Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.45

A Holistic Approach to Subject Correlation Analysis in Secondary Education

Buddhi Ayesha
Rowan University
rathna55@rowan.edu

Adessha Jayasooriya
University of Moratuwa
adeesha.14@cse.mrt.ac.lk

Wishmitha Mendis
University of Moratuwa
wishmitha.14@cse.mrt.ac.lk

Bhanuka Mahanama
University of Moratuwa
bhanuka.14@cse.mrt.ac.lk

Malaka Dayasiri
University of Moratuwa
malaka.14@cse.mrt.ac.lk

Umashanger Thayasivam
Rowan University
thayasivam@rowan.edu

Uthayasanker Thayasivam
University of Moratuwa
rtuthaya@cse.mrt.ac.lk

ABSTRACT

This study presents a holistic investigation of subject correlations in secondary education by drawing on performance data from more than 600 students across grades 6 to 8 in Sri Lanka. By using correlation analysis, regression models, factor analysis, and hierarchical clustering, we reveal key interrelationships among core subjects, such as mathematics, science, and language studies, alongside broader disciplines, such as citizenship education and art. Our results confirm the robust influence of reading proficiency on science achievement, outpacing the traditionally studied mathematics-science link, and underscores the value of language skills in mastering diverse subjects. Factor analysis identifies a dominant general academic ability that spans multiple areas, particularly language and humanities, whereas clustering underscores that some subjects, such as art and practical and technical skills, cluster distinctly. These findings advocate for interdisciplinary teaching methods and targeted interventions, shedding light on students' varied learning trajectories and informing policy to enhance overall educational outcomes.

Keywords: *Educational data mining, subject correlation, holistic analysis, factor analysis, hierarchical clustering.*

1. Introduction

Exploring associations among academic subjects is a significant area of research in educational data mining, with correlation analysis widely used to investigate these relationships. This study presents a holistic approach to identifying subject correlations at the secondary school level, which encompasses all significant aspects of the educational experience (Mahmoudi et al. 2012). By analyzing correlations across all subjects, we aim to improve the learning experience for middle school students and inform educational policy changes.

One of the main challenges of a holistic approach is the high dimensionality due to the increased number of subjects, which requires a large sample size for accurate clustering and analysis. Although significant work has been done on subject-level correlations (Wang 2005), most studies focused on specifically selected subjects, potentially overlooking influences from other areas. Because an academic term can include many subjects, disregarding some subjects may lead to incomplete or biased results. Therefore, a holistic analysis is necessary to accurately reflect the true nature of subject interrelationships.

This paper presents a framework and methodology for analyzing correlations among different subjects by using a holistic approach. By using advanced data mining techniques, we identify patterns and correlations that may be missed in studies confined to part of the syllabus (Wang 2005; O'Reilly and McNamara 2007; Maerten-Rivera et al. 2010). Recent studies emphasize the effectiveness of holistic and interdisciplinary methods in educational settings, such as advanced machine learning algorithms for predictive analysis (Yağcı 2022), holistic educational frameworks (Miseliunaite et al. 2022), and the integration of science, technology, engineering, arts, and mathematics (STEAM) into education (Marín-Marín et al. 2021). These findings affirm the need for a holistic perspective in educational research to enhance learning outcomes across various environments.

This paper is organized as follows: Section 2 provides a comprehensive review of related studies on correlation analysis in education. Section 3 details the methodology adopted for this research, including data collection, preprocessing, and analytical techniques. Section 4 presents the empirical findings and offers a critical discussion of the results. Section 5 draws the conclusions of the study and highlights potential directions for future research.

2. Literature Review

The exploration of subject correlations in education has been a focal point for researchers aiming to enhance student performance and inform educational policies. However, traditional studies often focus on specific subject pairs or small sample sizes, potentially overlooking the complex interrelations within a broader curriculum. This literature review critically examines existing research on subject correlation analysis, highlights the methodologies used, and identifies gaps that the current study aims to address.

2.1. Subject Correlations and Academic Performance

Understanding how proficiency in one subject area may influence or predict performance in another is a key concern in educational research. For example, Ünal et al. (2023) conducted a meta-analysis that explored the sources of correlation between reading and mathematics achievement. They concluded that both domain-general cognitive abilities and domain-specific skills contribute to the observed correlations, highlighting the multifaceted nature of academic performance. Similarly, Jindra et al. (2022) provided observational evidence of the reciprocal relationship between reading and mathematics proficiency over time.

O'Reilly and McNamara (2007) conducted a longitudinal study that investigated how reading comprehension skills impact science achievement. By using hierarchical linear modeling, they found that reading skills significantly predict science performance, especially in understanding complex scientific texts. Cruz Neri et al. (2021) also emphasized that language proficiency is crucial for students to articulate scientific concepts effectively, which suggests that interventions to improve language skills could positively affect science achievement.

The relationship between mathematics and science has also been extensively studied. Wang (2005) analyzed data from eighth-grade students by using structural equation modeling and found a bi-directional relationship between mathematics and science achievement. Strong mathematical skills provide a foundation for understanding scientific concepts, whereas engagement in science reinforces mathematical understanding. However, these studies often focus on specific grades or education levels, which may limit the generalizability of their findings.

2.2. Methodological Approaches in Subject Correlation Studies

Methodologies in subject correlation research have evolved to incorporate more sophisticated statistical and data mining techniques. Traditional methods such as Pearson and Spearman correlation coefficients are commonly used

for initial exploratory analyses (Barnard-Brak et al. 2017). Although effective for measuring linear and monotonic relationships, these methods may not capture the complexity inherent in educational data, which often involve high-dimensional and non-linear interactions.

Advanced statistical methods, such as factor analysis and structural equation modeling, have been used to uncover latent variables that influence multiple subjects simultaneously. For instance, Barnard-Brak et al. (2017) used confirmatory factor analysis to examine underlying constructs that affect reading and mathematics proficiency, revealing significant roles of cognitive and non-cognitive factors.

Machine learning techniques are increasingly adopted in educational data mining to predict student performance and identify patterns in large datasets. Yağcı (2022) applied ensemble learning methods to predict academic success, demonstrating that algorithms such as random forests and gradient boosting outperform traditional regression models in handling complex, non-linear relationships.

Clustering algorithms, such as hierarchical clustering and k-means, have been used to group students or subjects based on performance metrics. Mahanama et al. (2018) used hierarchical clustering to identify subject groupings, which provides insights into curriculum development and personalized learning strategies.

Despite these advancements, many studies remain limited in scope, often focusing on specific subjects or small, homogeneous samples. External factors such as socioeconomic status, parental education, and learning styles are frequently overlooked, potentially confounding the relationships among subjects (Beylik and Genç Kumtepe 2021).

2.3. Holistic Educational Frameworks and Interdisciplinary Approaches

The shift toward holistic education emphasizes integrating personal, social, and academic development. Mahmoudi et al. (2012) argue that holistic approaches consider the whole learner by promoting interconnected learning experiences across disciplines. This perspective aligns with the integration of STEAM education, which advocates for interdisciplinary learning to foster creativity and critical thinking (Marín-Marín et al. 2021).

Recent studies support the effectiveness of holistic and interdisciplinary methods. Miseliunaite et al. (2022a) conducted a systematic literature review and found that holistic education frameworks contribute to improved student engagement and motivation. The inclusion of arts in science, technology, engineering, and mathematics (STEM) education (forming STEAM) has been shown to enhance problem-solving skills and innovation (Yakman and Lee 2012).

However, implementing holistic education faces challenges, such as curriculum rigidity and assessment practices that favor subject-specific achievements. Miseliunaite et al. (2022) highlight the need for systemic changes to fully realize the benefits of holistic educational approaches.

2.4. Critical Analysis of Existing Research

Although significant progress has been made, several gaps persist in the literature:

Limited Scope of Studies: Many studies focus narrowly on specific subject pairs or education levels, which limit the applicability of findings across different contexts. This narrow focus may lead to incomplete understandings of how various subjects interrelate within a comprehensive curriculum.

Methodological Constraints: Traditional statistical methods may not adequately capture the complex, non-linear relationships in educational data. There is a need for more sophisticated analytical techniques that can handle high-dimensional data and uncover deeper insights.

Underrepresentation of External Factors: Socioeconomic status, parental education, and learning styles are often underrepresented in analyses, despite their significant impact on student performance. Ignoring these factors can result in biased findings and limit the effectiveness of proposed educational interventions.

Geographic and Cultural Limitations: The majority of research is concentrated in Western countries, which may not account for cultural and educational differences in other regions. This limits the generalizability of findings and the development of globally applicable educational strategies.

2.5. Contribution of the Current Study

The present study addresses the aforementioned gaps by adopting a holistic approach to analyzing subject correlations in secondary education. Key contributions include the following:

Comprehensive Subject Analysis: By analyzing correlations across all subjects within the curriculum, the study provides a more complete picture of subject interrelationships, identifying overarching patterns and highly correlated subject categories.

Advanced Methodological Framework: Using a combination of correlation analysis, regression, factor analysis, and hierarchical clustering allows for a nuanced examination of complex relationships in the data. This methodological rigor enhances the reliability of the findings.

Inclusion of External Factors: The study incorporates variables such as socioeconomic status, parental education, and learning styles, providing a more comprehensive understanding of factors that influence student performance.

Diverse Sample and Context: By collecting data from more than 600 students across urban, suburban, and rural regions in Sri Lanka, the study adds valuable insights from a non-Western context, which contributes to the global discourse on educational data mining.

2.6. Relevance to Educational Policy and Practice

The findings of this study have practical implications for educators and policymakers. By identifying significant patterns of subject correlations, the research can inform curriculum development, teaching strategies, and resource allocation. Emphasizing a holistic approach aligns with contemporary educational goals of fostering well-rounded learners equipped with interdisciplinary skills.

In summary, although existing research has laid the groundwork for understanding subject correlations in education, there is a clear need for more holistic, methodologically robust studies that consider a wider range of subjects and external factors. The current study sought to fill this gap by offering a comprehensive analysis that can contribute to improved educational strategies and student outcomes.

3. Methodology

This study used a comprehensive methodological framework designed to holistically analyze subject correlations in secondary education. The methodology encompassed detailed data collection procedures; meticulous data preprocessing; and the application of various advanced analytical techniques, including correlation analysis, regression analysis, factor analysis, and hierarchical clustering. Each component was elaborated to ensure clarity and depth, addressing the complexities involved in the research process and responding to the reviewers' recommendations.

3.1. Data Collection Process

The data collection was conducted across multiple government schools in Sri Lanka, which involved a sample of more than 600 students from diverse urban, suburban, and rural regions (Mahanama et al. 2018). Schools were selected based on their geographic representation and willingness to participate, which ensured a broad spectrum of socioeconomic backgrounds. The student sample was balanced in terms of gender, with approximately equal numbers of male and female students, enhancing the generalizability of the findings within the Sri Lankan educational context.

By focusing on students in grades 6, 7, and 8 (ages 11 to 14 years), performance data were collected over three consecutive academic years. This longitudinal approach provided insights into student progress and developmental trends over time. Specifically, end-term examination marks for all the subjects were gathered for each student, which encompassed core academic areas such as mathematics, science, Sinhala (primary language), English (secondary language), religion, history, health, citizenship education, geography, practical and technical skills (PTS), Tamil (secondary national language), and art.

In addition to academic performance data, comprehensive information on student learning backgrounds and learning styles was collected to enrich the analysis. This included socioeconomic indicators (parents' education levels and occupations), family background details, participation in supplementary educational support such as private tuition, and engagement in extracurricular activities such as sports, clubs, and arts programs.

To ensure the reliability and validity of the data, a stratified random sampling method was used. Schools were stratified based on geographic location and type (urban, suburban, rural), and within each stratum, schools were randomly selected. Within the selected schools, students were randomly chosen from the relevant grades to participate in the study. This sampling strategy aimed to minimize selection bias and ensure that the sample was representative of the broader student population.

Potential biases, for example, non-response bias, were addressed by encouraging participation through clear communication of the study's purpose, ensuring confidentiality, and obtaining the necessary ethics approvals. However, it is acknowledged that schools that declined participation might share characteristics that influence the findings, and this limitation was considered in the interpretation of results.

Ethical considerations were paramount throughout the data collection process. Ethics approval was obtained from the institutional review board of the affiliated university, which adhered to international standards for research that involves minors. Informed consent was secured from both the students and their parents or legal guardians. Confidentiality was maintained by anonymizing personal identifiers and securely storing data, and participants were informed of their right to withdraw from the study at any point without repercussions.

3.2. Data Collection Dimensions and Techniques

The data collection encompassed three primary dimensions: student performance data, student learning background data, and student learning style data. Student performance data included examination marks and assignment scores for each subject, obtained directly from official school records. Student learning background data were collected via structured questionnaires, which captured variables such as socioeconomic status, family background, access to additional educational support, and engagement in extracurricular activities. Student learning style data were assessed by using a questionnaire modeled after the Learning Connections Inventory (LCI) model, which evaluates individual learning preferences across scales such as sequence, precision, technical reasoning, and confluence.

To accommodate the large sample size and diverse participant backgrounds, a combination of data collection techniques was used. Both paper-based and digital questionnaires were administered to collect learning background and style data. Multiple-choice and Likert-scale questions facilitated ease of response and efficient data processing. Academic performance data were collected by obtaining permission from school administrators to access official records, with data extraction conducted on-site to ensure data integrity and adherence to confidentiality protocols.

Given the prevalence of handwritten records in schools, data digitalization was a critical step. Data were manually entered into secure electronic databases, and, to enhance accuracy and efficiency, optical mark recognition technology was used for processing questionnaire responses when feasible by using tools such as scripts for data acquisition with paper-based surveys. Pilot testing of questionnaires and data collection procedures was conducted to refine instruments and ensure clarity and appropriateness for the target age group.

The choice of data collection medium was influenced by the technological infrastructure available at each school. In schools with adequate IT facilities, digital questionnaires were administered by using computers or tablets. In contrast, paper-based questionnaires were used in schools that lacked such resources. This dual approach ensured inclusivity and maximized participation rates, with the familiarity of students with the chosen medium reducing response bias and enhancing data quality.

3.3. Data Preprocessing

Before analysis, the collected data underwent meticulous preprocessing to ensure validity and reliability. Data cleaning involved scrutinizing the dataset for inconsistencies, missing values, and outliers. In cases of incomplete records, efforts were made to retrieve missing information, and, if retrieval was not possible, then missing values were handled by using mean substitution or appropriate imputation techniques.

Examination scores were standardized to account for variations in grading scales across different subjects and schools. Z-scores were calculated to normalize the data, which facilitated meaningful comparisons. Categorical data from questionnaires, such as parental occupation and learning style preferences, were encoded by using numerical representations. For nominal variables, one-hot encoding was applied, whereas ordinal variables were encoded based on their inherent order.

The internal consistency of the questionnaires was assessed by using Cronbach's alpha, with a high reliability coefficient of 0.98, which indicated strong internal consistency among the questionnaire items. This process ensured that the data were suitable for subsequent advanced analytical techniques.

3.4. Analytical Techniques

To thoroughly analyze the data and address the research objectives, several advanced analytical techniques were applied, including correlation analysis, regression analysis, factor analysis, and hierarchical clustering. These methods are widely used in educational data mining to uncover patterns and relationships within complex datasets (Romero and Ventura 2010).

3.4.1. Correlation analysis

Correlation analysis was used to examine the relationships between different academic subjects. Specifically, the Pearson correlation coefficient, Spearman rank correlation coefficient, and Kendall tau were calculated to capture both linear and monotonic associations (Khamis 2008). Using multiple correlation measures allowed for a robust understanding of the inter-subject relationships, accommodating potential non-linearities and the ordinal nature of some data. Correlation matrices and heatmaps were generated to visualize these relationships and identify significant patterns among subjects.

3.4.2. Regression analysis

Multiple linear regression models were constructed to predict student performance in key subjects based on their scores in other subjects and background variables. This approach enabled the exploration of predictive relationships and the quantification of the impact of various factors on academic outcomes (Cohen et al. 2013). Standard diagnostic tests were conducted to ensure the validity of the regression models, including checks for multicollinearity, heteroscedasticity, and normality of residuals.

3.4.3. Factor analysis

Exploratory Factor Analysis was conducted to identify latent constructs that underlie students' performance across different subjects. The suitability of the data for factor analysis was assessed by using the Kaiser-Meyer-Olkin measure and Bartlett Test of Sphericity (Kaiser 1974). Principal axis factoring with promax rotation was used to extract factors, which allowed for correlated factors, which is appropriate given the interconnected nature of academic abilities (Fabrigar et al. 1999). Factors with eigenvalues greater than 1 were retained based on the Kaiser criterion.

3.4.4. Hierarchical cluster analysis

Hierarchical clustering was applied to group subjects based on similarities in student performance patterns. Cosine similarity was used as the distance metric due to its effectiveness in high-dimensional spaces (Tan et al. 2021). Agglomerative hierarchical clustering with average linkage was performed, and dendrograms were generated to visualize the clustering process and identify natural groupings among subjects.

3.4.5. Principal component analysis

Principal component analysis (PCA) was used as a dimensionality reduction technique to simplify the data while retaining most of the variance (Jolliffe and Cadima 2016). Standardized data were used for PCA to ensure each variable contributed equally. The principal components that explain significant variance were retained and used as inputs for clustering algorithms to enhance computational efficiency and mitigate the curse of dimensionality.

3.5. Software and Tools

A web application was developed to collect data, manage the dataset, and visualize the results of the study. The front end of the application was built by using Angular, which provides an interactive interface for data entry and real-time visualization. The back end was developed by using Node.js with the Express.js framework, handling data processing and ensuring secure and efficient management of the collected information. MongoDB was used as the database to store and manage the data due to its flexibility and scalability in handling large datasets.

For data analysis, Python was used due to its comprehensive libraries suitable for statistical analysis and machine learning. Libraries such as pandas were used for data manipulation and cleaning, NumPy for numerical computations, SciPy for statistical functions, scikit-learn for implementing machine learning algorithms, and statsmodels for advanced statistical modeling. Visualization was performed by using Matplotlib and Seaborn libraries, which were instrumental in generating plots, heatmaps, dendrograms, and other visual representations essential for interpreting the data.

This integrated software environment facilitated efficient data handling from collection to analysis, ensuring the integrity and accessibility of data throughout the research process.

4. Analysis

The analysis consists of two main parts: correlation analysis of subjects and hierarchical cluster dendrogram analysis. The correlation analysis shows a year-by-year analysis of subjects, whereas the hierarchical clustering provides an overall analysis of subjects.

4.1. Correlation Analysis

We used three fundamental correlation techniques, Pearson, Spearman, and Kendall, to examine relationships among subjects across different grades. Although Pearson and Spearman indicated higher correlation scores, Kendall produced moderate scores and provided complementary insights into monotonic (rather than strictly linear) relationships. Overall, all three methods revealed a consistent pattern of correlations, indicating that a multi-method approach offers a more comprehensive understanding of the varying strengths and nature of subject interrelationships.

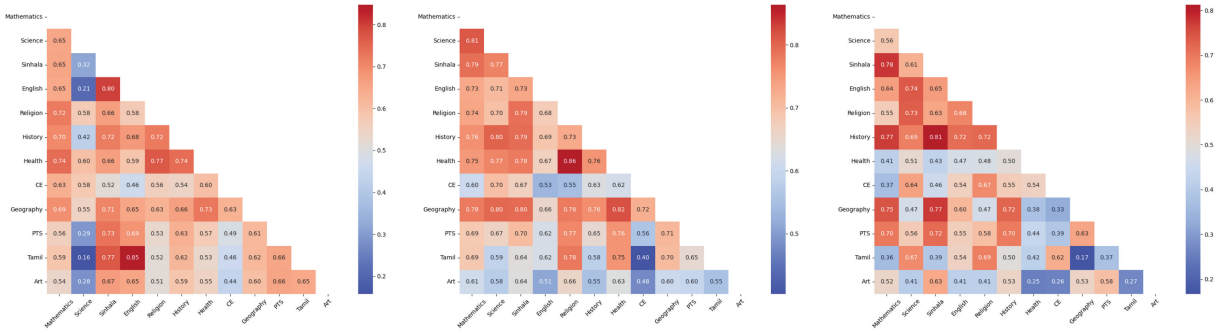
4.1.1. Correlation analysis on subjects

Shown in Figures 1, 2, and 3 are that Pearson and Spearman uncovered more high-correlation pairs among subjects compared with Kendall, which identified fewer but still meaningful, high correlations. Notably, Pearson and Spearman highlighted several strongly correlated pairs, as reflected in Tables 1, 2, and 3. Whereas Kendall coefficients were generally smaller, they confirmed the same overall distribution of subject relationships and underscored the importance of analyzing different types of associations.

The Spearman and Pearson methods identified more relationships between the subjects, which the Kendall method did not, as shown in Tables 1, 2, and 3. This suggests that the subjects do not have a completely linear relationship but rather a monotonic relationship.

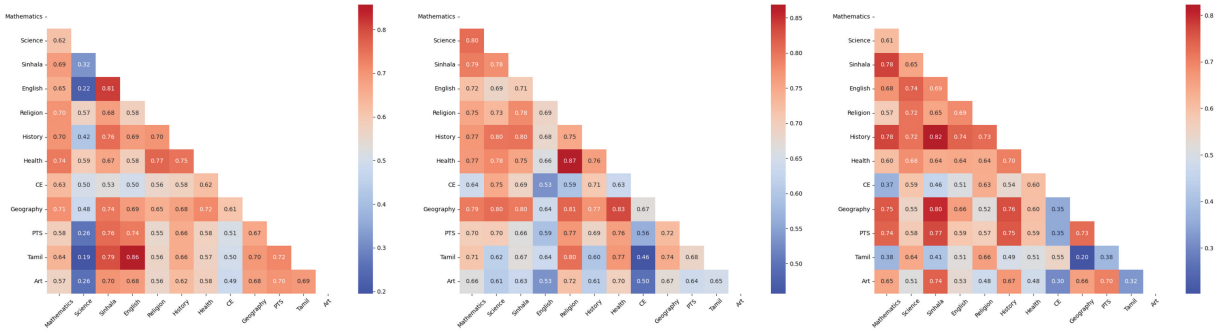
4.1.2. Correlation analysis on all grades

In this analysis, we included data from all grades (grade 6, grade 7, and grade 8). Figure 4 shows the correlation matrix heatmap, in which we identified several key patterns of relationships among different subjects.



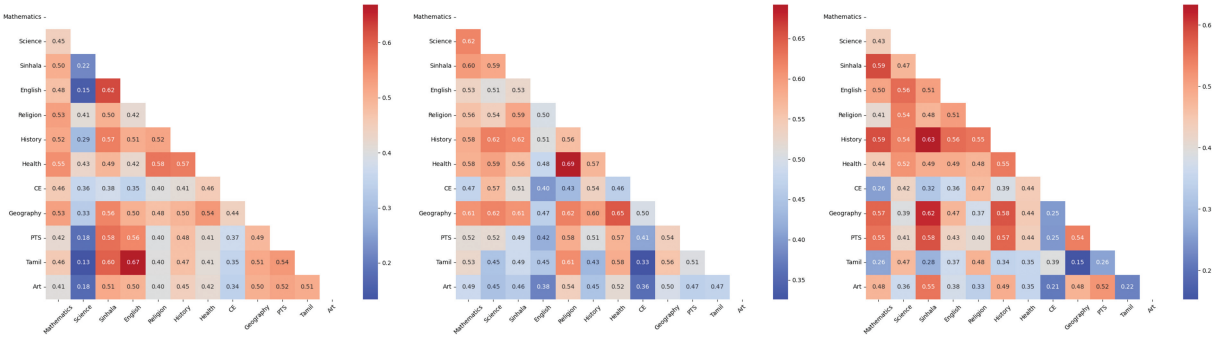
(a) Grade 6 (b) Grade 7 (c) Grade 8

Figure 1: Pearson correlation: (a) grade 6, (b) grade 7, (c) grade 8.



(a) Grade 6 (b) Grade 7 (c) Grade 8

Figure 2: Spearman correlation: (a) grade 6, (b) grade 7, (c) grade 8.



(a) Grade 6 (b) Grade 7 (c) Grade 8

Figure 3: Kendall correlation: (a) grade 6, (b) grade 7, (c) grade 8.

Table 1: Highly correlated subjects in grade 6.

Subject 1	Subject 2	Pearson	Spearman
Sinhala	English	0.80	0.81
English	Tamil	0.85	0.86

Table 2: Highly correlated subjects in grade 7.

Subject 1	Subject 2	Pearson	Spearman
Mathematics	Science	0.81	0.80
Science	History	0.80	0.80
Science	Geography	0.80	0.81
Sinhala	History	0.79	0.80
Sinhala	Geography	0.80	0.80
Religion	Health	0.86	0.87
Religion	Geography	0.78	0.81
Religion	Tamil	0.78	0.80
Health	Geography	0.82	0.83

Table 3: Highly correlated subjects in grade 8.

Subject 1	Subject 2	Pearson	Spearman
Sinhala	History	0.81	0.82
Sinhala	Geography		0.80

First, there were high positive correlations between mathematics and a range of other subjects. Specifically, strong positive correlations were observed between mathematics and science (0.857), Sinhala (0.882), English (0.882), religion (0.839), history (0.881), health (0.837), and geography (0.867). This suggests that students who excel in mathematics tend to perform well in these subjects as well. Similarly, science also exhibited strong positive correlations with mathematics (0.857), Sinhala (0.866), religion (0.833), history (0.883), health (0.873), and geography (0.863). Furthermore, Sinhala showed high correlations with religion (0.920), history (0.934), and health (0.893).

Moderate positive correlations were observed between citizenship-education and several subjects, including mathematics (0.739), science (0.848), Sinhala (0.809), religion (0.781), history (0.821), and geography (0.777). In addition, English demonstrates moderate correlations with mathematics (0.882), Sinhala (0.868), religion (0.849), and history (0.847).

Lower positive correlations were noted between academic subjects and extracurricular activities, such as club activities, racquet sports, and monitor roles. These lower correlations indicate a weaker relationship between academic performance and participation in these extracurricular activities.

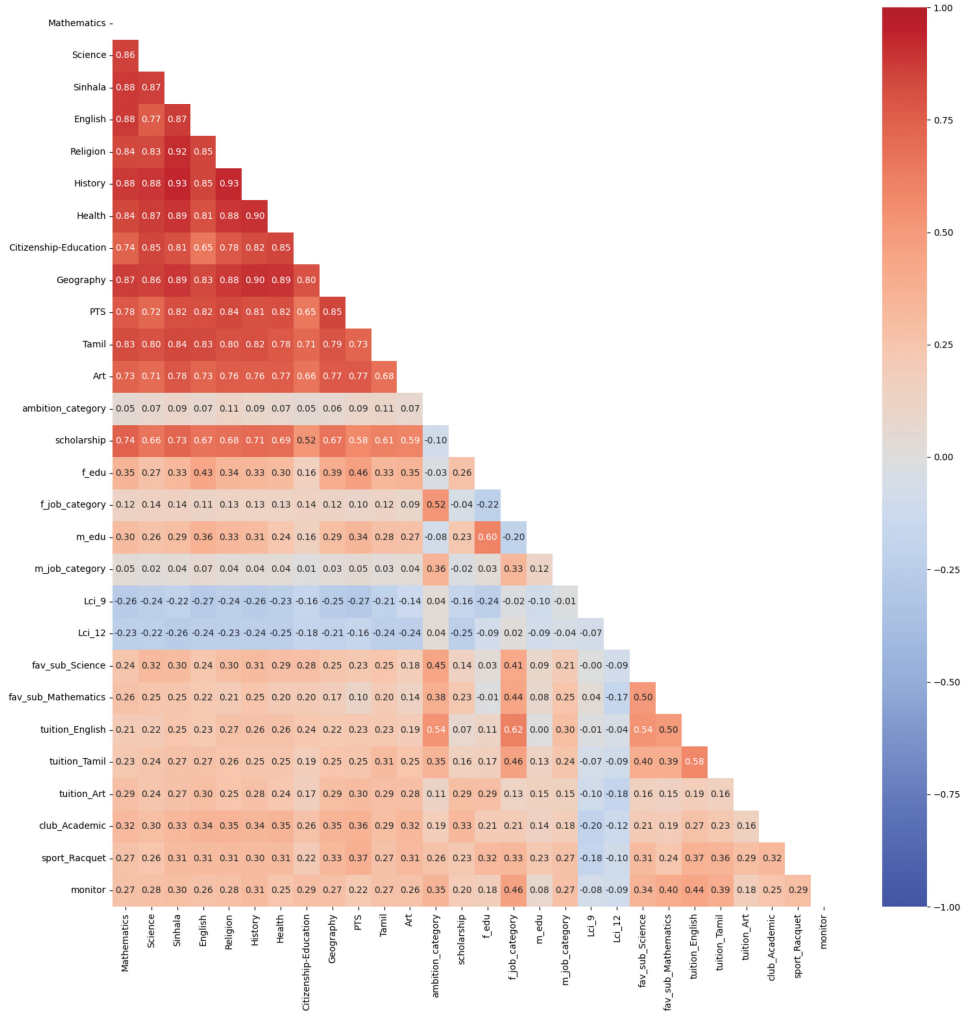


Figure 4: Correlation matrix of all grades.

Negative correlations were found with the variables LCI_9 and LCI_12, which show negative correlations with most subjects. This suggests that these variables may represent factors that inversely affect academic performance.

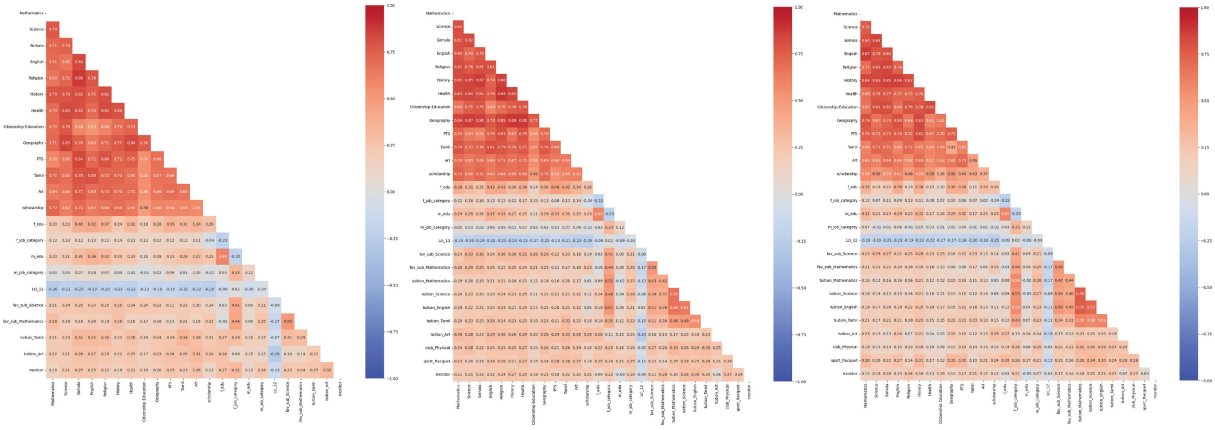
Lastly, interesting observations include small-to-moderate positive correlations between students' favorite subjects (science and mathematics) and their overall academic performance. This indicates that a student's favorite subjects can have a positive impact on his or her performance across other subjects.

4.1.3. Correlation analysis for grade 6

The correlation matrix for grade 6 (Figure 5(a)) revealed significant insights into the relationships among various subjects and other factors. Mathematics exhibited strong positive correlations with science (0.78), history (0.78), and Sinhala (0.71). Science showed high correlations with mathematics (0.78), history (0.79), and health (0.83). Similarly, Sinhala had strong correlations with English (0.84), religion (0.88), and health (0.81). English showed notable correlations with Sinhala (0.84), religion (0.78), and mathematics (0.71). Religion had high correlations with Sinhala (0.88), science (0.72), and health (0.82). In addition, extracurricular activities and parental education levels exhibited moderate correlations with academic performance. For instance, parental education (*f_edu*, *m_edu*) showed a moderate positive correlation with academic subjects, which indicated the influence of parental background on student performance.

4.1.4. Correlation analysis for grade 7

In grade 7, the correlation matrix (Figure 5(b)) indicated a pattern similar to grade 6 but with varying strengths. Mathematics demonstrated high positive correlations with science (0.82), history (0.83), and Sinhala (0.81). Science



(a) Correlation Matrix Grade 6 (b) Correlation Matrix Grade 7 (c) Correlation Matrix Grade 8

Figure 5: Correlation matrices for grades 6, 7, and 8: (a) correlation matrix, grade 6; (b) correlation matrix, grade 7; (c) correlation matrix, grade 8.

showed strong correlations with mathematics (0.82), history (0.85), and geography (0.79). Sinhala had high correlations with religion (0.83), history (0.82), and mathematics (0.81). English exhibited significant correlations with science (0.70), Sinhala (0.76), and religion (0.75). Religion had high correlations with Sinhala (0.83), mathematics (0.82), and history (0.82). Additional factors, such as extracurricular activities (tuition, clubs) and parental occupation categories, showed moderate correlations, which reflects their impact on students' academic achievements.

4.1.5. Correlation analysis for grade 8

The correlation matrix for grade 8 (Figure 5(c)) reveals that mathematics had strong positive correlations with English (0.87), history (0.84), and Sinhala (0.80). Science showed high correlations with Sinhala (0.84), mathematics (0.76), and religion (0.82). Sinhala had notable correlations with religion (0.85), history (0.85), and science (0.84). English demonstrated significant correlations with mathematics (0.87), history (0.86), and Sinhala (0.80). Religion showed high correlations with Sinhala (0.85), history (0.82), and science (0.82). Furthermore, analysis of the data indicated that extracurricular activities and parental education continue to have moderate correlations with students' performance across various subjects.

These analyses underscore the interconnectedness of different subjects and the influence of external factors, providing a comprehensive understanding of the academic dynamics for each grade.

4.2. Regression Analysis

In this study, we performed regression analyses to predict the scores of mathematics, science, and Sinhala for grades 6, 7, and 8 by using the marks of other subjects as predictors. The analyses were conducted separately for each grade. The results of these regression analyses are summarized in Table 4.

Table 4: Summary of regression analysis results.

Grade	Target	Mean Squared Error	R^2
6	Science	219.01	0.55
6	Sinhala	146.76	0.66
6	Mathematics	92.75	0.79
7	Science	97.37	0.81
7	Sinhala	46.15	0.87
7	Mathematics	113.78	0.79
8	Science	116.52	0.72
8	Sinhala	40.82	0.87
8	Mathematics	120.13	0.78

The results indicate that the regression models for predicting mathematics, science, and Sinhala scores in grades 6, 7, and 8 have varying degrees of success. The R^2 values for the models ranged from 0.55 to 0.87, which suggests that a significant proportion of the variance in the target scores can be explained by the predictor variables. The models for predicting Sinhala scores generally performed better, with R^2 values above 0.85 for grades 7 and 8.

For grade 6, the regression model for predicting mathematics scores had an R^2 value of 0.79, which indicates that approximately 79% of the variance in mathematics scores is explained by the scores in other subjects. The models for science and Sinhala had R^2 values of 0.55 and 0.66, respectively.

In grade 7, the models showed improved performance with R^2 values of 0.81 for science, 0.87 for Sinhala, and 0.79 for mathematics. This suggests that the predictor variables are more effective in explaining the variance in target scores for this grade level.

The regression models for grade 8 also demonstrated a strong performance, particularly for Sinhala, with an R^2 value of 0.87. The models for science and mathematics had R^2 values of 0.72 and 0.78, respectively.

Overall, these regression analyses provide valuable insights into the relationships among different subjects' scores and highlight the effectiveness of using other subjects' marks as predictors for the target scores in each grade.

4.3. Factor Analysis

To identify the underlying relationships among the different subjects, a factor analysis was conducted. A factor analysis was performed by using the following steps:

1. **Data Standardization:** The marks for each subject were standardized to ensure that all variables were on the same scale.
2. **Correlation Matrix Calculation:** The correlation matrix of the subjects was calculated to understand the relationships between them.
3. **Determining the Number of Factors:** Eigenvalues were calculated, and a scree plot was used to determine the number of factors. It was found that one factor had an eigenvalue greater than 1.
4. **Factor Extraction:** Factor analysis was performed with one factor, and the factor loadings were obtained.
5. **Handling Missing Values:** Missing values in the dataset were handled by filling them with the mean of each column.

The following table, Table 5, summarizes the factor loadings for each subject across three factors:

The factor analysis reveals the following insights:

- **Factor 1:** This factor has strong negative loadings for all the subjects, which indicates a general academic performance factor. The highest loadings are observed for Sinhala (–0.961), religion (–0.943), and history (–0.962).
- **Factor 2:** This factor shows very low positive loadings across all the subjects, which suggests that it might not significantly contribute to the variance in academic performance.
- **Factor 3:** This factor has a moderate positive loading for English (0.292), which indicates a specific factor that might be related to language skills or preferences.

These results highlight the underlying structure of the academic performance data, with factor 1 representing a general academic ability and factors 2 and 3 capturing more-specific dimensions of student performance. The factor loadings plot (Figure 6) and the scree plot (Figure 7) further illustrate the contribution of each factor to the overall variance in the dataset.

This factor analysis provides a comprehensive understanding of the underlying dimensions of academic performance, aiding in the identification of key areas for targeted interventions and support.

4.4. Hierarchical Cluster Analysis

Hierarchical clustering was applied to identify similar subjects in the syllabus. For this purpose, each subject was represented by a vector of 12 dimensions. Each dimension's value was calculated based on the Pearson correlation coefficient of each subject against other subjects. Cosine similarity was used to calculate the similarity between

Table 5: Factor loadings for various subjects and variables.

Subject/Variable	Factor 1	Factor 2	Factor 3
Mathematics	–0.917	0.075	0.105
Science	–0.911	0.049	–0.185
Sinhala	–0.961	0.032	0.017
English	–0.895	0.048	0.292
Religion	–0.943	0.025	0.036
History	–0.962	0.033	–0.034
Health	–0.937	0.043	–0.108
Citizenship-education	–0.848	0.041	–0.379
Geography	–0.937	0.074	–0.002
Practical and technical skills	–0.858	0.050	0.188
Tamil	–0.863	0.032	0.096
Art	–0.803	0.056	0.080
Ambition category	–0.117	–0.625	0.014
Scholarship	–0.588	0.159	0.107
Father's job category	–0.171	–0.702	–0.034
Mother's job category	–0.060	–0.379	0.090
Learning Connections Inventory 9	0.197	–0.081	–0.119
Learning Connections Inventory 12	0.196	–0.009	–0.043
Favorite subject: Sinhala	0.144	–0.433	–0.203
Favorite subject: mathematics	–0.264	–0.540	–0.043
Favorite subject: science	–0.327	–0.541	–0.125
Tuition: mathematics	–0.236	–0.878	0.029
Tuition: science	–0.251	–0.811	0.034
Tuition: English	–0.301	–0.816	–0.012
Tuition: Tamil	–0.300	–0.581	0.115

subjects. The dendrograms obtained by this approach provide an overview of the grouping of subjects. Due to the use of cosine similarity, the subjects in the same cluster require similar skill sets.

In addition to PTS, art, and citizenship education, which have a larger distance from other subjects, the hierarchical cluster dendrogram identified two clear clusters among the subjects. Mathematics and English language are in the same cluster, as shown in Figure 8, which have the highest failure rates in the ordinary level examination (Department of Examinations, Sri Lanka 2017). Moreover, Cruz Neri et al. (2021) state that mathematics and foreign languages are associated with each other. Furthermore, Bergen (2017) explains that mathematics can be thought of as a foreign language, with its unique terminology and symbol system. Therefore, the results shown in Figure 8 confirm that students in Sri Lanka also showed a similar attitude toward secondary languages and mathematics.

Religion, geography, history, and Sinhala have low distances from each other and are related to reading and memorizing. The next closest distances to these subjects are science and health, which are also related to reading skills, as described in the literature. The cluster dendrogram confirms that science has the closest distance to the above-identified reading cluster.

4.5. Spatial Cluster Analysis

Cluster analysis was performed by using various clustering algorithms to identify distinct groups within the student data. The following algorithms were used: DBSCAN, OPTICS, Mean Shift, HDBSCAN, GMM, and Spectral Clustering. This section summarizes the findings and provides an overview of the clusters identified. Among these, DBSCAN provided better clustering results in this case, identifying distinct clusters and noise points effectively.

DBSCAN (Figure 9) was applied to the PCA-reduced data with an epsilon value of 0.6 and a minimum sample size of 2. The clustering process resulted in the identification of several clusters, including noise points (denoted by cluster label –1). The characteristics of each identified cluster were analyzed, and the results are summarized below.

The analysis of DBSCAN clusters (Table 6) revealed distinct patterns in the academic performance and aspirations of students. Clusters 0, 1, and 2 exhibited higher average mathematics scores and a strong inclination toward

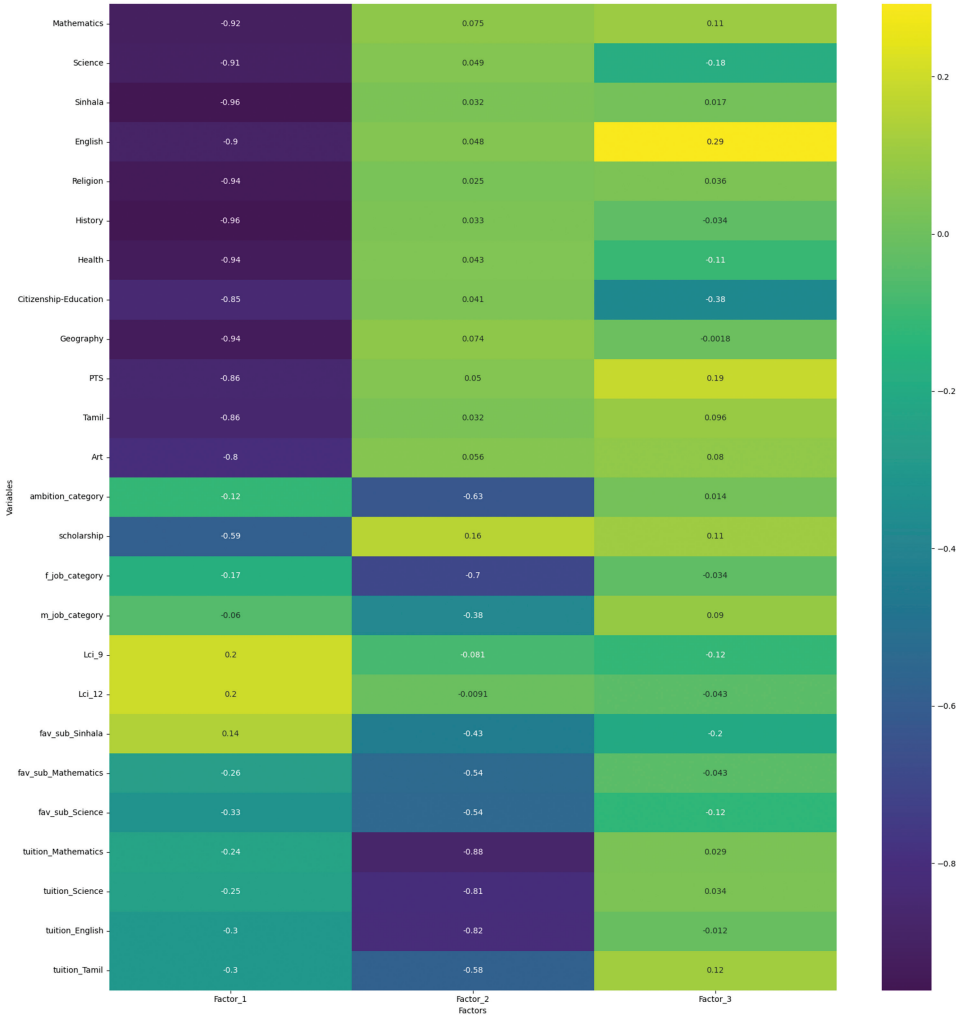


Figure 6: Factor loadings.

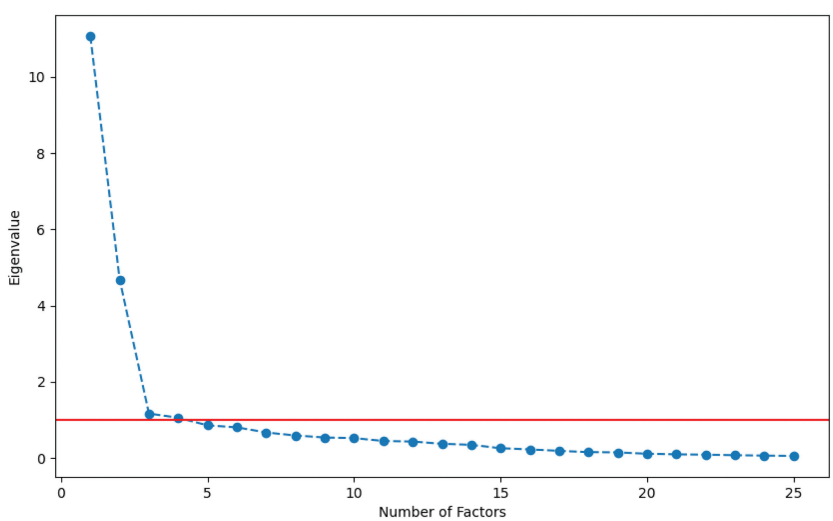
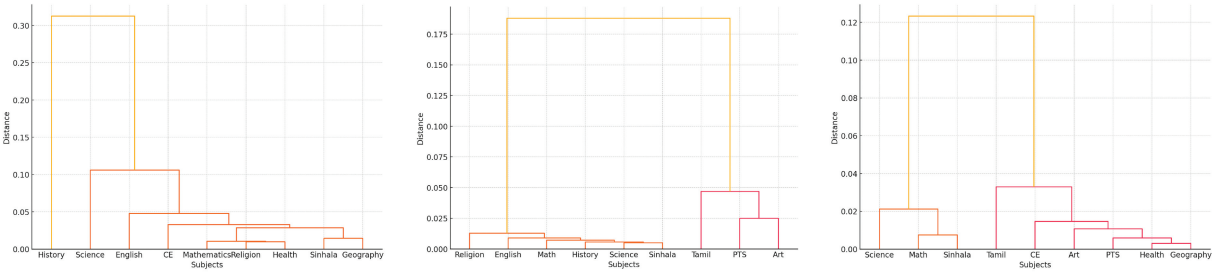


Figure 7: Scree plot.



(a) Cluster Dendrogram for Grade 6 (b) Cluster Dendrogram for Grade 7 (c) Cluster Dendrogram for Grade 8

Figure 8: Hierarchical clustering dendrograms for grades 6, 7, and 8: (a) cluster dendrogram for grade 6, (b) cluster dendrogram for grade 7, (c) cluster dendrogram for grade 8.

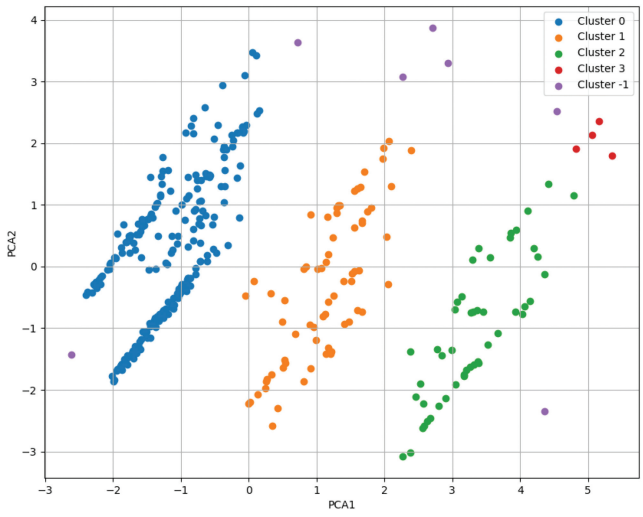


Figure 9: DBSCAN clusters visualization in two-dimensional plane.

health-care–related ambitions, with parents commonly engaged in sales or educational professions. Conversely, cluster 3, which had the lowest average mathematics score, also showed a divergence in parental occupations, which indicated a possible correlation between parental profession and student academic performance.

The noise points identified by DBSCAN (cluster –1) highlight students whose characteristics did not align closely with any other cluster, which suggests unique or outlier profiles.

5. Conclusion

This study presents a comprehensive analysis of subject correlations in secondary education through a holistic approach, encompassing a wide range of academic disciplines. By analyzing performance data from more than 600 Sri Lankan students across grades 6 to 8, we used advanced data mining techniques, including correlation analysis, regression, factor analysis, and hierarchical clustering, to uncover significant patterns in subject interrelationships.

Our findings reveal strong associations between reading and science achievement, consistent with existing literature that emphasizes the critical role of language skills in understanding scientific concepts (O'Reilly and McNamara 2007; Barnard-Brak et al. 2017). Notably, the correlation between reading and science was found to be stronger than that between science and mathematics, which highlights the importance of literacy in science education (Beylik et al. 2021; Jindra et al. 2022; Ünal et al. 2023). This suggests that interventions aimed at improving reading skills may have a substantial impact on students' performance in science.

The moderate association between science and mathematics observed in our analysis indicates that, although these subjects are related, they may require different cognitive skills or learning approaches at the lower secondary level.

Table 6: Characteristics of identified clusters by using DBSCAN.

Characteristic	Cluster –1 (Noise Points)	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Average mathematics score	66.52	70.97	72.07	69.81	57.50
Predominant ambition category	Educational instruction and library occupations	Health-care practitioners and technical occupations	Health-care practitioners and technical occupations	Health-care practitioners and technical occupations	Health-care practitioners and technical occupations
Common father's job category	Farming, fishing, and forestry occupations	Sales and related occupations	Sales and related occupations	Sales and related occupations	Installation, maintenance, and repair occupations
Common mother's job category	Management occupations	Educational instruction and library occupations	Educational instruction and library occupations	Educational instruction and library occupations	Educational instruction and library occupations
Father's education level	High school	Bachelor's degree	Bachelor's degree	Bachelor's degree	High school
Mother's education level	Bachelor's degree	Bachelor's degree	Bachelor's degree	Bachelor's degree	High school
Summary	Students with diverse ambitions and lower academic performance	Students with higher academic performance and ambition toward health-care professions	Students with the highest academic performance, aiming for health-care professions	Students with moderate academic performance, also inclined toward health-care professions	Students with the lowest academic performance, differing parental professions

This insight underscores the need for educators to tailor instructional strategies to address the specific demands of each subject. Given that our analysis did not include data beyond the eighth grade, future research could explore whether the association between science and mathematics strengthens in higher grades, as supported by another study (Wang 2005).

When comparing different correlation techniques, the Spearman correlation was more suitable for our dataset, identifying more relationships between the subjects than did the Pearson correlation coefficient. This suggests that the relationships among the subjects are more monotonic rather than strictly linear, which emphasizes the importance of selecting appropriate statistical methods to accurately capture the nuances in educational data.

The hierarchical clustering analysis revealed that subjects such as citizenship education, art, and PTS had greater distances from other subjects. This indicates that these subjects may assess unique skill sets or inherent abilities not directly linked to performance in other academic areas. Recognizing these distinctions can help educators design curricula that acknowledge the diverse talents and interests of students, potentially enhancing engagement and learning outcomes.

Factor analysis identified a general academic performance factor with strong negative loadings across all subjects, particularly Sinhala, religion, and history. This underscores the significance of a broad academic ability that spans multiple disciplines, reinforcing the idea that foundational skills in language and humanities are integral to overall academic success. Factors 2 and 3, although contributing less to the variance, highlighted specific dimensions related to language skills and other attributes, providing deeper insights into the components of student performance.

Cluster analysis by using various algorithms identified distinct groups of students with varying academic performance levels and parental education backgrounds. These clusters offer valuable insights into student profiles, enabling educators and policymakers to develop targeted interventions and support mechanisms that address the specific needs of different student groups. For instance, understanding that certain clusters of students may benefit from additional support in mathematics or language subjects can inform resource allocation and instructional strategies.

Overall, our study contributes to the field of educational data mining by demonstrating the value of a holistic approach in uncovering complex interrelationships among academic subjects. By integrating multiple analytical techniques, we provide a nuanced understanding of how various disciplines interact, which can inform the development of more effective educational strategies and policies. These findings highlight the interconnectedness of academic disciplines and the necessity for interdisciplinary approaches in education.

Future research should aim to expand the dataset to include a larger and more diverse sample, encompassing different regions and educational contexts. In addition, longitudinal studies could provide insights into how subject correlations evolve over time and across different educational stages. Exploring causal relationships by using advanced machine learning techniques would further enhance our understanding of the factors that influence student performance, ultimately contributing to improved educational outcomes and student success across various contexts.

References

Barnard-Brak, L., T. Stevens, and W. Ritter. 2017. "Reading and Mathematics Equally Important to Science Achievement: Results from Nationally-Representative Data." *Learning and Individual Differences* **58**, no. 1–9. doi:10.1016/j.lindif.2017.07.001.

Bergen, S.L. 2017. "Mathematics and Foreign Language: Authentic Texts in Mathematics." Accessed February 25, 2025. https://api.semanticscholar.org/CorpusID:189339966.

Beylik, A., and E. Genç Kumtepe. 2021. "Examining Transactional Distance in Synchronous Online Learning Environments." In *Motivation, Volition, and Engagement in Online Distance Learning*, edited by Hasan Uçar and Alper T. Kumtepe, 147–167. Hershey, PA: IGI Global. doi:10.4018/978-1-7998-7681-6.ch007.

Cohen, J., P. Cohen, S.G. West, and L. Aiken. 2013. "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences." 3rd edition. New York, NY: Routledge. doi: 10.4324/9780203774441.

Cruz Neri, N., K. Guill, and J. Retelsdorf. 2021. "Language in Science Performance: Do Good Readers Perform Better?" *European Journal of Psychology of Education* **36**, no. 1: 45–61. doi:10.1007/s10212-019-00453-5.

Department of Examinations, Sri Lanka. 2017. G.C.E. (O/L) Examination 2017 – Performance of Candidates. Accessed February 25, 2025. https://doenets.lk/documents/statistics/G.C.E.(OL)%20%20Examination%202017%20Perfomance%20of%20Candidates.pdf.

Fabrigar, L., D. Wegener, R.C. MacCallum, and E. J. Strahan. 1999. "Evaluating the Use of Exploratory Factor Analysis in Psychological Research." *Psychological Methods* **4**, no. 3: 272–299. doi:10.1037//1082-989X.4.3.272.

Jindra, C., K. Sachse, and M. Hecht. 2022. "Dynamics between Reading and Math Proficiency over Time in Secondary Education – Observational Evidence from Continuous Time Models." *Large-Scale Assessments in Education* **10**, no. 1: 12. doi:10.1186/s40536-022-00136-6.

Jolliffe, I.T., and J. Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, no. 2065: 20150202.

Kaiser, H.F. 1974. "An Index of Factorial Simplicity." *Psychometrika* **39**, no. 1: 31–36. doi:10.1007/BF02291575.

Khamis, H. 2008. "Measures of Association: How to Choose?" *Journal of Diagnostic Medical Sonography* **24**, no. 3: 155–162. doi:10.1177/8756479308317006.

Maerten-Rivera, J., N. Myers, O. Lee, and R. Penfield. 2010. "Student and School Predictors of High–Stakes Assessment in Science." *Science Education* **94**, no. 6: 937–962. doi:10.1002/sce.20408.

Mahanama, B., W. Mendis, A. Jayasooriya, V. Malaka, U. Thayasivam, and U. Thayasivam. 2018. "Educational Data Mining: A Review on Data Collection Process." In *Proceedings of the 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, September 27–28. doi: 10.1109/ICTER.2018.8615532.

Mahmoudi, S., E. Jafari, H. Nasrabadi, and M. Liaghatdar. 2012. "Holistic Education: An Approach for 21 Century." *International Education Studies* **5**, no. 3: 178–186. doi:10.5539/ies.v5n3p178.

Marín-Marín, J., A. Moreno Guerrero, P. Dúo Terrón, and J. López-Belmonte. 2021. "Steam in Education: A Bibliometric Analysis of Performance and Co-Words in Web of Science." *International Journal of STEM Education* **8**, no. 1: 41. doi:10.1186/s40594-021-00296-x.

Miseliunaite, B., I. Kliziene, and G. Cibulskas. 2022. "Can Holistic Education Solve the World's Problems: A Systematic Literature Review." *Sustainability* **14**, no. 15: 9737. doi: doi:10.3390/su14159737.

O'Reilly, T., and D. McNamara. 2007. "The Impact of Science Knowledge, Reading Skill, and Reading Strategy Knowledge on More Traditional" High-Stakes" Measures of High School Students' Science Achievement." *American Educational Research Journal* **44**, no. 1: 161–196. doi:10.3102/0002831206298171.

Romero, C., and S. Ventura. 2010. "Educational Data Mining: A Review of the State of the Art." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **40**, no. 6: 601–618. doi:10.1109/TSMCC.2010.2053532.

Tan, P.-N., M. Steinbach, A. Karpatne, and V. Kumar. 2021. *Introduction to Data Mining*. 2nd edition. Pearson: Boston, MA, USA. ISBN 9780137506286.

Ünal, Z., N. Greene, X. Lin, and D. Geary. 2023. "What is the Source of the Correlation between Reading and Mathematics Achievement? Two Meta-Analytic Studies." *Educational Psychology Review* **35**, no. 1: 4. doi:10.1007/s10648-023-09717-5.

Wang, J. 2005. "Relationship between Mathematics and Science Achievement at the 8th Grade." *Online Submission* **5**: 1–17.

Yağcı, M. 2022. "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms." *Smart Learning Environments* **9**, no. 1: 11. doi:10.1186/s40561-022-00192-z

Yakman, G., and H. Lee. 2012. "Exploring the Exemplary STEAM Education in the U.S. as a Practical Educational Framework for Korea." *Journal of the Korean Association For Research in Science Education* **32**, no. 6: 1072–1086. doi:10.14697/jkase.2012.32.6.1072.