International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I7P126

Volume 74 Issue 7, 428-448, July 2026
© 2026 Seventh Sense Research Group®

Original Article

Uncovering Latent Academic Drivers: A Multivariate Factor and Regression Analysis Across Different Board of Education

Monisha S[1], Edwin Raj A[2]

[1,2] *Hindustan Institute of Technology & Science (HITS), Rajiv Gandhi Salai (OMR), Padur, Chennai, Tamil Nadu, India.*

[2]*Corresponding Author : ewin.raj@gmail.com*

Received: 28 January 2026 Revised: 20 June 2026 Accepted: 24 June 2026 Published: 28 July 2026

Abstract - *This paper looks at the interdependence and independency of subject-wise academic scores, as well as identify factors that affect student performance, across Matriculation, CBSE and Government school systems. A descriptive statistical analysis, correlation analysis, graphical visualization, Exploratory Principal Component Analysis and Factor Analysis, regression models, and machine-learning techniques were used to analyze the data of 540 higher secondary students. The analysis revealed two large latent academic dimensions, namely, STEM ability and Language Ability. STEM ability was found to have high loadings in Physics, Chemistry, Biology and Mathematics, whereas Language ability showed high loadings in English and Language. The two combined explained 77.4% of the overall Variance, indicating that student performance can be captured on fewer and more understanding academic dimensions. Models of prediction of latent academic scores included OLS, Ridge, Lasso, Random Forest, and XGBoost. The OLS and Ridge Regression models have the best testing results, with the value of R to 0.974 and 0.973, respectively, and in most cases, the mean absolute error is about 2.10. Permutation-importance analysis showed that science-related subjects strongly contributed to latent academic performance. Propensity-score-weighted analysis indicated that section assignment had no statistically significant effect on latent performance scores. The findings support the use of interpretable linear models for structured educational data and show that latent academic patterns can assist in targeted academic interventions. However, as the study is cross-sectional, future research should include longitudinal data and socioeconomic factors to improve causal interpretation and explanatory strength.*

Keywords - *Academic Performance, Educational Data Analysis, Exploratory Factor Analysis, Machine Learning, Principal Component Analysis, Regression Analysis.*

1. Introduction

Factor Analysis (FA) plays a critical role in explaining the importance of curriculum selection in higher secondary education by identifying latent configurations of the various factors that provide academic discretion. Students today are forced to select individual subjects: science, commerce and the arts, which significantly influence their further academic paths and employment. However, such choices are often informed by a plethora of factors, which include personal interests and academic achievements, peer pressure, parental pressure, school culture, as well as the opportunities in career choices. Quantitative differences in analytical approaches permit them to answer different questions of research and theoretical concepts, which leads to their extensive use in diverse fields. However, the versatility of the statistical methods also leads to ongoing debates regarding their proper application [1]. Factor analysis helps in the simplification of complex sets of items or variables through the identification of underlying dimensions by explaining the relationship among them. During analysis, the relationship of inter-items in the instrument, comprised of 20 items, can be able to generate 400 correlations in a basic analysis, which can be challenging to mental process. Factor analysis simplifies these correlation matrices, helping the researcher to a great extent in understanding the relationship in the scale among various items and may share the common underlying factors. Factor analysis is widely used and recommended to refine and development of clinical assessment tools as it provides evidence supporting the construct validity of the measures [2].

Factor analysis rests on the assumption that all variables are somewhat interrelated. Variables should be measured at a minimum ordinal level. Moreover, factor analysis requires a relatively large sample size, with a commonly accepted guideline being a ten-to-one ratio of observations to variables [3]. The impact of several factors to examine academic performance, concentrating on various stages of education [4]. Factor analysis should adhere to best practices found in the

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

literature in the education domain, especially regarding proper techniques and clear reporting of results. Reviewers must ensure that researchers explain their factor analysis methods accurately and use correct terminology. Both researchers and reviewers need to confirm that the chosen factor model aligns with the study's objectives. Oblique rotation should generally be applied unless there is a justified, theoretical reason for using orthogonal rotation. Before deciding which factors to retain, multiple solutions should be evaluated. Final interpretations of the factors must be informed by an understanding of the variables and a thorough review of all factor loadings [5].

Factor analysis primarily involves two approaches:

Confirmatory Factor Analysis (CFA)

Exploratory Factor Analysis (EFA)

Factor analysis is especially helpful in finding hidden patterns in the interactions between academic, social, and psychological factors in that domain [6, 7]. For example, factors like teacher influence, interest in particular subjects, and parental expectations may be grouped together to create a single latent factor called "External Motivation." Factor analysis enables determining salient variables based on which the academic preferences of students are formed, with the help of systematic identification and categorization of correlated indicators.

Factor analysis can be used to improve equity and the quality of education as it provides a necessary insight of performance differences among the various types of schools, such as government schools, Matriculation schools and CBSE schools. Exploratory Factor Analysis is generally used to evaluate the dimensionality and is commonly applied at an initial stage of research to investigate the inter-relations between a group of items or variables. Contrary, Confirmatory Factor Analysis is a more complex and sophisticated set of steps that are applied to empirical validation of the theoretical models about the very characteristics of the variables' connection.

EFA is typically used in education studies at the early phases of data collection development, especially to test the latent variables inaccessible in the measured form. When the investigation of the future EFA has already been carried out on the target tool, CFA develops on the previous results of the correlation, allowing the researchers to approve the dimensions or structures that were previously determined.

Factor analysis helps in informed decision-making on curriculum choice as it helps clarify the dimensions of intricate trends. The explanation of educational disparities also requires the recognition of performance differences according to the typology of schools, thus explaining how the institutional settings, quality of instruction, and resource distribution affect student achievement. This kind of information can guide the development of specific policies and procedures that would ensure that all learners are given an equal opportunity.

1.1. The Use of Factor Analysis

By focusing on key factors instead of an overwhelming number of potentially insignificant variables, factor analysis aids in categorizing variables being categorized as the right groups. Other applications of factor analysis include scaling, mapping, hypothesis testing, and data transformation [6].

- To conduct a factor analysis, the data must exhibit both univariate and multivariate normality [7].
- Additionally, there must be no univariate or multivariate outliers present [8].
- A key assumption in factor analysis is that there will be a linear correlation between the variables and factors by calculating the correlations [9].
- For a factor to be identified, it typically must include three variables, and it varies depending on the study's design [10].
- Generally, it is prudent to interpret rotated factors that contain two or less variables. It ought to involve 5-10 observations of each variable, and the volume of the sample ought to be composed of 300 participants [11].

2. Literature Review

Factor analysis has been widely applied for identifying hidden structures, reducing data complexity, and supporting theory development in research problems involving multiple observed variables [12]. Its use in recognizing hidden factors also demonstrates its relevance for decision-making problems where several criteria influence prioritization and interpretation [13]. In institutional and educational contexts, multiple factor analysis has been used to explore university performance, showing that multivariate approaches can organize complex educational indicators into meaningful dimensions for evaluation [14].

Multivariate statistical procedures provide the methodological foundation for examining relationships among variables, validating assumptions, and interpreting results in a systematic manner [15]. Practical data analysis guidance also emphasizes the importance of data screening, reliability checking, sample adequacy, and correct use of statistical software before conducting factor-based or predictive analysis [16]. Determining the appropriate number of factors is a critical step because incorrect factor retention can weaken the validity of the extracted structure and reduce the reliability of interpretation [17]. Similarly, multivariate analysis requires careful attention to correlation patterns, factor loadings, model assumptions, and overall suitability of the dataset for valid statistical interpretation [10].

A growing number of recent educational data mining projects have concentrated on applying computational models to forecast students' academic achievement. Machine learning methods have been shown to support Early detection of pupils who could need academic support, and they provide useful tools for improving academic monitoring and institutional decision-making [18]. Comparative studies indicate that prediction performance is strongly affected by algorithm selection, dataset characteristics, preprocessing strategy, feature selection, and evaluation metrics [19]. Broader reviews of machine learning in educational data show that the field has expanded rapidly, but many studies still face limitations related to generalizability, methodological consistency, feature interpretation, and reproducibility [20].

Interpretability has become an important concern in academic performance prediction because educational models should not only provide accurate outputs but also explain the factors influencing learner outcomes [21]. Predictive studies using educational data mining and hyperparameter optimization show that model tuning can improve performance; however, the effectiveness of such techniques depends on the quality of the dataset, the selected variables, preprocessing decisions, and the validation strategy used [22]. Data-driven approaches also demonstrate that predictive modeling can help institutions monitor student progress, detect academic risk, and design timely interventions for improving learning outcomes [23]. Similar educational data mining research confirms that student performance datasets can reveal useful academic patterns, although many models remain focused mainly on performance metrics rather than detailed explanatory interpretation [24].

The reviewed studies show that factor-based analysis and educational data mining are both important for academic performance research. Factor analysis supports the identification of latent dimensions and helps explain how observed variables are grouped into meaningful constructs [10, 12-17]. Machine learning and educational data mining approaches support prediction, classification, and decision-making based on student-related data [18-24]. However, several recent studies still provide limited discussion of model assumptions, feature justification, interpretability, and practical application in educational settings [19-24]. This creates a unique research gap of approach that combines the explanatory capability of factor analysis and the predictive capability of education data mining. A model that can identify the meaningful academic factors, validate the methods choices and representing the results which can be interpreted can have greater practical value as compared to those approaches that only focus on identifying the accurate predictors. The proposed research occupies this vacuum on the one hand, by focusing on structured identification of factors, transparent analysis, and performance-oriented interpretation; and, on the other hand, by concentrating on a more accurate educational implementation.

3. Novelty and Comparison

The novelty of this study lies in its integrated use of EFA/PCA, regression modelling, machine-learning comparison, permutation-importance analysis, and propensity-score-weighted analysis to examine academic performance across Matriculation, CBSE, and Government school systems. Unlike existing studies that mainly focus either on factor analysis or prediction, this work combines latent factor identification with predictive model validation in a single framework.

The proposed method reduces six subject-wise scores into two meaningful academic dimensions: STEM ability and Language Ability. These two latent factors explain 77.4% of the total Variance, showing that the method captures the major structure of student performance with fewer and more interpretable variables. This makes the approach stronger than methods that depend only on individual subject marks or overall scores.

The study further compares multiple regression and machine-learning models, including OLS, Ridge, Lasso, Decision trees, Random Forest, Gradient Boosting, and XGBoost, KNN, SVR, and Bagging Regressor. The results show that OLS Regression and Ridge Regression achieved the best testing performance, with R^2 values of 0.9742 and 0.9734, respectively, and MAE values close to 2.10. This proves that the proposed linear modelling approach is more suitable for the structured academic dataset than complex models that showed overfitting.

The proposed method is superior to prediction-only methods since the method offers interpretable accuracy. It delineates the key academic aspects, clarifies the role of subject variables as a contributor to latent academic performance via the process of permutation importance, and affirms that the notion of science-related subjects plays a huge role in influencing latent academic performance. Subject-level and latent ability patterns more fully explain differences in academic performance in the propensity-score-weighted analysis, even if section assignment has no discernible impact on latent performance. All in all, the proposed framework enhances the element of novelty by incorporating the elements of explanation, prediction, comparison, and interpretation into a single model. It offers a more reliable and practical approach for identifying academic strengths, weaknesses, and intervention needs across different education boards.

4. Factor analysis and Score Estimation

4.1. Model of Factor Analysis

A multivariate method, factor analysis, is used to examine whether a set of variables, denoted as $M_1, M_2, \ldots, M_i$ can be explained by a smaller number of underlying unobservable variables called factors, represented as $x_1, x_2 \ldots . x_a$, where the number of factors taken, like

Subjects, Year, Type of School, is less than the number of observed variables, meaning a < i.

Given observations on i variables, $M_1, M_2, \dots M_i$, with a mean vector μ and covariance matrix Σ, it is assumed that the factors $x_1, x_2 \dots . x_a$ are independent of each other and independent of the error terms ε_l for j= $1,2 , \dots , i$ [4]. The factors are also assumed to have a mean of 0 and a variance of 1. Each observed variable is modelled as a linear combination of these factors and an associated error term. This forms the basis of the factor analysis model for a given variable.

$$M_p = \mu + \lambda_{p1}x_1 + \lambda_{p2}x_2 + \cdots + \lambda_{pi}x_i + \varepsilon_l \quad (1)$$

Given that $\mu = 0$ when the observations are standardized, it can be rewritten as Equation (1) as:

$$M_p = \sum_{k=1}^{a} \lambda_{pk}x_k + \varepsilon_l \quad (2)$$

for $l = 1,2, \dots i\ ;\ k = 1,2, \dots a$

where λ_{jk} represents the factor loading on the j^{th} variable on k^{th} factor, and x_k are the fundamental causes. The unique or specific factors are denoted by ε_l .

Factor loadings λ_{pk} provide insight into the contribution of each variable to a given factor [25]. These loadings are analogous to multiple regression analysis weights and indicate the correlation strength between the factor and the corresponding variable [26].

Equation (1) can be represented in matrix form as:

$$M = \Lambda x + \varepsilon, \quad (3)$$

Where,

$x = x_1, x_2 \dots . x_a; \varepsilon = \varepsilon_1, \varepsilon_2, \dots \varepsilon_i$

And

$$\Lambda = \left[\begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1a} \\ \vdots & \ddots & \vdots \\ \lambda_{i1} & \cdots & \lambda_{ia} \end{pmatrix}\right].$$

The Variance of (1) is obtained as;

$$var(M_p) = {\lambda_{p1}}^2 var(x_1) + {\lambda_{p2}}^2 var(x_2) + \cdots + {\lambda_{pa}}^2 var(x_a) + var(\varepsilon_l). \quad (4)$$

Given that, the factor is assumed to have unit variance, Equation (3) can therefore be rewritten as:

$$var(M_p) = {\lambda_{p1}}^2 + {\lambda_{p2}}^2 + \cdots + {\lambda_{pk}}^2 + \Psi_l, \quad (5)$$

where considered as $\sum_{k=1}^{a} \lambda_{jk}^2$ communality of the a^{th} variable, and Ψ_l is considered as specific Variance of k^{th} variable Σ can be expressed as the covariance matrix of M [27].

$$\Sigma = \mathrm{E}(\mathrm{M} - \mu)(\mathrm{M} - \mu)' \quad (6)$$

Hence,

$$\mathrm{E}(\mu) = 0 \quad (7)$$

$$\Sigma = \mathrm{E}(\mathrm{M})(\mathrm{M})' \quad (8)$$

From (1), it can be written as;

$$\Sigma = \mathrm{E}(\mathrm{M})(\mathrm{M})' = E(\Lambda\, x + \varepsilon)(\Lambda\, x + \varepsilon)' \quad (9)$$

$$= \Lambda \mathrm{E}(xx')\Sigma\Lambda' + E(\varepsilon\varepsilon') \quad (10)$$

$$\Sigma = \Lambda\, \mathrm{var}(x)\Lambda' + var(\varepsilon) \quad (11)$$

Since the $var(\ x) = 1$, it follows that

$$\Sigma = \Lambda\Lambda' + \Psi, \quad (12)$$

Where,

$$\Psi = \left[\begin{pmatrix} \Psi_1 & 0 & 0 \\ \dots & \Psi_2 & \dots \\ 0 & 0 & \Psi_p \end{pmatrix}\right]. \quad (13)$$

4.2. Factor Score estimation

The factor score is the sum of each variable's scores on each factor [28]. The factor score can be calculated by taking one standard each variable's score and multiplying it by the corresponding factor, and the product is summed up. By using the factor model equation, the following expression is obtained:

$$\varepsilon = \mathrm{M} - \Lambda x \quad (14)$$

When (8) is squared on both sides and the sum is obtained,

$$\sum_{k=1}^{a} \varepsilon^2\, pk = \sum_{k=1}^{a} (M_p - \Lambda x_k)^2 \quad (15)$$

Reducing the aforementioned equation in relation to x_k and setting the derivation to zero gives

$$-2 \sum_{k=1}^{a} (M_p - \Lambda x_k)\Lambda = 0 \quad (16)$$

Then it becomes the following expression, when $-2 \neq 0$;

$$\sum_{k=1}^{a} (M_p - \Lambda x_k)\Lambda = 0 \quad (17)$$

This then gives

$$\Lambda a\widehat{M}_p - \Lambda\Lambda a\hat{x}_k \quad (18)$$

Dividing the above equation by $'a'$ and setting it in terms of $\hat{x}_k$, is obtained as,

$$\Lambda\Lambda\hat{x}_k = \Lambda\widehat{M}_p \quad (19)$$

The equation is multiplied by an inverse coefficient of $\hat{x}_k$ thus ; [29]

$$\hat{x}_k = (\Lambda\Lambda)^{-1}\Lambda\widehat{M}_p \quad (20)$$

Hence, Equation (9) is considered the factor score.

4.3. The Framework of the Multivariate Factor Analysis Model

The multivariate version of multiple linear regression is employed to model the relationship between the responses, $A_1, A_2, \dots A_a$ and a single group of predictor variables, $b_1, b_2, \dots b_r$ [27]. Each response was considered to follow its own individual regression model, ensuring that

$$A_i = \beta_{0i} + \beta_{1i}b_i + \cdots + \beta_{ri}b_r + \varepsilon_i, \ \forall\, i = 1,2,3, \dots, a \quad (21)$$

The error term ε has

$$\varepsilon_l = \begin{bmatrix} \varepsilon_{l1} \\ \varepsilon_{l2} \\ \dots \\ \varepsilon_{lm} \end{bmatrix} = 0; \quad \mathrm{Var}(\varepsilon) = \Sigma \quad (22)$$

Defining the response (dependent) variables as matrix A, the fixed but unknown parameters as matrix β, and the errors as matrix ε.

$$A_{p\times m} = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1a} \\ A_{21} & A_{22} & \dots & A_{2a} \\ \dots & \dots & \dots & \dots \\ A_{i1} & A_{i2} & \dots & A_{ia} \end{pmatrix} = \left(A_{(1)}\middle|A_{(2)}\middle|\dots|A_{(a)}\right) \quad (21)$$

$$\beta_{(r+1)m} = \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1a} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2a} \\ \dots & \dots & \dots & \dots \\ \beta_{i1} & \beta_{i2} & \dots & \beta_{ia} \end{pmatrix} = \left(\beta_{(1)}\middle|\beta_{(2)}\middle|\dots|\beta_{(a)}\right) \quad (22)$$

$$\mathcal{E}_{p\times m} = \begin{pmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} & \dots & \mathcal{E}_{1a} \\ \mathcal{E}_{21} & \mathcal{E}_{22} & \dots & \mathcal{E}_{2a} \\ \dots & \dots & \dots & \dots \\ \mathcal{E}_{i1} & \mathcal{E}_{i2} & \dots & \mathcal{E}_{ia} \end{pmatrix} = \left(\mathcal{E}_{(1)}\middle|\mathcal{E}_{(2)}\middle|\dots|\mathcal{E}_{(a)}\right) \quad (23)$$

for all $i = 1\ to\ n$. The design matrix is defined as $n * (r+1)$, a design matrix of explanatory variables or factors in Z.

Thus, the multivariate linear regression model is;

$$A_{p\times m} = \beta_{(r+1)m} + \mathcal{E}_{p\times m} \quad (24)$$

with $E(\mathcal{E}) = 0$ and

$$\mathrm{cov}(\mathcal{E}_i, \mathcal{E}_t) = (\sigma_{it}) \times I \quad (25)$$

for $i, t = 1,2, \dots, u$

The covariance matrix is;

$$\Sigma = (\sigma_{it}) \quad (26)$$

$$\hat{\beta} = (Z'Z)Z'\left(A_{(1)}\middle|A_{(2)}\middle|..|A_{(u)}\right) = (Z'Z)\hat{\beta}^{-1}B'A \quad (27)$$

The predicted values comprised the following matrix;

$$\hat{A} = Z\hat{A} = B(Z'Z)\ Z'A \quad \text{and}$$

$$\hat{\varepsilon} = Z - \hat{Z} = [I - Z(Z'Z)Z']Z \quad (28)$$

Residual $SSCP = \widehat{\varepsilon'}\hat{\varepsilon} = Z'Z - \hat{\beta}Z'Z\hat{\beta}$ and unbiased estimator can be written as;

$$\Sigma = \hat{\Sigma} = {\widehat{\varepsilon'}\hat{\varepsilon}}/{\mathrm{p} - \mathrm{r} - 1} \quad (29)$$

4.4. Principal Component Method

The covariance Σ for spectral decomposition is represented as eigenvector-eigenvalue pairs with (λ_1, e_i) as

$\lambda_1 > \lambda_2 > \cdots . \lambda_m > 0$ is given

$$\Sigma = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \lambda_3 e_3 e_3^T + \cdots . + \lambda_n e_c e_c^T.$$

The loading can be obtained;

$$L = \left(\sqrt{\lambda_1 e_1}, \sqrt{\lambda_2 e_2}, \dots, \sqrt{\lambda_c e_c}\right) \quad (30)$$

The covariance matrix can be observed as;

$$S_{11} + S_{22} + \cdots + S_{cc} = tr\,(S) \quad (31)$$

Trace of the sample correlation matrix

$$\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_c = \rho, \ i = 1,2,3, \dots, c \quad (32)$$

The proportion of total sample variance due to jth factor

$$= \begin{cases} \frac{\hat{\lambda}_j}{tr(S)}; & \text{sample covariance factor analysis} \\ \frac{\hat{\lambda}_j}{\rho}; & \text{correlation factor analysis} \end{cases} \quad (33)$$

$$\widehat{L^*} = \hat{L}T$$

Where,

$$TT' = T'T = I \quad (34)$$

$$\hat{L}\hat{L'} + \hat{\Psi} = \hat{L}TT'\hat{L'} + \hat{\Psi} = \widehat{L^*}\widehat{L'^*} + \hat{\Psi} \quad (35)$$

Based on each common factor, the factor score can be calculated for the individuals as useful by-product factor scores using the factor analysis method. The standardized measures are the mean value and standard deviation from the coefficient matrix of the factor score.

The principal component can be calculated as;

$$\widehat{f_{ik}} = \widehat{l_i}x_{1a} + \widehat{l_2}x_{2a} + \cdots . \widehat{l_c}x_{ca} \quad (36)$$

Where,

$$\widehat{f_{ik}} = \text{factor score } a^{th} \text{ student} \quad (37)$$

$$\widehat{l_j} = \text{principal component of variable } j \quad (38)$$

$$x_{ab} = \text{observation of } a^{th} \text{ student on } b^{th} \text{ work} \quad (39)$$

$$Y_a = \beta_{0a} + \beta_{1a}z_a + \cdots + \beta_{ra}z_r + \varepsilon_a \quad (40)$$

The error term calculated is

$$E(\varepsilon) = E\left(\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_p \end{bmatrix}\right) \quad (41)$$

The error terms are correlated with various responses. Let $z_{b0} + z_{b1} + \cdots + z_{br}$ be the predictor variables for b^{th} individual student and the errors and responses of the a^{th} The student is given below;

$$\varepsilon_b = \begin{bmatrix} \varepsilon_{b1} \\ \varepsilon_{b2} \\ \cdots \\ \varepsilon_{bp} \end{bmatrix} \text{ and } Y_b = \begin{bmatrix} Y_{b1} \\ Y_{b2} \\ \cdots \\ Y_{bp} \end{bmatrix} \quad (42)$$

5. Materials and Methods

5.1. Study Area and Population

The research was carried out based on the academic performance of 540 higher secondary students who were enrolled in three Indian school systems, namely Matriculation, CBSE, and Government schools. The sample size was 180 students of each type of school, and thus balanced representation of the three types of schools. The variables that were considered in the study were Gender, section, type of school and marks in subjects.

The sampling method was stratified, where the school type was taken as the stratification criterion. The sample was categorized into three strata: Matriculation, CBSE and Government schools. A total of 180 student records were chosen in each stratum to be analyzed. This method was used to have proportional representation of the three school systems and to make meaningful comparisons of the patterns of academic performance across the various educational boards.

5.2. Data Collection

The research utilized an anonymized secondary academic dataset acquired from student examination reports. Marks achieved in six major subjects were used to measure academic performance: English, Mathematics, Physics, Chemistry, Biology, and Language. The Language subject included regional or second-language subjects, such as Tamil or Hindi, depending on the student's curriculum.

Demographic and institutional Factors like Gender, section, and type of school were also included in the dataset. To preserve confidentiality, all the personally identifiable information, including the names of students, registration numbers, and other identifiable information, were eliminated prior to analysis. This data was to be analyzed within the frames of the academic research, and the analysis was to be conducted on the anonymized data. The subject-wise marks were used as the primary quantitative variables, which are subject to factor analysis, principal component analysis, regression modelling, and comparing by means of machine learning.

5.3. Data Pre-processing

Prior to the process of statistical analysis, the dataset was reviewed based on missing values, repetition of records, discrepancies in entries, and inappropriate data points. Records that had missing marks of a subject were checked. In the case where the valid information was not available, the records that contained a value that was missing or invalid were not included in the analysis to ensure data reliability.

Categorical variables such as Gender, section and type of school were coded into numerical form and subjected to statistical and machine-learning analysis. School type was coded to represent Matriculation, CBSE, and Government schools, while Gender and section were converted into appropriate categorical indicators.

Boxplots, distribution plots and checks of standardized scores were used in the analysis of outliers. Outliers were

checked to see whether the outliers indicated valid scores in academic performance, or they could be the effect of an error in data entry. Valid extreme scores were kept, and the inconsistent or invalid scores were eliminated.

All subject marks were standardized with the Z-score normalization to ensure consistency across subjects. Standardization was needed since multivariate techniques like EFA, PCA, regression, and machine-learn models can all be sensitive to changes in the variables' scale. The cleaned and standardized dataset was then used for further analysis.

5.4. Exploratory Data Analysis

Exploratory Data Analysis was conducted to examine the structure, distribution, and relationships within the dataset. Minimum, maximum, mean, median, and standard deviation are examples of descriptive statistics; skewness and kurtosis were calculated for each subject across the three school types.

Graphical methods were used to support the statistical interpretation of the data. Histograms and kernel density estimation plots were used to examine score distributions. Boxplots were used to identify dispersion and potential outliers. Q–Q plots were used to visually assess normality patterns. Violin plots were used to compare gender-wise score distributions. Correlation heatmaps were used to examine inter-subject relationships, while scree plots were used to support factor and component retention.

The EDA results provided an initial understanding of subject-wise performance patterns and supported main component analysis and factor analysis applications, regression modelling, and machine-learning model comparison.

5.5. Factor Analysis and Principal Component Analysis

Exploratory Factor Analysis and Principal Component Analysis were applied to identify latent academic performance dimensions among the six subject variables. The suitability of the data for factor-based analysis was assessed through the correlation structure among the subject variables. Factor and component retention were guided by the eigenvalue-greater-than-one criterion, scree plot interpretation, and explained Variance.

The extracted dimensions were interpreted using factor loadings, eigenvalues, communalities, uniqueness values, and percentage of Variance explained. Subjects with high loadings on the same factor were grouped into common academic dimensions.

The analysis identified two major latent dimensions: STEM-related ability and language-related ability. These dimensions summarized the major academic performance patterns and were later used for regression and machine-learning-based analysis.

5.6. Regression and Machine-Learning Modelling

The latent academic scores obtained from EFA/PCA were used for explanatory and predictive modelling. Regression and machine-learning models were applied to examine how subject-wise marks contributed to latent academic performance. The dataset was separated using an 80:20 split into training and testing subsets, where 80% of the records were used for model training, and 20% were used for testing. A fixed random seed was used to ensure reproducibility.

Coefficient of determination (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to assess the model's performance. The models evaluated included Ordinary Least Squares Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, Gradient Boosting, Decision Tree, Support Vector Regression, K-Nearest Neighbors, and Bagging Regressor. These models were selected to compare the performance of linear, regularized, tree-based, ensemble-based, and instance-based approaches.

Cross-validation was used to estimate model stability and reduce the possibility of overfitting. Permutation importance was used to examine feature importance and identify the subject variables that had the greatest influence on predicting latent academic performance.

5.7. Software Tools and Reproducibility

Python was used for all analyses. Data handling and numerical computation were performed using Pandas and NumPy. Preprocessing, train-test splitting, PCA, regression modelling, machine-learning algorithms, cross-validation, and performance evaluation were performed using scikit-learn.

Statistical modelling was conducted using statsmodels, while data visualization was performed using Matplotlib and Seaborn. Gradient-boosted tree modelling was conducted using XGBoost.

To ensure reproducibility, the analysis followed a fixed workflow consisting of data loading, column cleaning, missing-value checking, preprocessing, standardization, exploratory data analysis, EFA/PCA extraction, train-test splitting, model training, model evaluation, and visualization. When splitting trains into test and model training, a fixed random seed was utilized, and thus, the repeat experiment results remained constant. All the figures were prepared, and the image of high resolution was passed in the manuscript.

6. Results and Discussion

6.1. Descriptive statistics

Calculation of descriptive statistics was done to offer a summary of the performance of students attending the three school types: Government, Matriculation, and CBSE schools.

The measures of central tendency and dispersion (frequency, mean, median, standard deviation, minimum and maximum numbers) were analyzed. These statistics gave a general idea of the distribution of scores and the variability of the performance of the subjects and the categories of schools.

The preliminary basis of further multivariate analysis, including correlation analysis, Principal Component Analysis, Investigative Factor Analysis, and regression-based modelling, was also based on the descriptive results. Table 1 shows that most of the student records belonged to male students, accounting for 54.81% of the sample, while female students accounted for 45.19%.

Table 2 presents the skewness and kurtosis values for the subject-wise scores across Matriculation, CBSE, and Government schools. Skewness was used to assess the symmetry of the score distributions, while kurtosis was used to examine the degree of peakedness or flatness in the distributions.

In the present study, most Kurtosis and skewness values fell within a reasonable range, suggesting that the subject-wise score distributions did not show severe departures from normality. Therefore, the data were considered suitable for subsequent multivariate statistical analysis.

Table 1. Frequency of gender

Gender	Matric Schools		CBSE Schools		Government Schools		Total	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Male	105	58.33	91	50.56	100	55.56	296	54.81
Female	75	41.67	89	49.44	80	44.44	244	45.19
Total	180	100	180	100	180	100	540	100

Table 2. Descriptive Statistics of School Type

School Type	Subjects	N	Min	Max	Mean	Std. Deviation	Skewness	Kurtosis
Matric	English	180	2.33	98.67	45.72	29.15	0.18	-1.11
	Mathematics	180	2.67	98.00	51.31	27.65	0.04	-1.11
	Physics	180	10.33	96.67	56.43	21.61	-0.04	-1.00
	Chemistry	180	9.00	95.33	57.18	21.99	-0.25	-0.74
	Biology	180	11.33	96.67	55.18	22.42	-0.05	-0.85
	Language	180	7.33	93.67	44.25	24.30	0.05	-1.01
CBSE	English	180	2.00	99.67	50.88	30.53	-0.04	-1.27
	Mathematics	180	2.67	98.67	53.03	27.64	-0.08	-1.22
	Physics	180	9.67	96.00	53.85	21.34	-0.17	-0.57
	Chemistry	180	10.00	93.33	55.33	21.97	-0.21	-0.72
	Biology	180	10.67	96.67	56.39	22.23	-0.13	-0.83
	Language	180	6.33	99.00	53.16	28.28	-0.06	-1.32
Govt.	English	180	2.33	100.0	45.97	30.25	0.23	-1.22
	Mathematics	180	2.33	100.0	51.96	28.13	-0.03	-1.15
	Physics	180	11.33	95.33	56.95	20.47	-0.01	-0.46
	Chemistry	180	12.33	94.67	58.51	20.83	-0.33	-0.42
	Biology	180	7.00	96.00	56.74	21.64	-0.16	-0.58
	Language	180	6.33	99.00	49.42	27.92	0.20	-1.25

6.2. Correlation Matrix

Figure 1 shows the correlation matrix of the six subject-wise scores. English and Language showed a very strong positive correlation (r = 0.97), indicating a common language-related dimension. Physics, Chemistry, and Biology also showed strong positive correlations (r = 0.71–0.78), forming a clear science-related cluster. Mathematics showed comparatively weaker correlations with the other subjects (r = 0.16–0.33), suggesting that it represents a relatively distinct academic skill area. Overall, the correlation pattern supports the use of EFA and PCA for identifying broader academic performance dimensions.

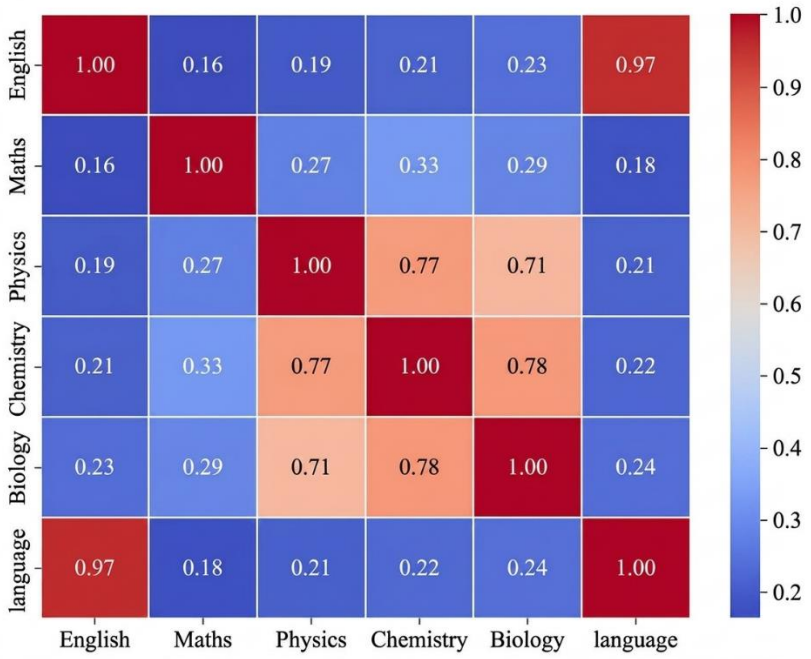

Fig. 1 Correlation matrix of subjects

Table 3. Distributional summary and KDE model locations for subject scores

Subjects	Distribution	KDE Peak	Observation
English	Slightly right-skewed, with more students scoring in the lower range	Around 0.1 to 0.2	A considerable number of students obtained lower scores, while fewer students scored in the higher range.
Mathematics	Fairly uniform with slight right skewness	Relatively flat, with a mild peak around 0.6 to 0.8	Student performance was widely distributed across score ranges, with a slightly higher concentration in the mid-to-high range.
Physics	Approximately bell-shaped and close to normal	Around 0.6	Scores were relatively balanced, with most students performing in the mid-to-high range.
Chemistry	Slightly left-skewed, indicating more students scoring in the higher range	Around 0.6 to 0.8	Students showed comparatively stronger performance in Chemistry than in English and Mathematics.
Biology	Slightly bimodal	Around 0.2 and above 0.6	The distribution suggests two performance groups, one with lower scores and another with higher scores.
Language	Slightly right-skewed, like English	Around 0.2	A larger number of students obtained lower scores, while relatively few students achieved very high scores.

Table 3 summarizes the distributional characteristics and KDE peak locations of the six subject-wise scores. English and Language show right-skewed distributions, indicating a higher concentration of lower scores, whereas Physics and Chemistry show more balanced or slightly higher-performing distributions. Biology shows a slightly bimodal pattern, suggesting the presence of two performance groups. Overall, the table indicates that subject-wise score distributions vary across disciplines, supporting the need for graphical and multivariate analysis. Figure 2 shows the histogram and kernel density estimation plots for the six subject-wise scores. The KDE curves help identify distributional patterns that may not be fully visible from the histograms alone. Chemistry and Physics showed comparatively higher average performance, with distributions that were approximately normal or slightly left-skewed. English and Language subjects showed lower average performance, with a larger concentration of students in the lower score ranges. Mathematics and Biology showed more varied distributions, with Mathematics displaying a relatively broad spread and Biology showing indications of bimodality. These distributional patterns provide a preliminary basis for further multivariate analysis by showing how subject-wise performance varies across the dataset.

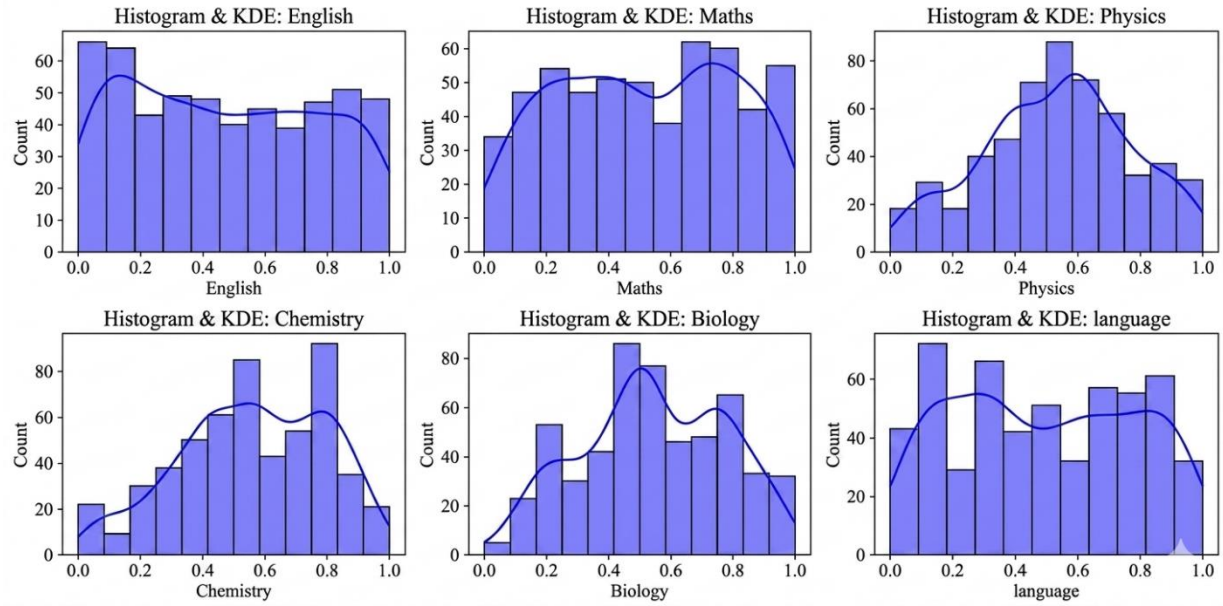

Fig. 2 Histogram with KDE of the subjects

Table 4 provides an overall comparative summary of subject-wise distribution quality, normality pattern, boxplot consistency, KDE pattern, and average correlation. Chemistry obtained the highest overall score of 0.75, followed by Physics and Biology with scores of 0.73, indicating stronger and more stable distributional patterns among the science subjects. Mathematics recorded the lowest overall score of 0.55, suggesting greater variability and weaker average correlation with the other subjects. English and Language showed moderate overall scores; mainly, these results support the later factor-analysis finding that science-related subjects and language-related subjects form separate academic dimensions. Figure 3 demonstrates the dispersion, distribution, and possible outliers of all six subjects. Physics, Chemistry and Biology have the highest median scores of between 0.55 and 0.60, and English and Language subjects have lower median scores of around 0.45, which indicates the relatively poor performance in them.

The median of the mathematics discipline is 0.50, which indicates an average level of achievement. No major outliers are identified, which demonstrates that the distribution of normalized scores is uniform, as the inter-quartile ranges of the subjects are similar. The range in the English and language disciplines is seen to be more varied at the lower end, with many students scoring low. This homogeneity is assumed by this consistency. There are no notable deviations in any of the subjects, and this testifies to the reliability of the dataset to be used in the comparative and multivariate analysis. exhibiting comparable Interquartile Ranges (IQR), reflecting considerable heterogeneity in student performance. The English and Language disciplines exhibit a greater dispersion in the lower half, indicating that numerous students attained low scores. This consistency supports the notion that data is uniform across all subjects. No subject exhibits significant change; thus, the dataset is reliable and suitable for comparative and multivariate analyses.

Table 4. Overall Score Table: Summarizing the Findings from All the Plots

Subject	Histogram Distribution	Q–Q Plot Normality	Boxplot Consistency	KDE Pattern	Average Correlation	Overall Score (Avg)
English	0.65	0.70	0.80	0.72	0.35 (avg of r: 0.16, 0.19, 0.21, 0.23, 0.97)	0.64
Mathematics	0.50	0.60	0.80	0.60	0.25 (avg of r: 0.16,0.27.0.33,0.29,0.18)	0.55
Physics	0.75	0.85	0.82	0.80	0.43 (avg of r: 0.19, 0.27, 0.77, 0.71, 0.21)	0.73
Chemistry	0.78	0.88	0.83	0.82	0.46 (avg of r: 0.21, 0.33, 0.77,0.78,0.22)	0.75
Biology	0.76	0.84	0.81	0.81	0.45 (avg of r: 0.23, 0.29, 0.71, 0.78, 0.24)	0.73
Language	0.70	0.72	0.79	0.70	0.36 (avg of r: 0.97, 0.18,0.21,0.22,0.24)	0.65

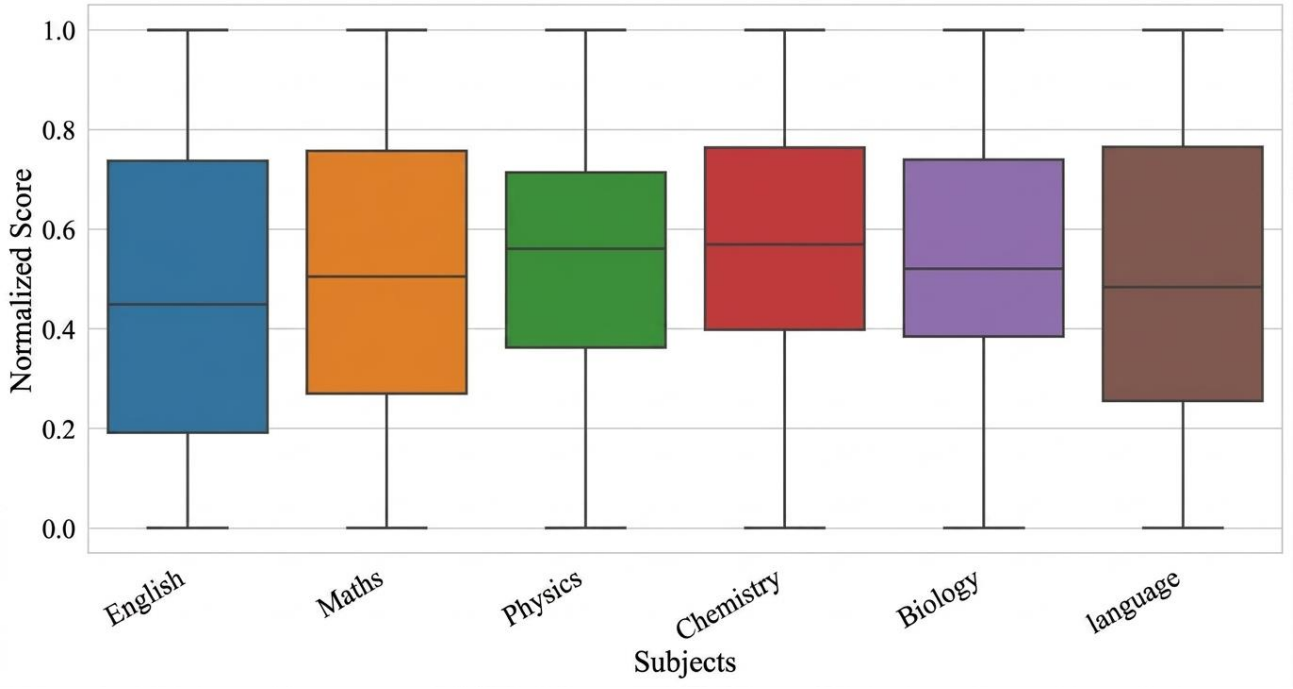

Fig. 3 Boxplot of normalized data

Figure 4 shows Q and Q graphs that evaluate the normality of the individual scores. Blue dots represent empirical quantiles, and the red line represents the theoretical normal distribution. In Physics, Chemistry and Biology, the points are close to the red line, and thus it is approximately normal, though with a higher frequency of lower-end scores than it would be under the assumption of an ideal normal distribution.

Mathematics portends the least variability, which can be indicative of an abnormality; even a small deviation of normality can be indicative of small skewness to the right or platykurtosis. The skewness of chemistry is mild left, indicating the presence of more high performers. These findings warrant using parametric methods like EFA and regression, and indicators of the need to carry out additional analysis on the subjects during the modelling.

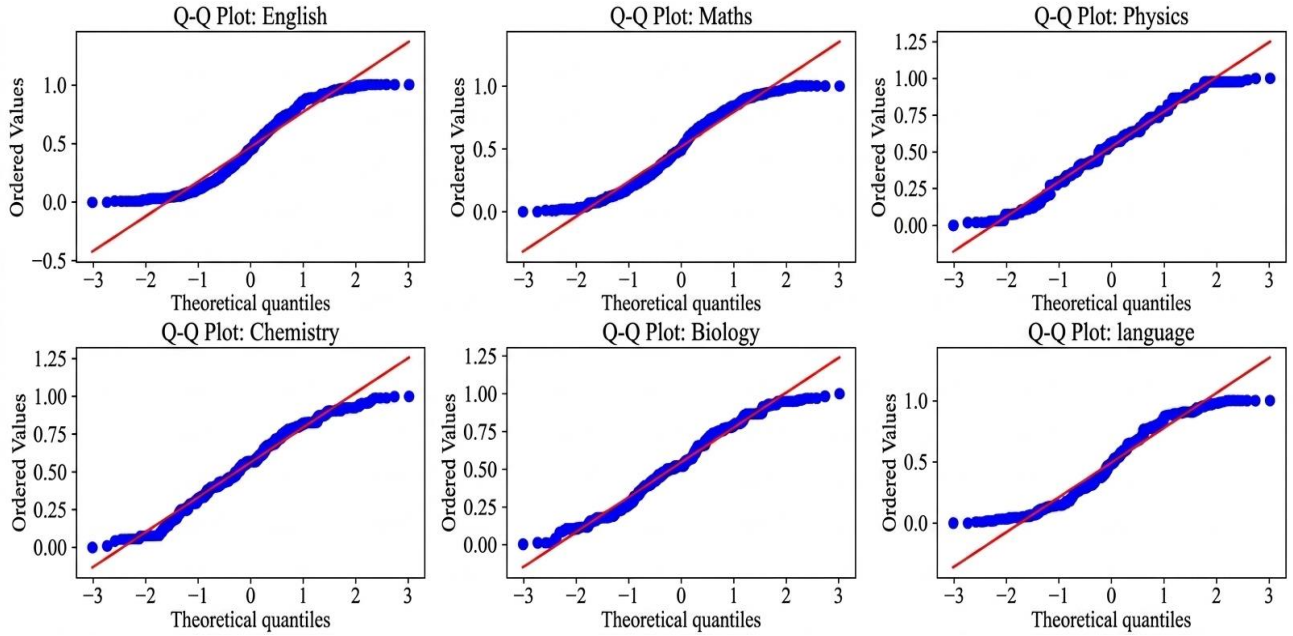

Fig. 4 Q-Q plot of subjects

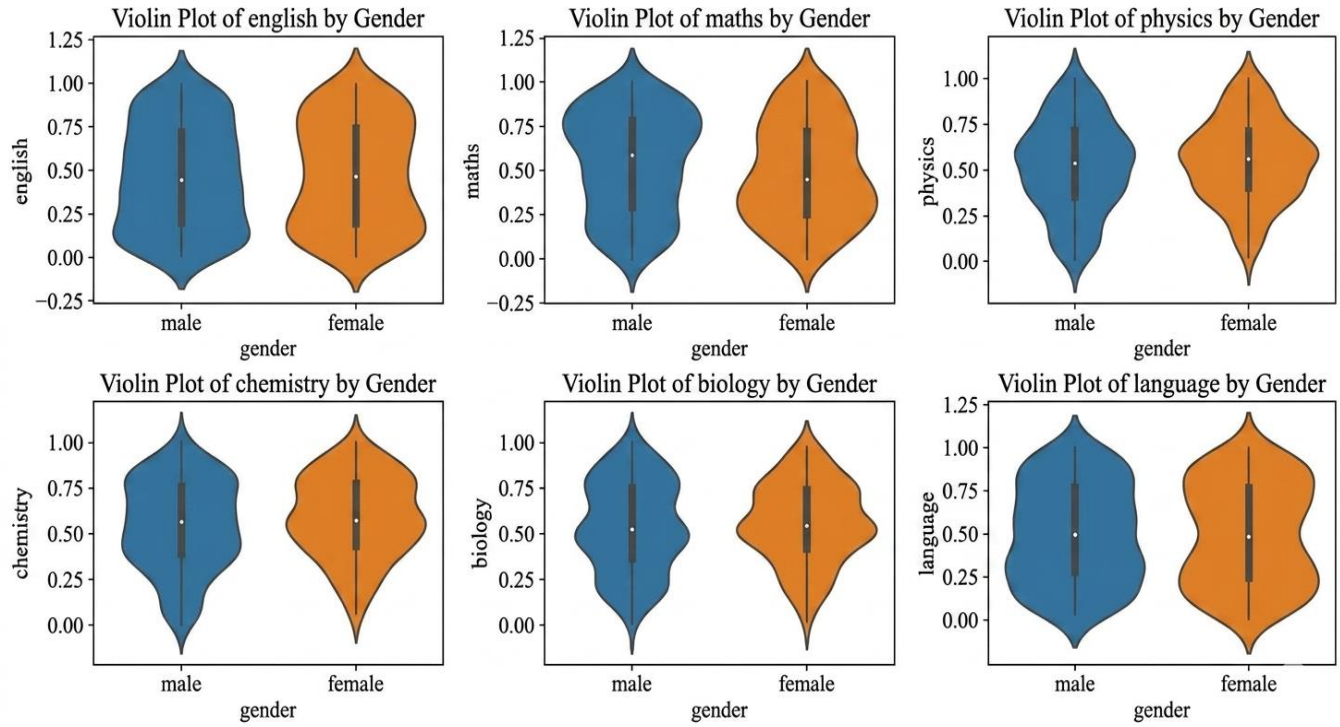

Fig. 5 Violin plot based on gender

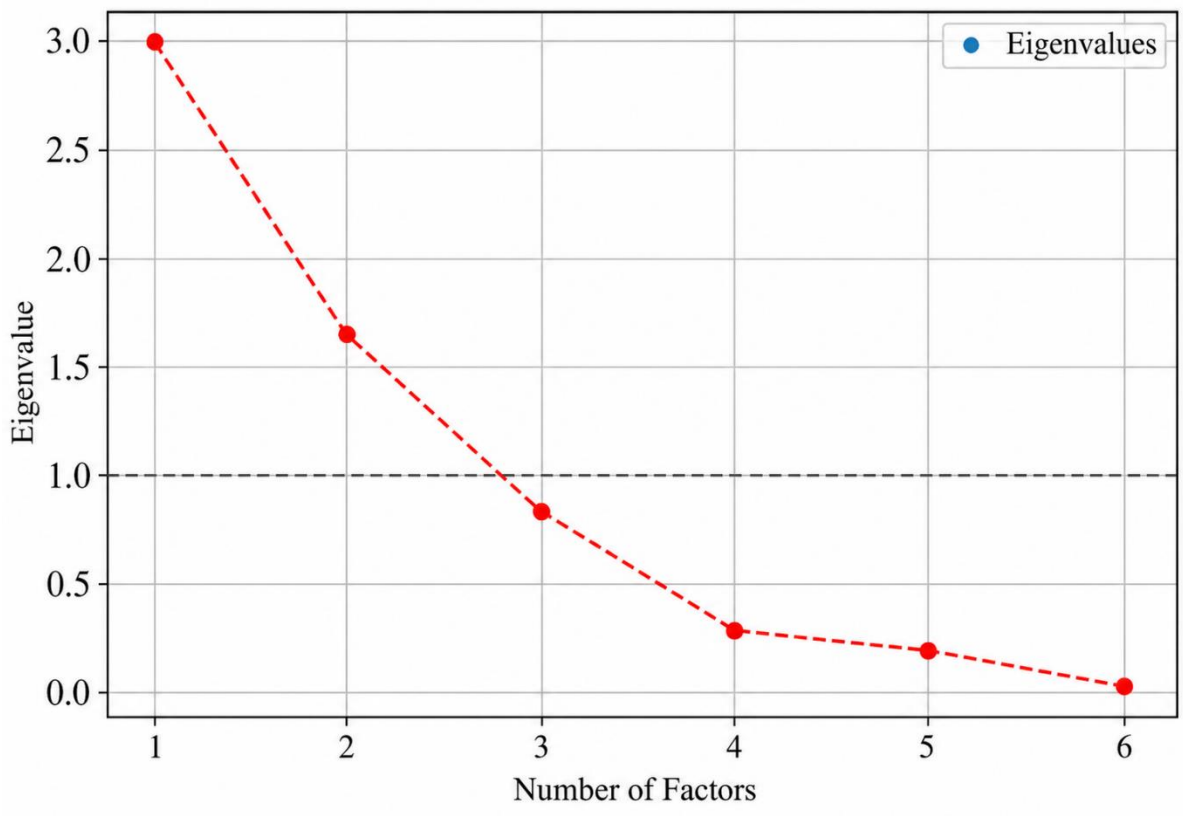

Fig. 6 Scree plot

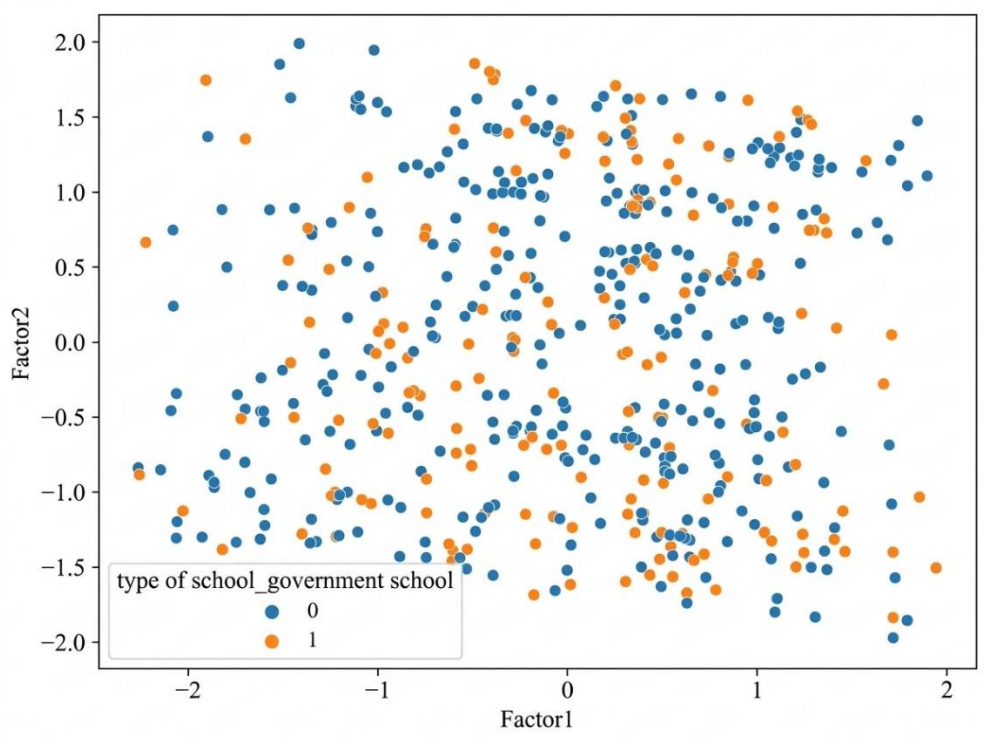

Fig. 7 Factor score by school type

Figure 5 shows how male and female students have distributed and have scored in each discipline, and female students are more likely to focus on the mid-to-high scoring range within the English and Language subjects. In mathematics and physics, male and female students exhibit comparable central tendencies; however, boys demonstrate much broader tails. In chemistry and biology, women possess a slight advantage due to their greater representation in senior positions. In general, girls perform better on assessments, particularly in subjects that necessitate extensive linguistic skills. In Figure 6, the scree plot, the eigenvalues of the elements retrieved are displayed in descending order. Elbow criteria or Kaiser rule (Eigenvalue greater than 1) gives you an idea about which factors to retain. The first two factors have eigenvalues greater than 1 (Factor 1≈: 3.0, Factor 2≈: 1.7), and these two factors capture most of the Variance in the dataset. The dataset comprising marks in six subjects can be reduced to two underlying latent factors.

Factor 1: General academic ability or science-related strength (high loading on Physics, Chemistry, Biology).

Factor 2: Language and communication skills (high loading on English, Language)

This is a Factor Analysis (FA) projection with points distinguished by school type (government versus others) in Figure 7. Factor 1 and Factor 2 are latent variables identified by Exploratory Factor Analysis (EFA). The two types of schools (Blue dots representing Government Schools and Orange dots representing Private Schools) exhibit overlapping clusters and distinct distribution patterns. This may indicate a weak yet present latent structure related to school type. Students from both government and non-government schools exhibit a range of intellectual talents, with no predominant group in either category.

Performance variability is greater inside each school type than between them. No quadrant is predominant for either category, indicating that no school style regularly achieves superior scores on both factors.

Both sorts of schools exhibit high and low performance. This graph illustrates the distribution of data across the first two principal components identified using PCA in Figure 8.

First component = STEM latent factor = latent PCA1

Second component = LANG latent factor = latent PCA2

The first component is defined as a common factor underlying math, physics, chemistry, biology, and languages. This captures the shared Variance across science subjects. The second component is defined as a common factor underlying English. This captures the shared Variance in language performance.

STEM = $\sum$ Total subjects except English lang = English

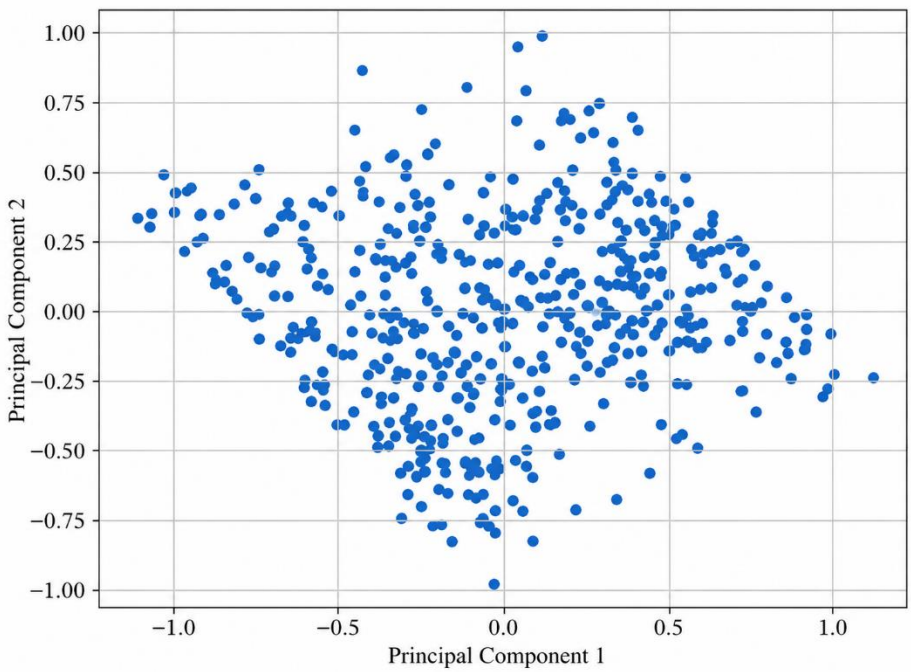

Fig. 8 PCA component scatter plot

In this analysis, student marks data were examined to identify latent academic factors that summarize overall performance across multiple subjects. Since the dataset has numeric variables of English, Mathematics, Physics, Chemistry, Biology and Language, it automatically treats the variables as primary subject variables. The initial was to conduct the Confirmatory Factor Analysis through the semopy package to divide performance into two interpretable latent factors, a STEM element of science-related subjects and a language element of language skill.

Nonetheless, as the semopy was not provided, the script selected the Principal Component Analysis. PCA minimized the six subject marks into fewer components that highlighted most of the Variance in performance. The first principal component is known to contribute to the Variance in the order of 49.9, and it is a general academic performance factor, and the second component is 27.5, which tends to represent opposing groups of subjects (e.g., stronger STEM compared to language performance). The latent scores obtained are saved to be later modelled, e.g. school or class level effects using mixed-effects models.

6.3. Integrated EFA/PCA-OLS-ML-SEM Framework.

Propensity-score-weighted analysis was conducted to examine whether class section was associated with overall academic achievement, represented by the latent principal component score (PC1). The propensity model was estimated using subject-wise scores in English, Mathematics, Physics, Chemistry, Biology, and Language. Before weighting, the comparison groups showed an imbalance in selected covariates. After weighting, the standardized mean differences were reduced close to zero, indicating improved covariate balance between the comparison groups. This reduced the influence of baseline covariate differences on the estimated association between section assignment and latent academic performance. The latent performance score was then used as the dependent variable and section as the explanatory variable in a weighted regression model. The estimated average treatment effect was 0.0053, with a confidence interval ranging from −0.299 to 0.310. The p-value of 0.973 indicated that the association was not statistically significant. Overall, the section assignment was not significantly associated with latent academic performance in this dataset.

Measurement:

$$Physics_{ij} = \lambda_{p_1\eta_1,ij} + \varepsilon_{p,ij}$$

$$English_{ij} = \lambda_{E_2\eta_2,ij} + \varepsilon_{E,ij}$$

Structural two-level:

$$\eta_{k,ij} = \beta_{0k}^{j} + x_{ij}\beta_k + u_{kj}$$

$$\beta_{0k}^{(school)_j} = \gamma_{00k} + z_j\gamma_{0k} + v_{okj}$$

Where i, student and j school, $k\epsilon\{1,2\}$

Factor loadings, model fit indices, ICCs for each latent variable, between-school Variance explained by school type, and credible/confidence intervals are shown in Figure 9.

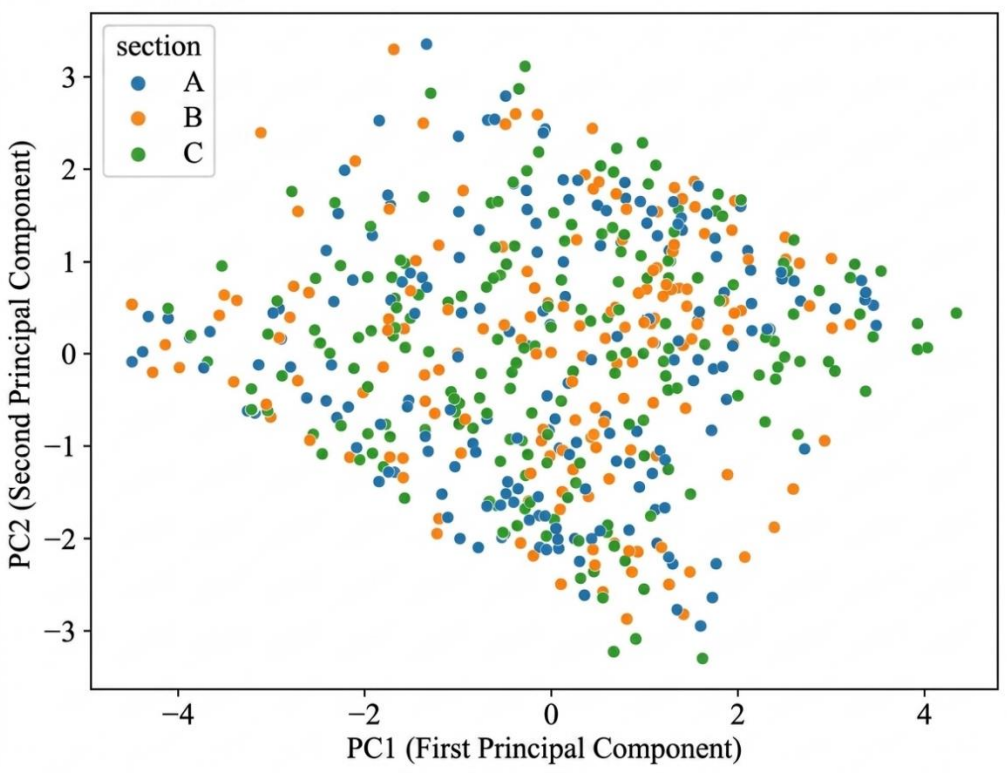

Fig. 9 Principal component of factor-based analysis

6.4. Propensity-Score-Weighted Analysis

A propensity-score-weighted analysis was conducted to examine whether section assignment was associated with students' latent academic performance scores. A Directed Acyclic Graph (DAG) was used conceptually to identify possible covariates that could influence both group assignment and academic performance. Based on this structure, inverse-probability weights were estimated using measured covariates, including gender, school type, section, and baseline academic performance indicators. Standardized Mean Differences (SMDs) were used to evaluate covariate balance following weighing.

The post-weighting balance diagnostics indicated that the measured covariates were well balanced across comparison groups, with SMDs reduced close to zero. This implies that the weighted groups were more similar as compared to the original unweighted groups. The latent academic performance score was then used as the outcome variable, and a weighted regression model was applied. The average treatment effect was estimated to be 0.005 with the 95% confidence interval of negative 0.299 to positive 0.310. This finding was not statistically significant (p=0.973; N=540) and therefore section assignment was not found to meaningfully associate with latent academic performance in this dataset.

Checks of sensitivity were also done by checking covariate balance and the stability of the estimates under different propensity-score specifications. The results remained consistent, suggesting that the finding was not strongly affected by extreme propensity scores or model specification changes. However, because the study is based on cross-sectional observational data, the findings should be interpreted as adjusted associations rather than definitive causal effects.

6.5. Explainable Model Comparison and Feature Importance Analysis

Regression and machine-learning models were evaluated to predict students' latent academic performance scores using subject-wise marks as predictor variables. The predictor variables included English, Mathematics, Physics, Chemistry, Biology, and Language.

The latent academic score obtained from EFA/PCA was used as the target variable for model training and evaluation. The model comparison included linear models, regularized regression models, tree-based models, ensemble models, and instance-based models. R^2, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were used to assess the model's performance. To evaluate the stability of the model and lower the possibility of overfitting, cross-validation was employed. For model explainability, permutation importance was used to identify the relative contribution of each subject to the prediction of latent academic performance. The results shown in Figure 10 indicate that science-related subjects contributed strongly to the prediction of overall academic performance.

This finding supports the factor-analysis results, where Physics, Chemistry, Biology, and Mathematics showed strong associations with the STEM-related latent dimension. The comparison of regression and machine-learning models showed that OLS Regression and Ridge Regression achieved the strongest predictive performance. These models recorded smaller error values and high R^2 scores as compared to the more complicated non-linear models. This implies that the academic score data is highly linear, and therefore, linear regression models were better suited to this dataset. While ensemble models such as Gradient Boosting, Random Forest, and XGBoost also performed very well, they were more susceptible to overfitting than OLS and Ridge Regression. The R^2 score comparison is presented in Figure 11, while the detailed model performance metrics are summarized in Tables 5 and 7. A better fit is indicated by higher R^2 scores and lower error values. OLS and Ridge Regression achieved the strongest testing performance, indicating that these models were well-suited to the linear structure of the dataset. In contrast, KNN and SVR produced lower testing R^2 scores, suggesting weaker suitability for this data structure. Decision Tree, XGBoost, and Random Forest showed high training performance, but their lower testing performance indicates a greater tendency toward overfitting. As shown in Table 7, OLS Regression achieved a testing R^2 value of 0.9742, while Random Forest achieved a training R^2 value of 0.9934. Table 5 summarizes the overall predicted effectiveness of the evaluated regression and machine-learning models. Ridge Regression and OLS Regression achieved the strongest performance, with the lowest MSE values of 6.507835 and 6.508920 and the highest R^2 values of 0.981610 and 0.981607, respectively. Their MAE values were also the lowest, close to 2.10, indicating smaller prediction errors. In contrast, KNN and SVR produced the weakest results, with lower R^2 values and higher error values. These findings show that linear regression-based models are more suitable for the structured academic performance dataset than the weaker instance-based models. Figure 12 shows the Mean Squared Error (MSE) comparison of machine learning models. XGBoost, Gradient Boosting, Ridge, and OLS are found to be the best performing models by having the lowest testing and training MSE values. SVR, KNN, and Decision Tree are highly overfitted due to a large gap between training and testing MSE. The remaining Random Forest, Bagging, and Lasso are moderate overfitting since there exists a low training error with high testing error. Figure 13 shows the Mean Absolute Error (MAE) comparison of machine learning models. Least mean absolute error was shown by the Multivariate OLS regression model and the Ridge regression. KNN has the highest testing MAE with ~5.72, which indicates low accuracy and high errors. Ridge and Lasso Regression provided with 2.10 and 2.57 MAE, respectively. XGBoost with moderate MAE during testing indicates strong learning with fewer errors. Based on the results, Decision Tree and XGBoost perform exceptionally well in training and not in testing, while models like SVR and Gradient Boosting have relatively lower performance.

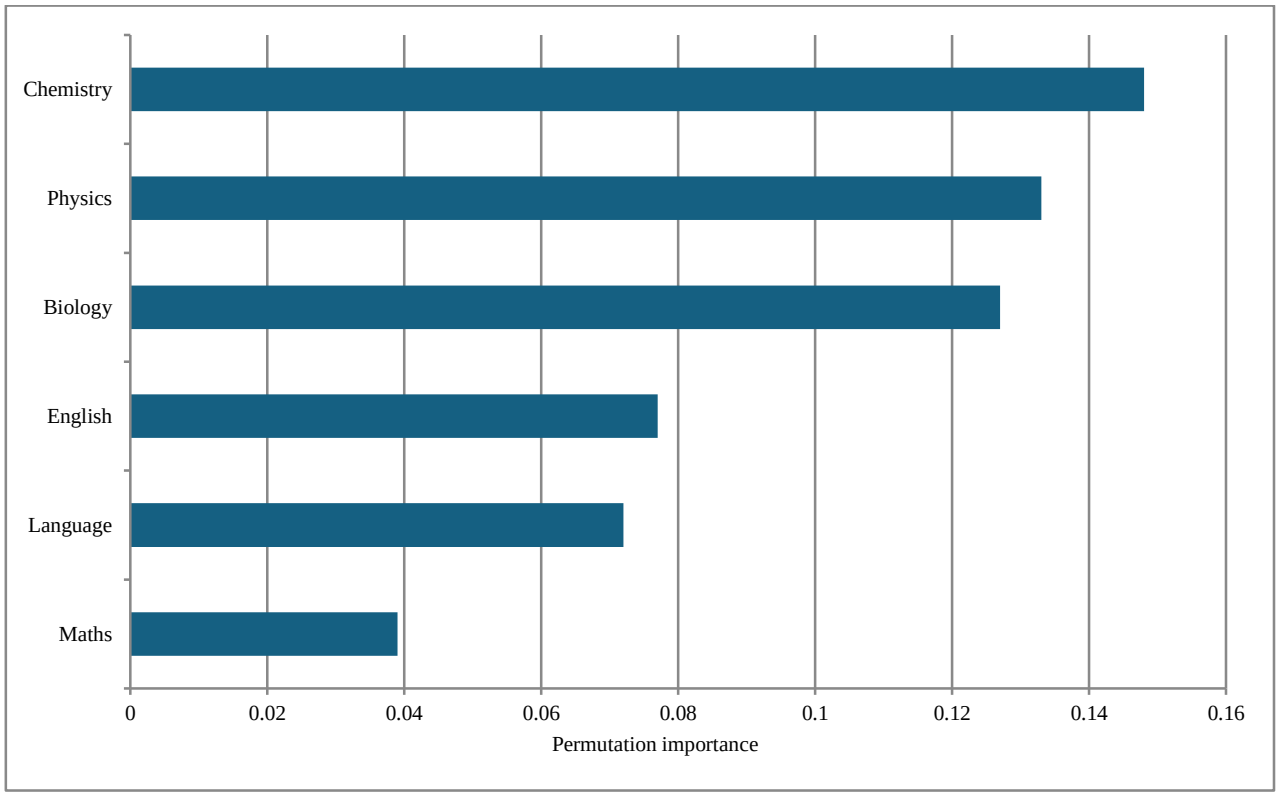

Fig. 10 Permutation importance based subject candidates with respect to subjects

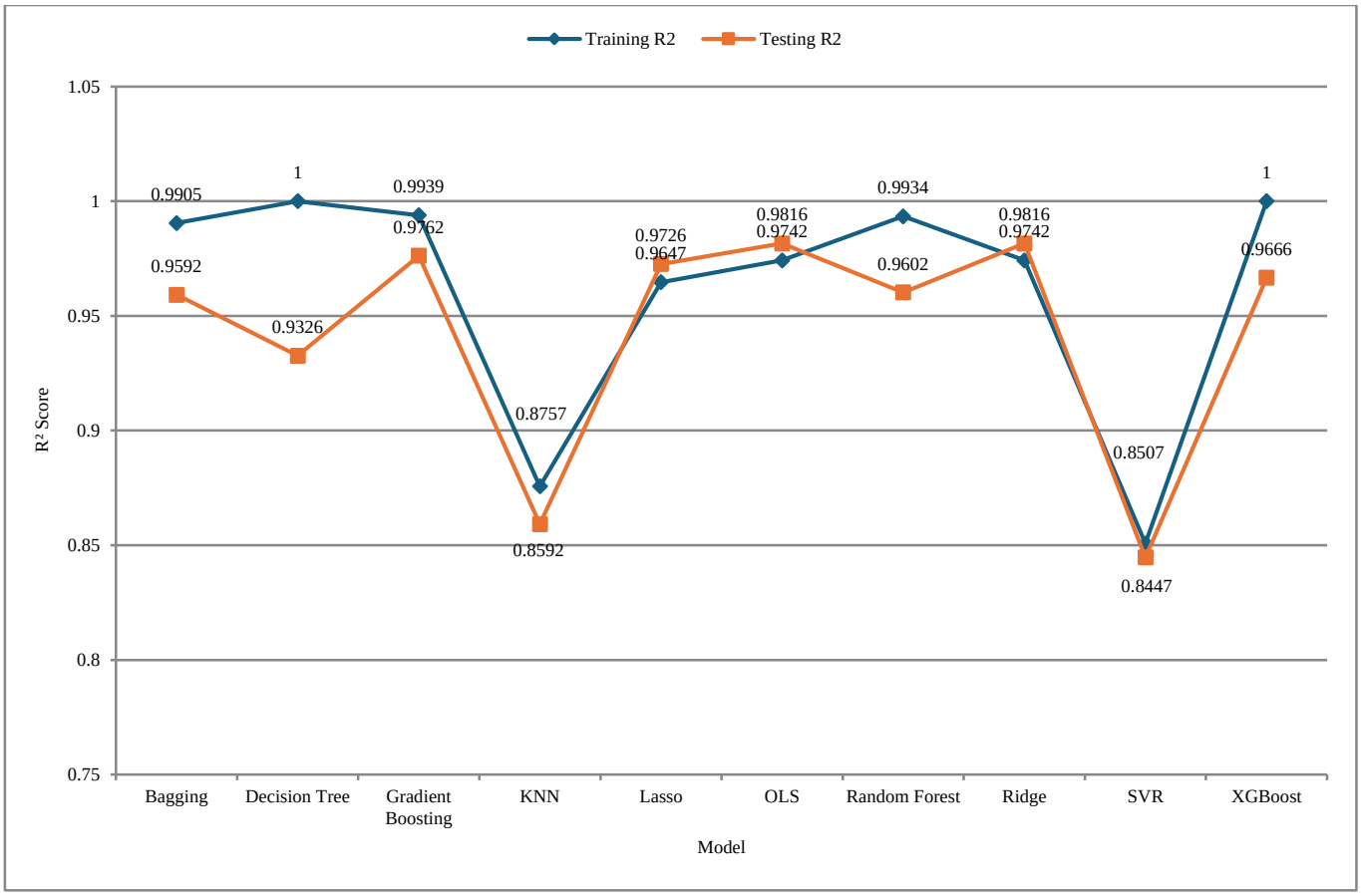

Fig. 11 R² score comparison

Table 5. Overall performance summary

No.	Model	MSE	R²	MAE
1	OLS	6.508920	0.981607	2.099077
2	Ridge	6.507835	0.981610	2.099710
3	Lasso	9.696176	0.972600	2.566450
4	XGBoost	11.834852	0.966557	2.738392
5	Gradient Boosting	8.334896	0.976447	2.393870
6	Random Forest	14.140834	0.960041	3.104830
7	Decision Tree	24.891975	0.929660	3.827160
8	KNN	49.830206	0.859189	5.724074
10	SVR	54.959150	0.844696	5.147931
11	Bagging	15.656134	0.955759	3.187500

OLS and Ridge Regression Model show very low MAE in training and testing with 2.0991 and 2.0997, respectively, in Table 5. OLS and Ridge regression outperform the other methods in the educational domain as they assume a linear relationship with structured data for examining academic performance. The consistent patterns in the dataset allow the OLS to provide accurate and reliable results. The effectiveness of OLS is enhanced by preprocessing to minimize bias. OLS avoids overfitting and performs well for this dataset.

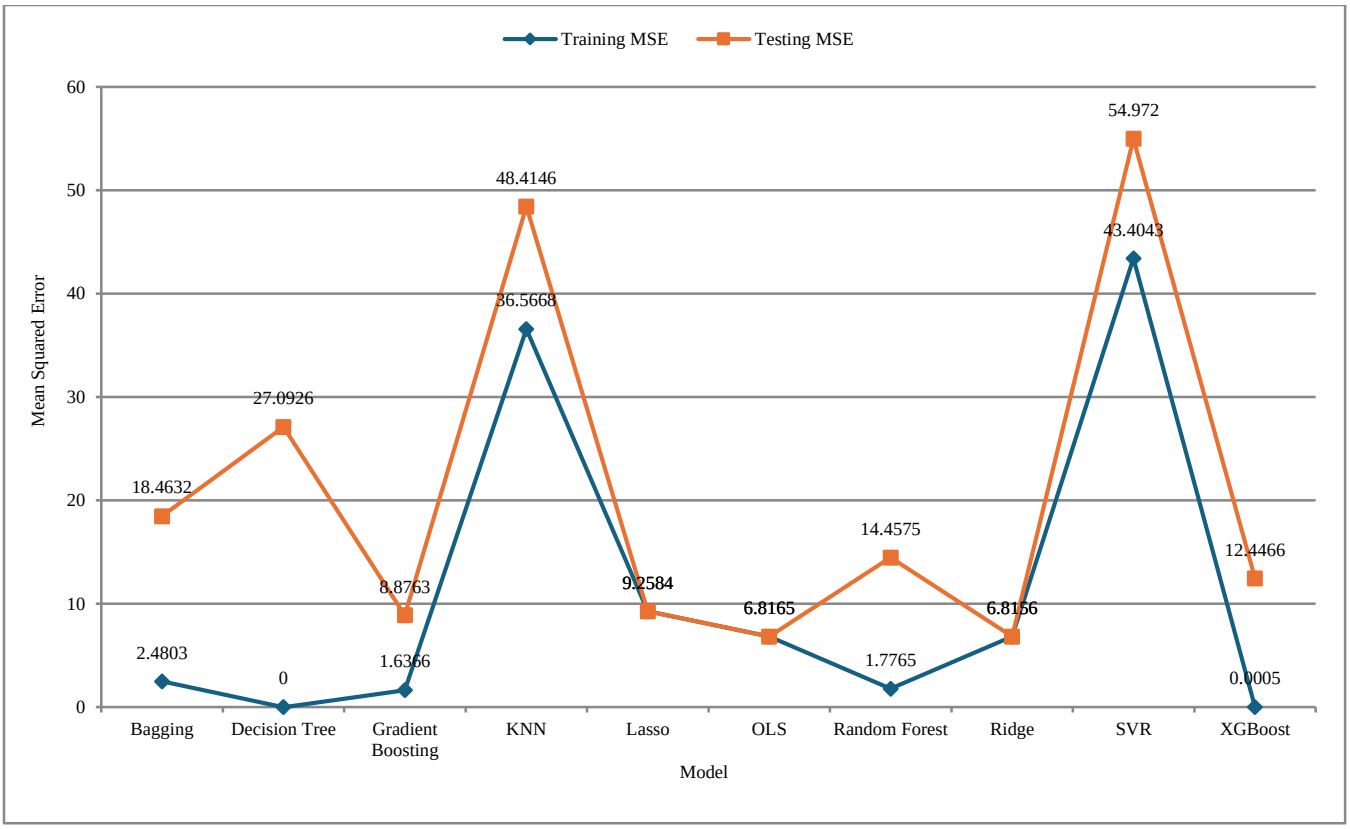

Fig. 12 Mean squared error comparison of regression models

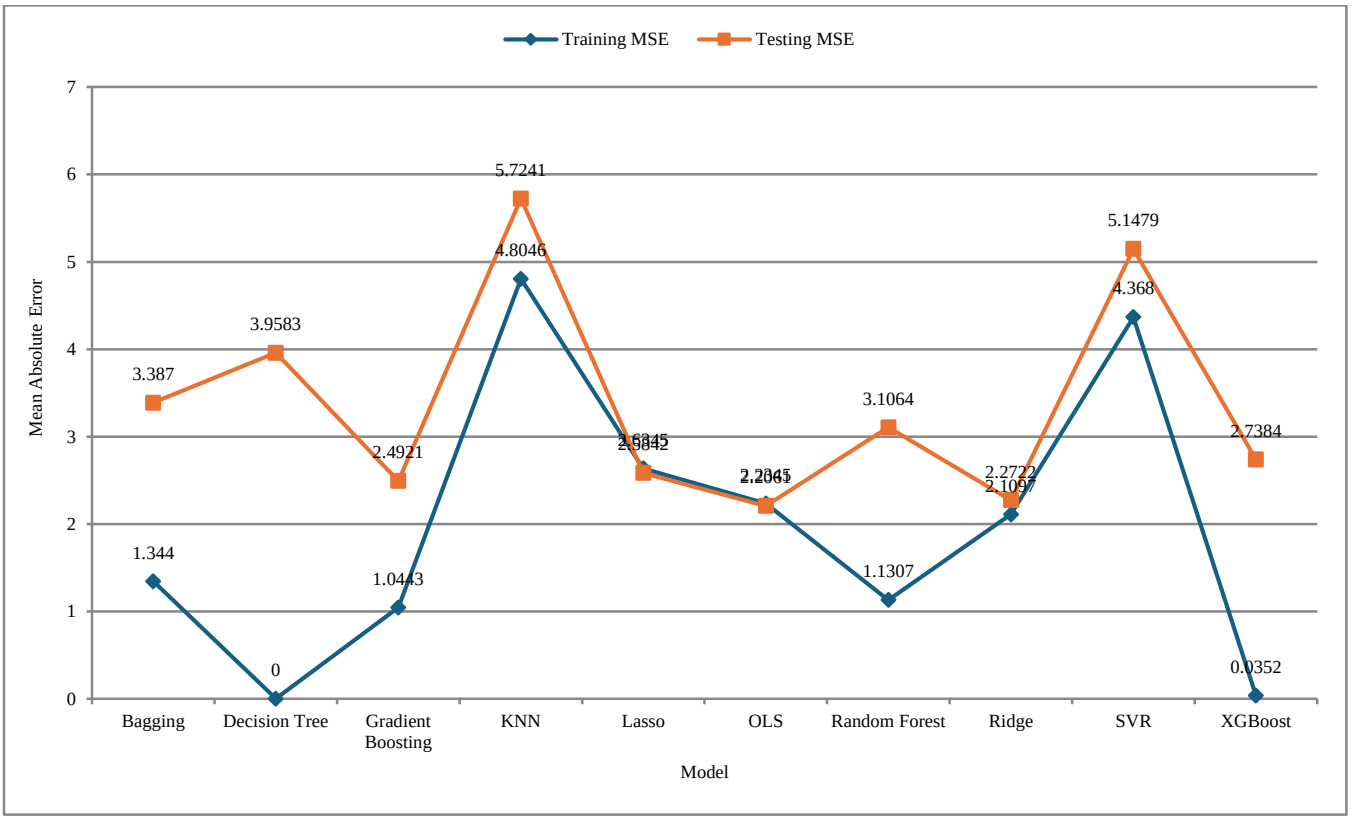

Fig. 13 Mean absolute error comparison of regression models

Table 6. Summary Of Factor Loading and Explained Variance (EFA/PCA)

Subject	Factor 1 (STEM Ability)	Factor 2 (Language Ability)	Communality	Uniqueness
Physics	0.86	0.18	0.78	0.22
Chemistry	0.83	0.22	0.74	0.26
Biology	0.80	0.25	0.72	0.28
Mathematics	0.75	0.30	0.68	0.32
English	0.27	0.84	0.79	0.21
Language	0.31	0.88	0.82	0.18
Eigenvalue	3.02	1.68	–	–
% Variance Explained	49.9%	27.5%	77.4% Total	–

Table 6 presents the results of the EFA and PCA, which extracted two dominant latent factors representing students' academic performance structure. Factor 1, named STEM Ability, had high loadings on Physics (0.86), Chemistry (0.83), Biology (0.80), and Mathematics (0.75), representing a general sense of scientific and analytical ability. Factor 2, marked as Language ability, had strong loadings on English (0.84) and Language (0.88), with the reflection of linguistic proficiency. A combination of these two factors explained 77.4% of the total Variance, which means that most of the changes in student achievement can be attributed to this two-dimensional interplay. The high communalities across all subjects suggest that the factor model effectively represents the observed data structure.

Table 7. Training and testing performance comparison

Model	Training R^2	Testing R^2	MAE	MSE	**RMSE**
OLS Regression	0.9951	0.9742	2.0991	4.407	2.0997
Ridge Regression	0.9945	0.9734	2.100	4.410	2.101
Lasso Regression	0.9824	0.9422	2.570	6.611	2.571
Random Forest	0.9934	0.9510	2.319	5.300	2.303
XGBoost	0.9962	0.9485	2.380	5.664	2.380
Gradient Boosting	0.9810	0.9352	2.615	6.890	2.624
Decision Tree	0.9980	0.9241	3.100	8.160	2.857
SVR	0.8945	0.8324	5.180	11.244	3.352
KNN	0.8680	0.7991	5.720	12.710	3.566

Table 7 summarizes the comparative performance of different machine learning and regression models used to forecast students' overall academic success. Among the evaluated models, Ordinary Least Squares (OLS) Regression and Ridge Regression achieved the best predictive accuracy, with testing R^2 values of 0.9742 and 0.9734, respectively, and mean absolute error values close to 2.10.

These results indicate a strong linear relationship between subject-wise scores and overall academic performance. The strong reliability and stability of the models of linear analysis shown in Table 7 support the fact that these models are applicable to structured academic data interpretation. Models such as Random Forest and XGBoost demonstrated slightly higher training accuracy but showed moderate overfitting, whereas instance-based models like KNN and SVR yielded weaker generalization performance.

Overall, Table 7 highlights the robustness and consistency of linear models in educational data analysis, validating their suitability for interpreting structured academic datasets.

7. Conclusion

This paper offers a model of analysis of academic performance that can be integrated and interpreted to analyze academic performance in Matriculation, CBSE, and Government school systems. The new novelty of the work is that it combines Exploratory Factor Analysis, Principal Component Analysis, regression modelling, machine-learn analysis, permutation-importance analysis, and propensity-score-weighted analysis within one framework. As opposed to the prediction-only studies, the proposed approach will identify the hidden academic dimensions and prove them with the help of predictive modelling, so that the results can be explained and used in practice. The results indicate that there are 6 subject-wise scores that can be simplified into 2 meaningful latent dimensions: STEM ability and Language ability. These two factors explained 77.4% of the total Variance which had shown that the proposed factor-based approach was effective in capturing the main structure of student academic performance using fewer and more interpretable variables. The comparison of the models also helps to prove the power of the offered method. Ordinary Least Squares Regression and Ridge Regression have the

highest testing performance in terms of high R 2 and low error values. This demonstrates that relationships between subject-wise marks and latent academic performance are, to a great extent, linear. Even though complex models like Random Forest, XGBoost, and Gradient Boosting demonstrated similar competition training results, they were more likely to overfit. The permutation-importance analysis revealed a high level of contribution of science related topics on prediction of latent academic performance.

The propensity-score-weighted model also indicated that section assignment had no discernible effect on latent performance. As a whole, the proposed framework enhances the study of academic performance because it consists of explanation, prediction, comparison, and interpretation, which support the analysis of academic performance in a data-driven manner, at an early stage, and through targeted application of academic interventions.

7.1. Limitations and Future Work

The current research work does have limitations that provide guidelines to future research work. The first one is that the analysis was performed with the assistance of cross-sectional scholarly data; it implies that the results suggest the correlations between the scores of the subjects, between the latent factors, as well as between the student performance, but they do not indicate strong causal relations. Longitudinal academic records should be used in future studies to investigate the nature of STEM ability and Language ability development over time.

Second, the dataset did not include valuable socioeconomic, behavioral, and institutional variables, such as parental education, family income, attendance, teacher quality, school infrastructure, classroom environment, learning resources, and student motivation. Including these factors in future work may improve model explanation and provide a broader understanding of academic achievement.

Third, the school systems that had been considered to conduct the study did not exceed the Matriculation, CBSE and Government school systems. Therefore, more research is necessary using other boards, regions, types of schools and other large schools of students to enhance the generalization, and one can make good comparisons across the school educational contexts.

Fourth, the study has been based on secondary scholarly reports; there may have been an institutional reporting and a point of disagreement. Future studies can make use of more standard forms of well-documented data sets.

Lastly, though the proposed linear models demonstrated great performance and interpretability, Future studies can look into the application of hybrid models that combine the transparency of the linear models using with the predictive power of advanced machine-learning methods. These models might be used in identifying the linear and non-linear performance patterns. Results can also be used to develop decision-support systems in schools, teachers, and policymakers to diagnose learning gaps early to design specific academic interventions.

Data Availability

The dataset supporting the conclusion of this article will be made available by the authors, without undue reservation.

Conflict of Interest

The authors disclose that they are not having any conflict of interest with publishing of this research.

Acknowledgement

The authors express their sincere gratitude to Hindustan Institute of Technology and Science (HITS), Chennai, for providing the necessary support and facilities to carry out this research.

References

[1] M. Settino et al., “Machine Learning for Predicting Post-Partum Anemia: A Comparative Performance Analysis,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mohsen Tavakol, and Angela Wetzel, “Factor Analysis: A Means for Theory and Instrument Development in Support of Construct Validity,” *International Journal of Medical Education*, vol. 11, pp. 245-247, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Juana Sanchez, “Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS,” *Journal of Statistical Software*, vol. 16, no. 4, pp. 1-3, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[4] Okeke Evelyn Nkiruka et al., “Application of Factor Analysis on Academic Performance of Pupils,” *American Journal of Applied Mathematics and Statistics*, vol. 5, no. 5, pp. 164-168, 2017. [Google Scholar] [Publisher Link]

[5] Matt R. Raven, “The Application of Exploratory Factor Analysis in Agricultural Education Research,” *Journal of Agricultural Education*, vol. 35, no. 4, pp. 9-14, 1994. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rudolph J. Rummel, *Applied Factor Analysis*, Northwestern University Press, 1970. [Google Scholar] [Publisher Link]

[7] Dennis Child, *The Essentials of Factor Analysis*, A and C Black, 2006. [Google Scholar]

[8] ADSUSISMre Field, *Discovering Statistics using SPSS: Introducing Statistical Method*, 2009. [Google Scholar]

[9] Richard L. Gorsuch, *Factor Analysis*, 2nd ed., Routledge, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[10] Barbara G. Tabachnick, and Linda S. Fidell, *Using Multivariate Statistics*, 5th ed., Allyn and Bacon/Pearson Education, 2007.[Google Scholar]
[11] Andrew L. Comrey, and Howard B. Lee, *A First Course in Factor Analysis*, 2nd ed., Psychology Press, 1992. [CrossRef] [Google Scholar] [Publisher Link]
[12] Olcay Kursun, "*SINBAD Automation of Scientific Process: From Hidden Factor Analysis to Theory Synthesis*," Electronic Thesis and Dissertations, University of Central Florida, United States, pp. 1-86, 2004. [Google Scholar] [Publisher Link]
[13] Seyed Ali Marjaie, and Vasundhara Kulkarni, "Recognition of Hidden Factors in Requirements Prioritization using Factor Analysis," *2010 International Conference on Computational Intelligence and Software Engineering*, Wuhan, China, pp. 1-5, 2010. [CrossRef] [Google Scholar] [Publisher Link]
[14] Delimiro Visbal-Cadavid, Mónica Martínez-Gómez and Rolando Escorcia-Caballero, "Exploring University Performance through Multiple Factor Analysis: A Case Study," *Sustainability*, vol. 12, no. 3, pp. 1-24, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[15] Keenan A. Pituch, and James P. Stevens, *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS*, 6th ed., Routledge, 2016. [Google Scholar] [Publisher Link]
[16] Julie Pallant, *SPSS Survival Manual: A Step-by-Step Guide to Data Analysis using IBM SPSS*, 7th ed., Routledge, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[17] John L. Horn, "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, vol. 30, no. 2, pp. 179-185, 1965. [CrossRef] [Google Scholar] [Publisher Link]
[18] Mustafa Yağcı, "Educational Data Mining: Prediction of Students' Academic Performance using Machine Learning Algorithms," *Smart Learning Environments*, vol. 9, no. 1, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[19] Yawen Chen, and Linbo Zhai, "A Comparative Study on Student Performance Prediction using Machine Learning," *Education and Information Technologies*, vol. 28, no. 9, pp. 12039-12057, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[20] Zara Ersozlu, Sona Taheri, and Inge Koch, "A Review of Machine Learning Methods used for Educational Data," *Education and Information Technologies*, vol. 29, no. 16, pp. 22125-22145, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[21] Rodrigo Guevara-Reyes et al., "Machine Learning Models for Academic Performance Prediction: Interpretability and Application in Educational Decision-Making," *Frontiers in Education*, vol. 10, pp. 1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[22] Mohamed Bellaj et al., "Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance," *International Journal of Online and Biomedical Engineering*, vol. 20, no. 3, pp. 55-74, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[23] Athanasios Angeioplastis et al., "Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach using Educational Data Mining Techniques," *Computers*, vol. 14, no. 3, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[24] Sazol Sarker et al., "Analyzing Students' Academic Performance using Educational Data Mining," *Computers and Education: Artificial Intelligence*, vol. 7, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[25] Harry H. Harman, *Modern Factor Analysis*, 2nd ed., University of Chicago Press, 1976. [Google Scholar]
[26] Paul Kline, *An Easy Guide to Factor Analysis*, 1st ed., Routledge, 1994. [CrossRef] [Google Scholar] [Publisher Link]
[27] Mohamed El Jihaoui, Oum El Kheir Abra, and Khalifa Mansouri, "Factors Affecting Student Academic Performance: A Combined Factor Analysis of Mixed Data and Multiple Linear Regression Analysis," *IEEE Access*, vol. 13, pp. 15946-15964, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[28] Xianchao Xie et al., "Matrix-Variate Factor Analysis and its Applications," *IEEE Transactions on Neural Networks*, vol. 19, no. 10, pp. 1821-1826, 2008. [CrossRef] [Google Scholar] [Publisher Link]
[29] Giacomo Della Riccia, and Alexander Shapiro, "Fisher Discriminant Analysis and Factor Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 1, pp. 99-104, 1983. [CrossRef] [Google Scholar] [Publisher Link]