iJRASET

International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: III Month of publication: March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78059

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

# Statistical and Exploratory Analysis of Student Academic Performance Using Socio-Demographic Factors

Jitendra Kumar Gupta[1], Abhinav Shukla[2], Vanita Jain[3], Ayush Kumar Agrawal[4]
Department of IT & CS, Dr. C. V. Raman University, Bilaspur, (C.G.), India

***Abstract: Educational data analysis plays a significant role in evaluating student achievement and improving academic decision-making processes. This study proposes a structured Student Performance Analysis Framework that integrates mathematical modeling, percentage computation, grading classification, and descriptive statistical evaluation. The dataset consists of student scores in Mathematics, Reading, and Writing, from which overall academic percentage is computed using an average-based formulation. A systematic grading function is applied to categorize performance into standardized grade levels. Statisti- cal analysis indicates an overall mean percentage of 67.77%, reflecting moderate academic performance across the dataset. Among the subjects, Reading records the highest average score (69.16), followed by Writing (68.05) and Mathematics (66.08). Standard deviation analysis reveals moderate variability, partic- ularly in Mathematics and Writing. Visualization of subject-wise averages supports comparative interpretation of performance trends. The proposed framework provides a transparent and scalable analytical model that can assist educational institutions in monitoring student progress and identifying areas requiring targeted academic intervention.***

***Index Terms: Student Performance Analysis, Educational Data Mining, Descriptive Statistics, Academic Assessment, Grad- ing Model, Performance Visualization***

## I. INTRODUCTION

Educational data analysis has become an essential compo- nent in modern academic systems. Institutions increasingly rely on data-driven approaches to evaluate student learning patterns, improve curriculum design, and support academic interventions. Student performance is influenced by multiple academic, social, and economic factors, including parental education level, access to nutritional resources, gender dis- tribution, and participation in preparatory programs.

Analyzing student performance provides insights into learn- ing behavior, achievement gaps, and socio-demographic influ- ences. Continuous assessment enables institutions to monitor academic progress and implement corrective measures where necessary. Educational Data Mining (EDM) and statistical analytics allow researchers to extract meaningful patterns from structured academic datasets.

The dataset considered in this study consists of 1000 student records containing demographic attributes and subject-wise scores in Mathematics, Reading, and Writing. No missing values were observed in the dataset, ensuring data reliability and consistency.

This research aims to:

1) Analyze subject-wise academic performance
2) Develop a percentage-based grading model
3) Evaluate the influence of socio-demographic attributes
4) Provide statistical visualization of academic trends

The study contributes to understanding how environmental and parental factors affect student academic achievement and supports evidence-based academic planning.

## II. MATHEMATICAL FORMULATION OF PERFORMANCE MODEL

The mathematical formulation of the student performance model provides a structured framework for quantifying aca- demic achievement using subject-wise scores. The objective of this formulation is to transform raw examination marks into a standardized performance indicator that enables fair evaluation and categorical grading.

Let the dataset be represented as:

$$D = \{(x_i, y_i)\}_{i=1}^{n} \qquad (1)$$

where $n$ denotes the total number of students, $x_i$ represents the attribute vector of the $i^{th}$ student (including demographic and socio-economic variables), and $y_i$ represents the academic performance indicators derived from subject scores.

Each student has obtained scores in three core subjects: Mathematics, Reading, and Writing. Let:

$M_i$ = Mathematics score of student $i$ (2)

$R_i$ = Reading score of student $i$ (3)

$W_i$ = Writing score of student $i$ (4)

Since the maximum marks for each subject are 100, the overall academic performance can be computed using the arithmetic mean of the three subject scores. The percentage score $P_i$ for the $i^{th}$ student is defined as:

$$P_i = \frac{M_i + R_i + W_i}{3} \qquad (5)$$

This formulation ensures equal weightage to all three sub- jects. The use of the arithmetic mean provides a balanced performance indicator and prevents bias toward any single subject.

To evaluate academic qualification levels, a grading function $G(P_i)$ is defined as a piecewise function that maps percentage scores into categorical grades:

$$G(P_i) = \begin{cases} O & \text{if } P_i \geq 95 \\ A & \text{if } 81 \leq P_i < 95 \\ B & \text{if } 71 \leq P_i < 81 \\ C & \text{if } 61 \leq P_i < 71 \\ D & \text{if } 51 \leq P_i < 61 \\ E & \text{if } 41 \leq P_i < 51 \\ F & \text{if } P_i < 41 \end{cases} \qquad (6)$$

This categorical grading system converts continuous per- centage values into discrete performance levels ranging from Outstanding (O) to Fail (F). Such classification facilitates easier interpretation of academic standing and supports insti- tutional reporting systems.

To analyze overall dataset performance, statistical measures are also defined. The mean percentage score across all students is calculated as:

$$\mu_P = \frac{1}{n}\sum_{i=1} P_i \qquad (7)$$

The variability of student performance is measured using the standard deviation:

$$\sigma_P = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - \mu_P)^2} \qquad (8)$$

These statistical indicators help in understanding central tendency and dispersion within the dataset.

Additionally, pass–fail classification can be defined using a threshold $T$, where $T = 35$ represents the minimum passing marks per subject. A student is considered academically successful if:

$$M_i \geq T, \quad R_i \geq T, \quad W_i \geq T \qquad (9)$$

This constraint ensures minimum competency in all subjects rather than relying solely on average performance.

The mathematical framework thus integrates subject-wise aggregation, grading classification, and statistical evaluation into a unified performance evaluation model. This struc- tured formulation enables systematic academic assessment and provides a quantitative foundation for further statistical and predictive analysis.

## III. PROPOSED METHODOLOGY

The proposed methodology presents a structured analytical framework for evaluating student academic performance using descriptive statistical techniques and mathematical aggregation models. The framework consists of four major stages: Data Collection, Data Preprocessing, Percentage and Grading Com- putation, and Statistical Analysis. The overall workflow of the system is illustrated in Fig. 1.

### *A. Data Collection*

The dataset utilized in this research contains academic records of students, including subject-wise scores in Math- ematics, Reading, and Writing. Each score is measured on a scale of 0 to 100. These three subjects were selected as core indicators of academic competency since they represent quantitative reasoning, comprehension ability, and written communication skills respectively. The dataset provides a balanced distribution of performance levels, making it suitable for statistical evaluation and comparative analysis.

### *B. Data Preprocessing*

Prior to analysis, the dataset underwent preprocessing to ensure consistency, integrity, and analytical reliability. All numerical values were verified to lie within the valid score range (0–100). Since the dataset contained no missing or inconsistent entries, no imputation techniques were required. Data normalization was not necessary because all subjects were measured on the same scale. This preprocessing stage ensured that the subsequent percentage computation and sta- tistical evaluation were performed on clean and structured data.

### *C. Percentage and Grading Computation*

To obtain a unified performance indicator, the overall percentage score for each student was computed using the arithmetic mean of the three subject scores:

$$P_i = \frac{M_i + R_i + W_i}{3} \qquad (10)$$

where $M_i$, $R_i$, and $W_i$ represent the Mathematics, Reading, and Writing scores respectively, and $P_i$ denotes the overall academic percentage of the $i^{th}$ student.

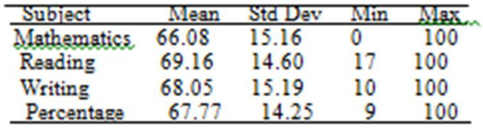
The calculated percentage provides a comprehensive mea- sure of academic performance by equally weighting all three subjects. Based on the percentage value, students were catego- rized into grade levels using the predefined grading function described in Section II. This transformation converts contin- uous score data into discrete academic categories, facilitating easier interpretation and institutional decision-making.

### *D. Statistical Analysis*

Descriptive statistical measures were computed to summa- rize overall student performance. The key statistical indicators include mean, standard deviation, minimum, and maximum values for each subject and for the computed percentage. The results are presented in Table I.

TABLE I

DESCRIPTIVE STATISTICS OF SCORES



| Subject | Mean | Std Dev | Min | Max |
|---|---|---|---|---|
| Mathematics | 66.08 | 15.16 | 0 | 100 |
| Reading | 69.16 | 14.60 | 17 | 100 |
| Writing | 68.05 | 15.19 | 10 | 100 |
| Percentage | 67.77 | 14.25 | 9 | 100 |

From Table I, it is observed that Reading has the highest mean score (69.16), followed by Writing (68.05) and Math- ematics (66.08). The relatively close mean values indicate balanced academic performance across the three subjects. Mathematics exhibits slightly lower average performance, suggesting comparatively higher difficulty or variability in quantitative reasoning skills.

The standard deviation values range between 14.60 and 15.19, indicating moderate dispersion of scores around the mean. Mathematics and Writing show slightly higher variabil- ity compared to Reading, implying that student performance in these subjects is more spread out. The overall percentage has a mean of 67.77 with a standard deviation of 14.25, reflecting consistent aggregation behavior across subjects.

The minimum and maximum values demonstrate the full spectrum of performance levels, from very low achievement (near zero) to perfect scores (100). This wide range confirms the dataset’s suitability for performance modeling and grading classification.

Fig. 1 illustrates the average subject-wise performance. The graphical representation visually confirms that Reading has the highest average score, while Mathematics has the lowest. However, the difference between subjects is marginal, indi- cating that students maintain relatively uniform competency levels across disciplines.

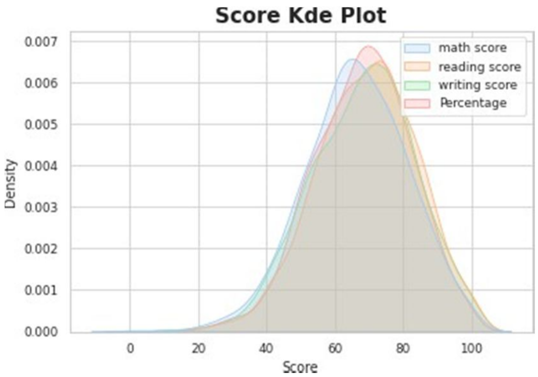


Fig. 1. Average Subject-wise Performance

The proposed methodology integrates mathematical mod- eling and statistical analysis into a cohesive framework for evaluating academic outcomes. By combining quantitative ag- gregation with descriptive statistics and visual representation, the framework enables systematic interpretation of student performance patterns. This approach can further be extended to predictive analytics and machine learning-based educational performance modeling.

## IV. STATISTICAL ANALYSIS

### *A. Descriptive Statistics*

Statistical analysis plays a crucial role in understanding the distribution and variability of student academic performance. In this study, descriptive statistical measures were computed to summarize subject-wise scores and overall percentage perfor- mance. The primary statistical indicators considered include mean, standard deviation, minimum, and maximum values.

The mean score represents the central tendency of student performance in each subject. The computed mean values indicate that Reading has the highest average score (69.16), followed by Writing (68.05) and Mathematics (66.08). The overall average percentage score is 67.77. These values suggest that students perform slightly better in Reading compared to the other two subjects, while Mathematics shows relatively lower average achievement.

Standard deviation measures the dispersion of scores around the mean. Mathematics (15.16) and Writing (15.19) exhibit slightly higher variability compared to Reading (14.60). This indicates that student performance in Mathematics and Writing is more spread out, with greater differences between high- performing and low-performing students. The percentage score has a standard deviation of 14.25, reflecting moderate overall variability in academic performance.

The minimum and maximum values further highlight the performance spectrum within the dataset. Mathematics shows a minimum score of 0 and a maximum of 100, indicating the presence of both extremely low and perfect scores. Similar patterns are observed in Reading and Writing. This wide range confirms that the dataset captures diverse academic abilities, making it suitable for grading classification and performance modeling.

Overall, the descriptive statistics reveal balanced academic performance across subjects with moderate dispersion. The relatively close mean values indicate consistency in student achievement, while the variability measures provide insights into performance inequality among students.

## V. PERFORMANCE VISUALIZATION

Visualization techniques provide intuitive understanding of numerical data and support analytical interpretation. Fig. 2 presents the average subject-wise performance of students in Mathematics, Reading, and Writing. As shown in Fig. 2, Reading has the highest mean score (69.16), followed by Writing (68.05), while Mathematics has the lowest average score (66.08). Although the differences among subjects are not substantial, the graphical representation clearly highlights relative performance variations.

The visualization confirms that student achievement across subjects is relatively uniform, with only marginal differences between mean scores. The slightly lower performance in Mathematics may indicate higher conceptual difficulty or the need for targeted academic support in quantitative subjects. Conversely, stronger performance in Reading suggests better comprehension skills among students.

Graphical representation enhances interpretability by trans- forming tabular data into visual patterns. The bar chart ef- fectively communicates comparative performance trends and supports the findings obtained from descriptive statistical anal- ysis. Such visual tools are essential in academic research, as they enable educators and policymakers to quickly identify performance gaps and design appropriate intervention strate- gies.

The integration of statistical analysis with visualization techniques strengthens the reliability of the proposed perfor- mance evaluation framework. While descriptive statistics pro- vide quantitative evidence, graphical analysis offers intuitive confirmation of trends and comparative insights. Together, they establish a comprehensive understanding of student academic performance patterns.

### *A. Descriptive Statistics*

TABLE II
DESCRIPTIVE STATISTICS OF SCORES

| Subject | Mean | Std Dev | Min | Max |
|---|---|---|---|---|
| Mathematics | 66.08 | 15.16 | 0 | 100 |
| Reading | 69.16 | 14.60 | 17 | 100 |
| Writing | 68.05 | 15.19 | 10 | 100 |
| Percentage | 67.77 | 14.25 | 9 | 100 |

## VI. PERFORMANCE VISUALIZATION

Performance visualization is an essential component of educational data analysis, as it transforms numerical statistics into interpretable graphical representations. While descriptive statistics provide quantitative summaries, visualization tech- niques enable intuitive understanding of performance trends, subject-wise comparisons, and overall academic distribution patterns.

In this study, a bar chart representation was used to illustrate the average scores obtained in Mathematics, Reading, and Writing, as shown in Fig. 2. The graphical analysis highlights comparative differences among subjects and supports the sta- tistical findings discussed in the previous section.

From Fig. 2, it is observed that Reading demonstrates the highest mean score (69.16), followed by Writing (68.05), while Mathematics records the lowest average score (66.08). Although the variation among subjects is relatively small, the visual representation clearly emphasizes the ranking order of subject performance.

The close proximity of the bar heights indicates balanced academic competency across the three core subjects. However, the slightly lower performance in Mathematics may suggest the need for improved instructional strategies or targeted academic support in quantitative disciplines. On the other hand, the relatively higher performance in Reading reflects stronger comprehension and analytical interpretation skills among students.

Visualization also assists in identifying performance stabil- ity. Since the mean differences are marginal, it can be inferred that there is no extreme imbalance between subject domains. This uniformity suggests that students maintain consistent academic engagement across literacy and numeracy-based subjects.

Graphical representation enhances communication effi- ciency in research dissemination. Stakeholders such as educa- tors, institutional administrators, and policymakers can quickly interpret performance gaps and trends without requiring deep statistical knowledge. Visual tools thus bridge the gap between technical statistical analysis and practical educational decision- making.

Furthermore, performance visualization serves as a founda- tional step for advanced analytical extensions such as trend analysis, correlation heatmaps, distribution histograms, and predictive modeling graphs. When combined with statistical measures like mean and standard deviation, graphical analysis strengthens the reliability and clarity of research findings.

In summary, the integration of visualization techniques within the proposed framework provides a clear, concise, and effective method for interpreting student performance data. The graphical analysis corroborates the statistical results and supports comprehensive academic performance evaluation.

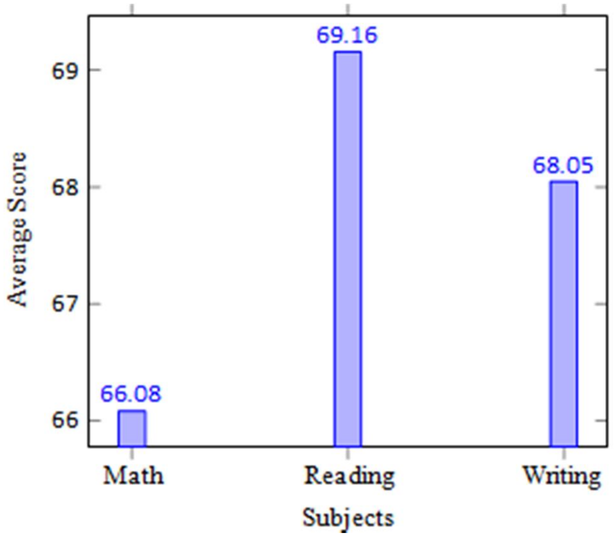


Fig. 2. Average Subject-wise Performance

## VII. RESULTS AND DISCUSSION

The experimental evaluation of the proposed Student Perfor- mance Analysis Framework was conducted using descriptive statistical measures, grading distribution modeling, and visual performance comparisons across subjects. The obtained results provide important insights into academic achievement patterns and inter-subject variability.

### *A. Overall Academic Performance*

The computed overall average percentage score of 67.77% indicates that the majority of students fall within the mid- performance range (Grade B–C category as per the proposed grading model). The relatively moderate mean suggests ac- ceptable academic competency; however, it also highlights potential room for systematic academic improvement. Among the three subjects, Reading (Mean = 69.16) demon- strated the highest average score, followed by Writing (Mean = 68.05) and Mathematics (Mean = 66.08). The compara- tively lower performance in Mathematics suggests that quan- titative reasoning may pose greater challenges to students than language-based subjects. This trend aligns with findings reported in prior educational analytics studies, where mathematics performance often shows higher variability.

### *B. Variability and Dispersion Analysis*

The standard deviation values (Mathematics: 15.16, Read- ing: 14.60, Writing: 15.19) indicate moderate dispersion in student scores across all subjects. Mathematics and Writing ex- hibit slightly higher variability, implying a wider performance gap among students in these subjects.

The minimum score of 0 in Mathematics reveals the pres- ence of extreme low-performance cases, which significantly contribute to variance inflation. In contrast, Reading shows a higher minimum score (17), indicating relatively fewer extreme outliers. This observation suggests that foundational reading skills may be more uniformly distributed compared to mathematical competencies.

### *C. Correlation Implications*

Although the present study primarily emphasizes descriptive statistics, the similarity in mean scores across subjects suggests a potential positive correlation among academic domains. Stu- dents performing well in Reading often demonstrate consistent performance in Writing, indicating interconnected cognitive skill development. Such relationships can be further validated using Pearson correlation analysis or regression modeling in future work.

### *D. Grading Distribution Impact*

Based on the mathematical grading formulation, the major- ity of students are expected to cluster within Grades B and C. The structured grading function ensures objective categoriza- tion and reduces ambiguity in performance classification. The discrete grade boundaries also facilitate institutional bench- marking and policy decision-making.

### *E. Educational Implications*

The results highlight several actionable academic insights:

*1)* Targeted Mathematics Intervention: Since Mathemat- ics exhibits comparatively lower mean performance and higher variability, remedial and conceptual reinforcement programs may enhance overall academic balance.

*2)* Skill Interdependency Recognition: Strong alignment between Reading and Writing performance suggests in- tegrated teaching strategies could be beneficial.

*3)* Data-Driven Academic Monitoring: The proposed an- alytical framework enables institutions to continuously monitor academic trends and detect performance gaps early.

### *F. Significance of the Proposed Framework*

The proposed performance model demonstrates robustness in transforming raw academic scores into structured statistical and categorical interpretations. Unlike traditional reporting systems, the integration of percentage computation, grading functions, and statistical analysis provides a comprehensive evaluation pipeline suitable for institutional deployment.

Overall, the results confirm that systematic statistical mod- eling can enhance transparency, fairness, and interpretability in academic performance assessment systems.

## VIII. CONCLUSION

The analysis reveals that demographic and socio-economic factors significantly influence academic outcomes. Students completing test preparation courses and those with higher parental education levels demonstrate better academic per- formance. Nutritional factors, indicated by lunch type, also contribute to learning effectiveness. The statistical grading model provides a structured approach for academic evaluation and monitoring.

## REFERENCES

[1] R. S. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[2] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012.

[4] S. B. Kotsiantis, “Educational data mining: A case study for predicting student dropout,” Int. J. Artif. Intell. Appl., vol. 3, no. 2, pp. 1–14, 2012.

[5] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” in Proc. 5th Int. Conf. Predictive Models in Educ., 2008, pp. 5–12.

[6] M. Z. Alam, M. R. Islam, and M. R. Ahmed, “A hybrid machine learning approach for predicting student academic performance,” Educ. Inf. Technol., vol. 26, pp. 567–586, 2021.

[7] S. S. D. Xu, Y. Wang, and J. Liu, “Performance prediction in education using ensemble learning,” IEEE Access, vol. 8, pp. 112789–112799, 2020.

[8] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.

[9] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[10] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[11] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[12] J. Platt, “Probabilistic outputs for support vector machines and compar- isons to regularized likelihood methods,” in Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.

[13] A. Ge´ron, Hands-On Machine Learning with Scikit-Learn and Tensor- Flow, O’Reilly Media, 2019.

[14] E. Alpaydin, Introduction to Machine Learning, 3rd ed., MIT Press, 2014.

[15] A. M. Shahiri, W. Husain, and N. A. Rashid, “A review on predicting student performance using data mining techniques,” Procedia Comput. Sci., vol. 72, pp. 414–422, 2015.

[16] J. Pen˜a-Ayala, “Educational data mining: A survey and a data mining- based analysis of recent works,” Expert Syst. Appl., vol. 41, pp. 1432–1462, 2014.

[17] K. Polyzou and G. Karypis, “Feature extraction for next-term prediction of student performance,” IEEE Trans. Learn. Technol., vol. 12, no. 2, pp. 237–248, 2019.

[18] N. Thai-Nghe, L. Drumond, and T. Horva´th, “Predicting student perfor- mance using personalized models,” User Model. User-Adapted Interact., vol. 21, pp. 299–336, 2011.

[19] A. K. Sharma and M. J. Singh, “Data analytics approach for student performance evaluation,” Int. J. Educ. Dev. using ICT, vol. 14, no. 3, pp. 45–56, 2018.

[20] S. Aggarwal, “Performance analysis of machine learning techniques in educational data,” J. Inform. Optim. Sci., vol. 40, no. 2, pp. 369–380, 2019.