results

An ordinal logistic regression analysis was conducted to assess whether the relationship between the chosen predictors and educational attainment would be visible with their combined consideration in the model. There were 24 predictors coded and 98,000 observations. The output from the optimization process suggested convergence of the model. Of the predictor coefficients listed, only one reached the 0.05 significance level – the Low category of the Socioeconomic Status with β=−0.0290, p=0.0446. Its odds ratio was equal to 0.9714, while its 95% confidence interval for the coefficient was [−0.057,−0.001] approximately. This corresponds to an odds ratio that is very close to 1, which indicates the existence of a small relationship on the odds scale. Parental Employment Status – Unemployed had β=-0.0225, OR=0.9778 and p=0.0564. Parental Internet Access - Yes did not reach the 0.05 significance threshold either (β=−0.0172, OR=0.9830, p=0.1447).

Table 4. Selected results from the ordinal regression model.

Predictor	Coefficient (β)	Odds ratio	p-value	Interpretation at 0.05
Socioeconomic Status: Low	-0.0290	0.9714	0.0446	Statistically significant; very small association
Parental Employment Status: Unemployed	-0.0225	0.9778	0.0564	Not statistically significant
Internet Access: Yes	-0.0172	0.9830	0.1447	Not statistically significant
Predictor	**Coefficient (β)**	**Odds ratio**	**p-value**	**Interpretation at 0.05**

Note: Categorical Coefficients listed are relative to the reference categories in the original coded regression. The full regression output table should contain all variables, standard errors, confidence intervals, p-values, and reference categories. The sign of the effect must be interpreted based on the precise cumulative-link convention followed in the model.

The odds ratio was calculated from the coefficient as:

$$OR = e^{\beta} \quad (19)$$

In regard to Socioeconomic Status—Low, the obtained odds ratio turned out to be 0.9714. Despite being statistically significant in the output presented above, this value is fairly close to 1; hence, this indicates the weak connection between the variable in question and the dependent one. Consequently, it cannot be considered as the evidence of a strong socioeconomic association. There appears to be not much information in the regression output regarding associations between the socioeconomic variables and the ordinal dependent variable after controlling for other predictors. The relation between Socioeconomic Status—Low turned out to be quite weak, while the relation between parental unemployment seemed promising yet not statistically significant.

4.5. Supplementary Classification Model Performance

Classification models were used as part of the additional analysis to see whether the selected variables could classify between the Low, Medium, and High performance levels. The evaluation of the model included accuracy, balanced accuracy, macro-averaged F1-score, and multiclass ROC-AUC. According

to the classification output reported here, accuracy is 0.3808, balanced accuracy is 0.3370, macro-F1 is 0.2896, and multiclass ROC-AUC is 0.5007. Those values show poor discrimination performance of the model. Specifically, the ROC-AUC value is close to 0.5, which implies that the model could hardly differentiate between the three groups. The recall for the Low class is 0, whereas for the Medium and High it is around 0.510 and 0.500 respectively. That implies that the model is not able to classify Low performing students among those three classes, although it obtained a higher than 0.33 accuracy.

Table 5. Summary of supplementary classification performance.

Metric	Reported value
Accuracy	0.3808
Balanced Accuracy	0.3370
Macro-F1	0.2896
Multiclass ROC-AUC	0.5007
Recall: Low	0.000
Recall: Medium	Approximately 0.510

Note: that the above three recall figures for each class are based on the classification output provided and therefore should not be referred to as ordinal-regression performance unless the actual code used was for an ordinal regression model. Also, the approximate values of Medium and High categories above should be substituted by their precise values.

Weak classification performance is in line with very weak correlation and categorical effect size found for most variables. It should be noted, however, that while association and prediction have something in common, there can be a statistically significant relationship between a variable and the dependent one, but the predictive value of the variable can be weak.

4.6. Comparison of XGBoost Feature-Selection and Class-Weighting Strategies

The following three settings of XGBoost were evaluated using stratified five fold cross-validation:
(A) Feature Selection of Select K Best where k = 10 and no class weight; (B) All the encoded features without any class weight; and (C) All the encoded features with class weight. There were two variables that were not considered when comparing the models: Previous Academic Performance and Student ID. In the case of unweighted models, the all feature model (B) provided better cross-validation accuracy (0.3783) and Macro F1 (0.2885). But in the case of class weighted models (C), although it provided poor cross-validation accuracy (0.3385), it offered much higher Macro F1 (0.3343) and Balanced Accuracy (0.3365). Further, the value of Low-class recall is 0.3212, which is much better than the other two models.

Table 6. XGBoost model comparison using stratified five-fold cross-validation.

Metric	A: SelectKBest, k=10k=10	B: All features, unweighted	C: All features, class-weighted
Accuracy, mean	0.3761	0.3783	0.3385
Accuracy, SD	0.0029	0.0053	0.0018
Balanced accuracy, mean	0.3329	0.3349	0.3365
Balanced accuracy, SD	0.0026	0.0047	0.0029
Macro-F1, mean	0.2863	0.2885	0.3343
Macro-F1, SD	0.0024	0.0041	0.0023
ROC-AUC, mean	0.5025	0.5013	0.5024
ROC-AUC, SD	0.0035	0.0044	0.0030
Low-class recall, mean	0.0004	0.0012	0.3212
Medium-class recall, mean	0.5015	0.5067	0.3468
High-class recall, mean	0.4968	0.4970	0.3416

Note: SD denotes standard deviation across the five cross-validation folds. Values are reproduced from the supplied comparison output.

The findings indicate a compromise between accuracy and performance by classes. Class balancing improved macro F1-Score and classification of Low-performers, but at the cost of accuracy. On the other hand, the non-class-balancing all-feature model attained the highest accuracy but could not classify Low-performers at all.

Test data exhibited a similar trend.

Table 7. XGBoost comparison on the held-out test set.

Metric	A: SelectKBest, k=10k=10	B: All features, unweighted	C: All features, class-weighted
Accuracy	0.3746	0.3789	0.3330
Balanced accuracy	0.3316	0.3356	0.3323
Macro-F1	0.2849	0.2890	0.3296
ROC-AUC	0.5021	0.5018	0.4989
Low-class recall	0.0000	0.0014	0.3271
Medium-class recall	0.5132	0.5197	0.3440
High-class recall	0.4816	0.4855	0.3258

In the held-out test set, the all-features unweighted model had the best accuracy score (0.3789), while the class weighted model had the best macro-F1 score (0.3296). The class weighted model also had a Low-class recall of 0.3271, while the all features unweighted model had a Low-class recall of 0.0014. However, all three ROC-AUC scores in the test sets were very close to 0.5, suggesting poor discrimination.

Therefore, the above results do not point to any one particular model as superior to the others. Instead, it shows that the choice of model depends on the objective: unweighted models performed better in terms of accuracy, while the class weighted models balanced out classification performance across all the three classes. As the aim of this research paper is to find relationships between variables, rather than to predict the outcomes accurately, the above results are considered supplementary to this end.

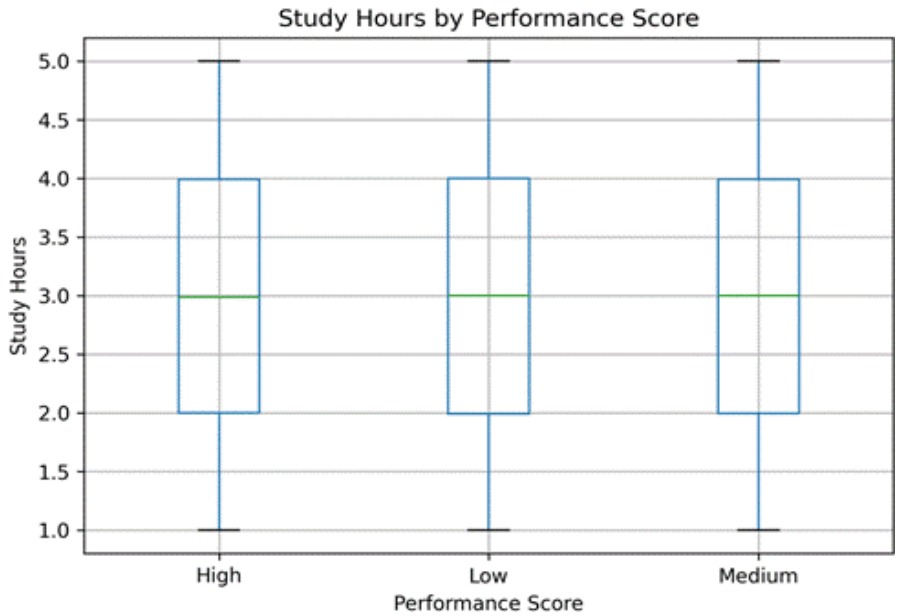


Figure 8. Distribution of Study Hours Across Academic Performance Categories.

Fig 8 illustrates the differences in study hours among the High, Medium, and Low categories of academic performance. The similarities in medians and distributions suggest that there is not much difference in the study hours among the categories.

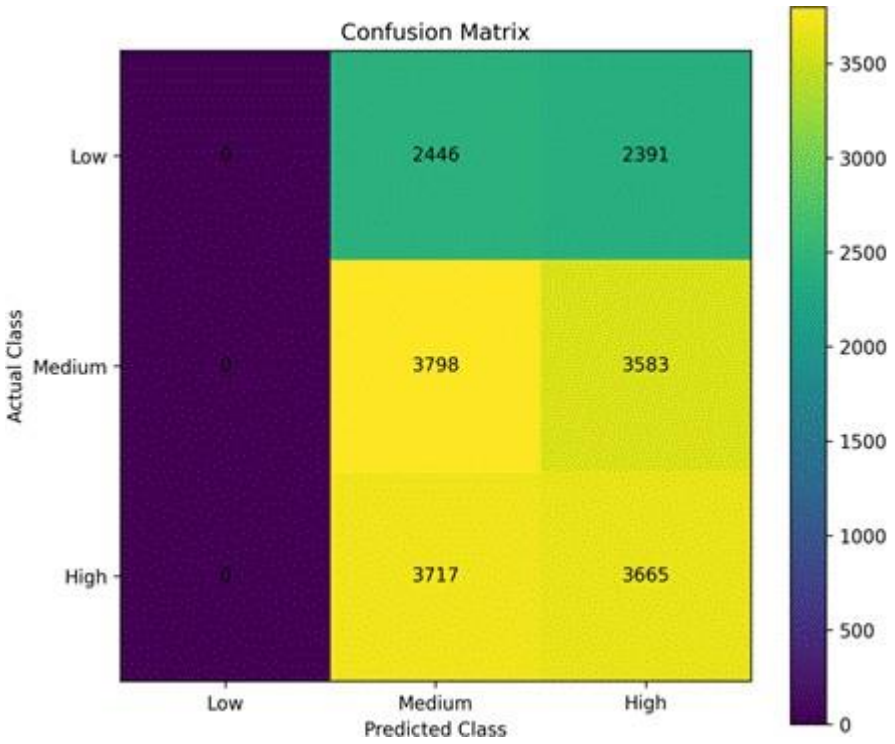


Figure 9. Confusion Matrix of the Academic Performance Classification Model.

The confusion matrix compares the actual classes to the classes predicted by the model. The model fails to recognize the Low-performance students and often categorizes them as Medium and High categories. This shows that the model has bad classification accuracy for the Low class.

4.7. Relative Importance of the Findings

Taken together, the results show different types of evidence on the research questions in the study. First, Previous Academic Performance had a very high correlation with the Performance Score. Since the magnitude of this association is exceptionally high compared with all other correlations, it must be interpreted separately from other variables that can serve as independent predictors without checking for any target overlap. Second, all other educational variables included in the analyses, such as Reading, Math, and Science Proficiency, Attendance Rate, Study Hours, and Homework Completion Rate, have correlations near 0 based on the reported Spearman analysis. This means that there is no evidence of strong monotonic relationships between the variables and the dependent variable in the whole sample. Third, out of the analyzed categorical variables, Parental Education Level turned out to be statistically significant, although its Cramér's V was equal to 0.0100, which indicates a very weak association. Also, out of the output of the ordinal regression, only the Low SES level was statistically significant with a small odds ratio near 1. It seems that statistical significance is not equal to practical importance. In the end, the machine learning comparisons showed that the selected predictors had little ability to classify performance groups effectively. Weighting for class did improve the ability to classify low performers and the macro-F1 measure but decreased overall accuracy. These findings highlight the differences between detecting statistical relationships and constructing a prediction model.

Generally, the results indicate that the vast majority of the educational and socioeconomic factors considered in this study have been found to show a relatively weak relationship with academic performance. However, Previous Academic Performance has been found to have a strong relationship which deserves further investigation. It is important to note that the results only indicate associations and not causations in the SPD24 data set.

5. CONCLUSION

The current study assessed the relationships between educational attributes, socioeconomic factors, and students' performance using the SPD24 Student Performance Dataset which included 98,000 records of students and 41 variables. The analysis used such statistical methods as descriptive statistics, Spearman rank correlation, test of independence chi-square, Cramér's V estimation of effect size, ordinal logistic regression, and additionally multinomial logistic regression and classification using XGBoost. It is reported that 37.66% of students were identified as High-performing, 37.66% as Medium-performing, and 24.68% as Low-performing. Academic performance in the previous year was found to have very strong relationship with the outcome (ρ=0.9371); thus, it was excluded from the regression model due to potential overlap with the outcome definition. Among categorical attributes, parental education was statistically significant in relation to students' performance although it had negligibly small effect size (Cramér's V=0.0100). In ordinal logistic regression, low socioeconomic status was the only attribute found to be statistically significant at 0.05 level ((β=−0.0290)); However, the small value of the coefficient needs to be interpreted with caution, and the proportional odds assumption must be validated. The models were shown to have low performance in classification. Multinomial logistic regression resulted in 38.08% accuracy, 33.70% balanced accuracy, 0.290 macro-F1, and 0.501 multiclass ROC-AUC. Unweighted XGBoost yielded 37.89% test accuracy, 0.289 macro-F1, and recall for the

Low category equal to 0.0014. Class weighting led to an increase in recall to 0.3271 and macro-F1 to 0.3296 but decreased the overall accuracy to 33.30%. There are practical implications of the research findings regarding educational data mining and student support. Individual socio-economic features should not be considered determinants of academic potential, and evaluation of predictive models should involve recall, balanced accuracy, macro-F1, etc.

The poor level of recall for Low-performers suggests that the models under evaluation cannot be utilized as standalone early warning systems because they need more work. Future research should take into account other contextual and longitudinal variables, validate the construction of the performance outcome variable, test the model assumptions and assess the generalizability on independent datasets. From the theoretical point of view, this study is a contribution to the field of educational data mining since it proves the importance of differentiating between significance and importance. A small effect size of the parent's education level and a small ordinal regression coefficient for the low socioeconomic status are examples of insignificant results despite statistical significance demonstrating that significant results do not automatically mean the presence of a relationship. Additionally, the poor predictive power proves that having a large dataset and using relevant predictors does not ensure enough information is available for separating the groups. Since the study is observational and relies on one dataset, the results cannot prove causation or prove that the variables studied are irrelevant in other educational settings. In conclusion, this study gives an insight into the specific educational and socioeconomic determinants of academic performance without making any unfounded assumptions.

REFERENCES

1. Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237.
2. Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. *Frontiers in psychology*, *12*, 698490.
3. Roslan, M. B., & Chen, C. (2022). Educational data mining for student performance prediction: A systematic literature review (2015-2021). *International Journal of Emerging Technologies in Learning (iJET)*, *17*(5), 147-179.
4. Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & education*, *143*, 103676.
5. Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting student academic performance at higher education using data mining: A systematic review. *Applied Computational Intelligence and Soft Computing*, *2022*(1), 8924028.
6. Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational technology research and development*, *68*(4), 1961-1990.
7. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International journal of educational technology in higher education*, *17*(1), 3.
8. Pisa, O. E. C. D. (2023). results (volume I): the state of learning and equity in education. *PISA. Paris*.
9. Rodríguez-Hernández, C. F., Cascallar, E., & Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review*, *29*, 100305.
10. Fouladvand, S., Noshad, M., Goldstein, M. K., Periyakoil, V. J., & Chen, J. H. (2023). Mild cognitive impairment: data-driven prediction, risk factors, and workup. *AMIA Summits on Translational Science Proceedings*, *2023*, 167.
11. Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and information technologies*, *28*(7), 8299-8333.
12. García-Martínez, I., Fernández-Batanero, J. M., Fernández-Cerero, J., & León, S. P. (2023). Analysing the impact of artificial intelligence and computational sciences on student performance: Systematic review and meta-analysis. *Journal of New Approaches in Educational Research*, *12*(1), 171-197.
13. Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., ... & Gates, K. M. (2023). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior research methods*, *55*(6), 3026-3054.
14. Antoninis, M., Alcott, B., Al Hadheri, S., April, D., Fouad Barakat, B., Barrios Rivera, M., ... & Weill, E. (2023). Global Education Monitoring Report 2023: Technology in education: A tool on whose terms?.
15. Chaka, C. (2021). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews.
16. Siddiqui, S., & Kureshi, N. (2026). Identifying At-Risk Students in Higher Education: A Multi-Institutional Study for Data-Driven Intervention and Student Success. *Sage Open*, *16*(3), 21582440261488832.

17. Oppong, S. O. (2023). Predicting students' performance using machine learning algorithms: a review. *Asian Journal of Research in Computer Science*, *16*(3), 128-148.
18. Rodrigues, L. S., dos Santos, M., Costa, I., & Moreira, M. A. L. (2022). Student performance prediction on primary and secondary schools-a systematic literature review. *Procedia Computer Science*, *214*, 680-687.
19. Carrillo-López, P. J., Constante-Amores, A., Arroyo-Resino, D., & Sánchez-Munilla, M. (2022). Self-Concept and Academic Achievement in Primary School: A Predictive Study. *International Journal of Education in Mathematics, Science and Technology*, *10*(4), 1057-1073.
20. Pirie, W. (2014). Spearman rank correlation: Introduction. *Wiley StatsRef: Statistics Reference Online*.
21. Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision analytics journal*, *7*, 100204.
22. Alalawi, K., Athauda, R., & Chiong, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, *5*(12), e12699.
23. Rakesh, D., Lee, P. A., Gaikwad, A., & McLaughlin, K. A. (2025). Annual Research Review: Associations of socioeconomic status with cognitive function, language ability, and academic achievement in youth: a systematic review of mechanisms and protective factors. *Journal of Child Psychology and Psychiatry*, *66*(4), 417-439.
24. Zhang, F., Jiang, Y., Ming, H., Ren, Y., Wang, L., & Huang, S. (2020). Family socio-economic status and children's academic achievement: The different roles of parental academic involvement and subjective social mobility. *British Journal of Educational Psychology*, *90*(3), 561-579.
25. Cui, T., & Liu, Q. (2025). Parental involvement and academic achievement in Year 4 and Year 8 students: A meta-analysis based on large-scale assessments in China. *British Educational Research Journal*, *51*(6), 2719-2739.
26. Wang, X., & Wei, Y. (2024). The influence of parental involvement on students' math performance: A meta-analysis. *Frontiers in Psychology*, *15*, 1463359.
27. Kim, S. (2022). Fifty years of parental involvement and achievement research: A second-order meta-analysis. *Educational Research Review*, *37*, 100463.
28. Otani, M. (2020). Parental involvement and academic achievement among elementary and middle school students: M. Otani. *Asia Pacific Education Review*, *21*(1), 1-25.
29. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(4), 1-21.
30. Cohausz, L., Tschalzev, A., Bartelt, C., & Stuckenschmidt, H. (2024). Investigating demographic features and their connection to performance, predictions, and fairness in EDM models. *Journal of Educational Data Mining*, *16*(2), 177-213.
31. Yanagiura, T., Yano, S., Kihira, M., & Okada, Y. (2023). Examining algorithmic fairness for first-term college grade prediction models relying on pre-matriculation data. *Journal of Educational Data Mining*, *15*(3), 1-25.