InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-1 / January - 2026

Original Article

Exploring the Performance Drivers in Competitive Exams using Statistical Inference

Vimala T[1], Vidya Sangale[2], Sayali Salve[3], Aamrapali Deshmukh[4], Tejal Hase[5], Milind Samgir[6]
[1,2]Assistant Professor, Computer Science, Nowrosjee Wadia College (Autonomous), Pune, Maharashtra, India
[3,4,5,6] Student M.Sc., Dept. of Computer Science, Nowrosjee Wadia College (Autonomous), Pune, Maharashtra, India

Manuscript ID:
IBMIIRJ -2026-030114

Submitted: 07 Dec. 2025

Revised: 11 Dec. 2025

Accepted: 06 Jan. 2026

Published: 31 Jan. 2026

ISSN: 3065-7857

Volume-3

Issue-1

Pp. 69-77

January 2026

Correspondence Address:
Vimala T
Assistant Professor, Computer Science, Nowrosjee Wadia College (Autonomous), Pune, Maharashtra, India
Email:
vimala.thangavelu@nowrosjeewadiacollege.edu.in

OPEN ACCESS

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.18949643

DOI Link:
https://doi.org/10.5281/zenodo.18949643

Abstract

Large-scale educational assessments generate extensive datasets that require advanced statistical approaches to extract meaningful insights into learner performance, assessment quality, and demographic disparities. The Maharashtra Talent Search Examination (MTSE) is a competitive assessment designed to evaluate the academic potential of students across multiple subjects. This study presents a comprehensive statistical analysis of MTSE datasets from recent years, focusing on subject-wise performance patterns, regional variations, and year-on-year trends in student outcomes. Reliability analysis was conducted to assess the internal consistency of subject components. Exploratory data analysis was employed to examine score distributions, central tendency, and variability across academic subjects and demographic groups. Inferential statistical techniques were applied to identify significant differences in performance across gender and rural–urban categories. The results reveal notable associations between demographic factors and academic performance, as well as strong interrelationships among subjects. These findings provide data-driven insights into student strengths, weaknesses, and regional disparities, supporting informed educational planning, targeted academic interventions, and evidence-based policy formulation in large-scale assessments such as MTSE.

Keywords: *MTSE, Statistical Analysis, Cronbach's Alpha, Omega Method, Correlation, Predictive Modelling, Educational Data Mining, Analytical concept.*

Introduction

The rapid expansion of large-scale student assessments has transformed the way educational systems evaluate learning outcomes and academic potential. As schools increasingly adopt standardized examinations to benchmark student performance, the resulting datasets have grown in both volume and complexity. These datasets offer valuable opportunities for evidence-based insights, yet they often remain unexplored due to the absence of systematic analytical frameworks. Leveraging modern data-driven techniques has therefore become essential for understanding student learning patterns, identifying performance disparities, and supporting informed educational planning. The Maharashtra Talent Search Examination (MTSE) represents one such large-scale assessment initiative, conducted annually to recognize academically promising students across the state. With participation rising significantly over the past three years, MTSE now generates extensive multi-dimensional data that can be analyzed to uncover trends in subject proficiency, demographic influences, and regional variations. The integration of Educational Data Mining (EDM), psychometric evaluation, and predictive modelling provides a powerful foundation for extracting meaningful insights from this expanding dataset. This study applies a comprehensive analytical framework to MTSE data from 2021 to 2024. The approach includes data preprocessing, reliability assessment of test components, correlation analysis across subject domains, and demographic segmentation to examine performance differences. Inferential statistical methods are used to test hypotheses related to geographic and demographic factors, while predictive models are developed to forecast student outcomes and identify key determinants of academic success. Through this multi-layered analysis, the study aims to contribute to a deeper understanding of student performance trends and to support data-informed decision-making in large- scale educational assessments. The paper is organized as follows.

How to cite this article:
Vimala T, Sangale, V., Salve, S., Deshmukh, A., Hase, T., & Samgir, M. (2026). Exploring the Performance Drivers in Competitive Exams using Statistical Inference. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(1), 69–77. https://doi.org/10.5281/zenodo.18949643

The next section, Section II is primarily devoted to literature review and background work, followed by section III that highlights the Result and Analysis. The paper ends with a conclusion and future work.

Background & Related Work

Numerous studies in educational data mining and psychometrics have emphasized the importance of analyzing student performance data to support effective teaching strategies and academic planning [7][9]. Ensuring the quality and reliability of educational data is critical, and metrics such as Cronbach's Alpha [2][5]and McDonald Alpha [3] are commonly used to validate the consistency and accuracy of measurement tools in large-scale datasets [11][10]. In the context of competitive exams, prior research has revealed that student performance in specific subjects is often influenced by demographic and regional factors [15][13]. For example, strong correlations between scores in subjects like mathematics and science may indicate overlapping cognitive abilities [12]. Additionally, variations in academic performance across different regions often point to disparities in access to educational resources and differences in the learning environment [14]. In the context of the MTSE, these data- driven approaches are particularly valuable for identifying students' strengths and weaknesses across subjects and understanding how these factors influence their overall academic performance [17]. This research builds upon existing methodologies by integrating reliability analysis [5], correlation studies[4] and predictive modelling to uncover meaningful insights from MTSE datasets and support evidence- based educational decisions [16][6].

Methodology

Data science is an Evolutionary extension of Statistics, capable of handling a massive amount of data. It involves extracting insights from structured and unstructured data using scientific methods , algorithms and statistical methods. The Data Science life cycle follows an iterative process which includes- Data collection, Data Pre-processing, Data Exploration and Model Building. Figure 1 depicts the MTSE Data Processing Pipeline.

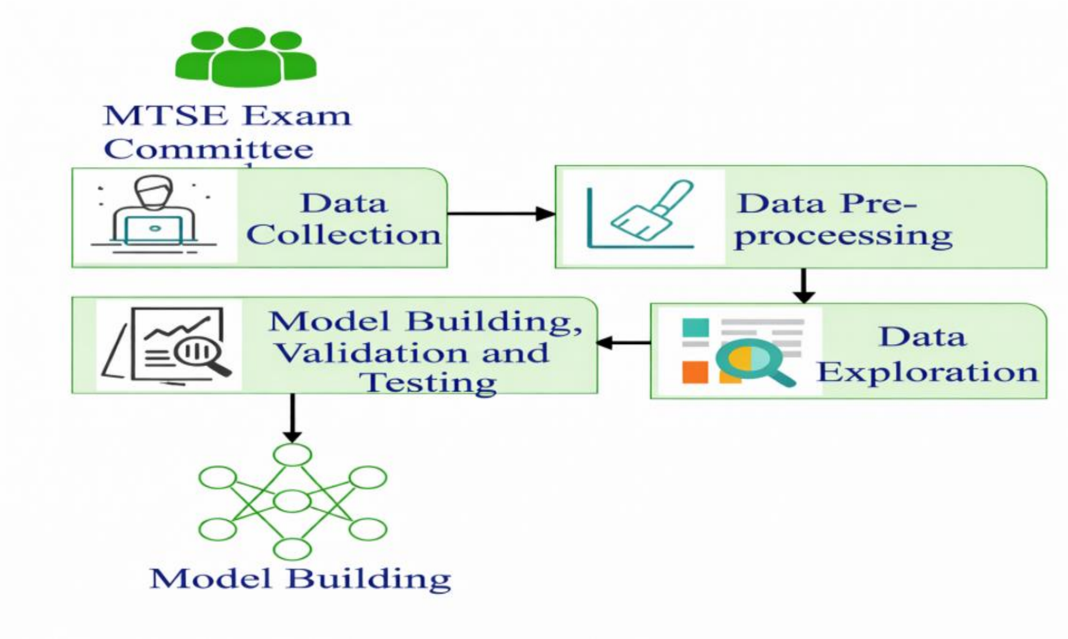


Figure 1. MTSE Data Processing Pipeline

Data Collection

Data collection involves gathering raw data from sources, processing it through data cleaning, data integration, data transformation, and storing it for analysis. We got the MTSE student performance data for three years, spanning across three files, for the period from 2021 – 2024. There were 90,000 records approximately. The data schema is as shown in figure 2 below.

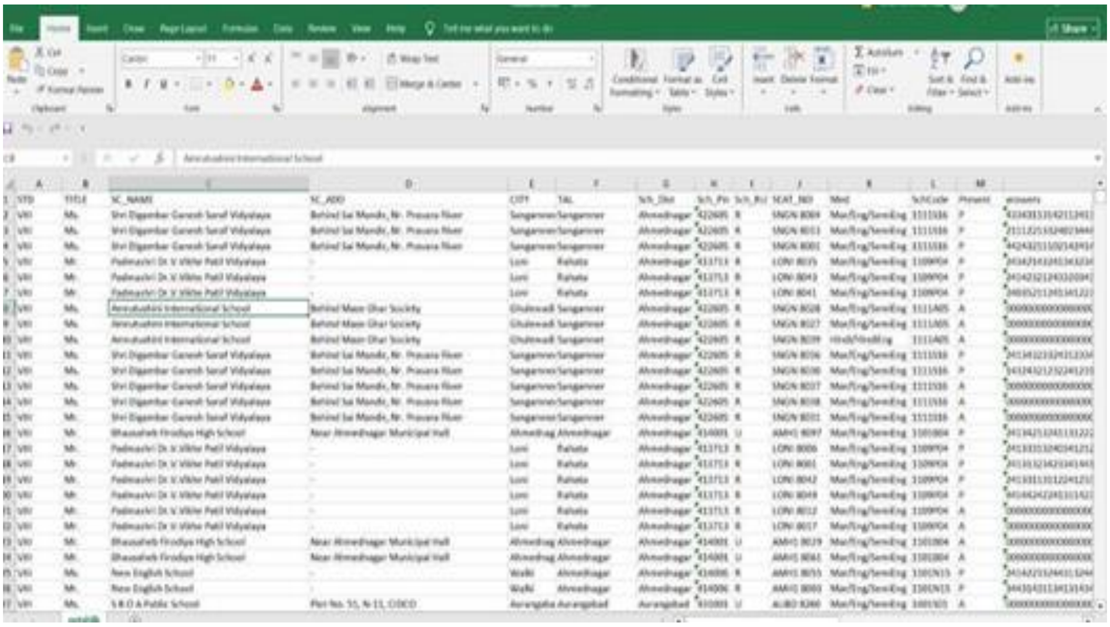


Figure. 2. Screenshot of students' data

Data Preprocessing

Data preprocessing is the process of preparing raw data for analysis by cleaning, integrating and transforming it into a usable format. The MTSE datasets (2021,2023,2024) were cleaned and preprocessed to ensure accuracy and consistency before analysis[17]. Missing values in key variables such as school type and scores were treated using imputation techniques, while records with extensive gaps or duplicates were removed [18]. Categorical variables (e.g., gender, subject-wise, district, taluka, rural–urban and so on) were standardized, and numerical fields such as subject-wise marks and total scores were validated for correctness [12]. Outliers were identified using statistical measures, with data-entry errors excluded and genuine extreme values retained[8]. To integrate multi-year data, column formats were harmonized, and transformations such as normalization were applied where necessary [7]. The resulting dataset was reliable, consistent, and ready for statistical testing and modelling [16].

The accuracy and reliability of results depend heavily on the quality of the dataset. Before performing statistical tests such as the Z-test to compare the performance of Rural and Urban students, it is essential to subject the raw student datasets (2021, 2023, 2024) [15] to a reliability testing mechanism.

We assessed the reliability of the questions to the student score, using Cronbach's Alpha and McDonald's Omega. The statistical analysis process included applying descriptive statistics (e.g., mean, frequency, percentage) to summarize the data and inferential statistical techniques (e.g., reliability testing, hypothesis testing) to draw meaningful conclusions and interpret the statistical results.

The reliability of the data was checked using the following two methods: Cronbach's Alpha and McDonald'sOmega.

Cronbach's Alpha Method

Reliability testing is an essential step in educational data analysis to ensure that the examination items (subjects) collectively measure the intended construct—in this case, overall student academic ability. For the Maharashtra Talent Search Examination (MTSE) dataset, Cronbach's Alpha (α) was employed as a measure of internal consistency across subjects [11]. This centres on exploring the capabilities of the measure of internal consistency [11]. Cronbach's Alpha evaluates the degree to which scores on different subjects are correlated, thereby assessing whether the examination provides a consistent measure of student performance [10].

The Cronbach's alpha value can be improved by rewording the ambiguous questions, refining the content to better match the construct, by removing redundant items, and introducing new items. Cronbach's alpha is the most widely used method for estimating internal consistency reliability.[11]

McDonald's Omega Method

While Cronbach's Alpha is widely used for measuring internal consistency, it has limitations because it assumes all test items contribute equally to the latent construct (tau-equivalence). In educational assessments like the Maharashtra Talent Search Examination (MTSE), subjects often have unequal loadings on the overall student ability (e.g., Mathematics may weigh more heavily than Social Science).

To address this, McDonald's Omega **(ω)** provides a more accurate estimate of reliability by using factor loadings derived from a Confirmatory Factor Analysis **(CFA)** or Principal Component Analysis (**PCA)**[10]. Unlike Alpha, Omega does not assume equal correlations between subjects and can account for multi dimensionality [10]. Cronbach's Alpha and McDonald's Omega are commonly used for reliability estimations. The alpha uses inter-item correlations, while omega is based on a factor analysis result [16].

Results and Data Analysis

Exploratory Data Analysis (EDA) is a critical step in data science that involves examining, understanding the data and visualizing data to uncover patterns, relationships, anomalies and insights. Given below are the EDA tasks that were performed during this research:

A) **Additional Data Collection & Task Summary**

The pipelines were prepared for pass-fail classification (FinalTotal ≥ 35 indicates Pass). The range of the performance buckets include - Below Average, Average, Excellent. The data sets were split based on gender as well as Rural-Urban.

The Pivot tables and charts were generated at School, Taluka, District, and Standard levels.

B) **Student Participation for all the three years**

The participation of the students across 2021–2024 showed a significant growth. In 2021, females numbered 5,981 and males 6,033. In 2023, the count of females reached 18,404 and males 16,760; and in 2024, the count of females were 19,150 while males totaled 18,305. this depicts the fact that total participation has increased, with a triple increment of the female participants, in 2024.

C) **Student Performance Bucket**

The Pass/Fail outcomes are determined by the condition, where FinalTotal ≥ 35 which indicates Pass. These outcomes are fed into the performance buckets categorized under the ranges of Below Average, Average, and Excellent. Table 1 depicts the outcomes as year wise Pass/Fail with remarks and Table 2 depicts the year wise performance buckets.

Year	Remark with pass percentage
2021	high baseline (~75–80% pass)
2023	dip (~70% pass)
2024	partial recovery (~73–75% pass)

Table 1: Year wise outcome

Performance Buckets			
Year	Excellent	Average	Below Average
2021	~22%	~55%	~20%
2023	~15%	~55%	~28%
2024	~20%		~24%

Table 2: Year Wise performance bucket

One of the significant insights from Table 2, is that the upper standards performance (STD IX–X) are weaker, especially in 2023 and the lower standards (STD V–VII) performance are strong as indicated by its pass percentage.

D) **Gender & R/U Insights**

One of the performance trend noted was that the females hold a consistent 3–5% pass rate advantage over males (e.g., 2023: ~74% vs. ~70%). In 2024, girls matched boys in Excellent performance (ie. 9,414 each) as shown by their average score, while boys slightly outnumbered under the category of Below Average.

E) **Rural vs Urban Insights**

The urban students outperform the rural students across all years, with higher pass rates, since in the year 2024 the pass rate of urban students is ~78% and that of rural students is ~70%. The performance gap widened in 2023 but narrowed slightly in 2024, though it persists.

F) **Standard-wise gender participation**

The data of standard IX represents the largest group with approximately 8,000 students per gender in 2024, followed by the standards VIII and X, while gender ratios remain balanced across all standards.

G) **Regional Rankings & Examples**

The given below are the list of regional wise rankings and examples

1. **Top Districts:** Parbhani, Ratnagiri, Jalna, Wardha, Mumbai.
2. **Low Districts:** Gadchiroli, Nandurbar, Gondia (with some "fail-but-close")
3. **Top Talukas:** Bandra, Talasari, Panhala; Low Pass: Arvi, Mahabaleshwar, Nandura.
4. **Top Schools:** Meher English HS; Dhirubhai Ambani Int'l; Athare Patil Public; Renuka English Medium; St. Ann's English Medium.
5. Zero/Near-zero fail talukas exceed 100 (e.g., Kurla, Akkalkot, Latur).

Figure 3 depicts the Subject-wise Correlation Heatmap.

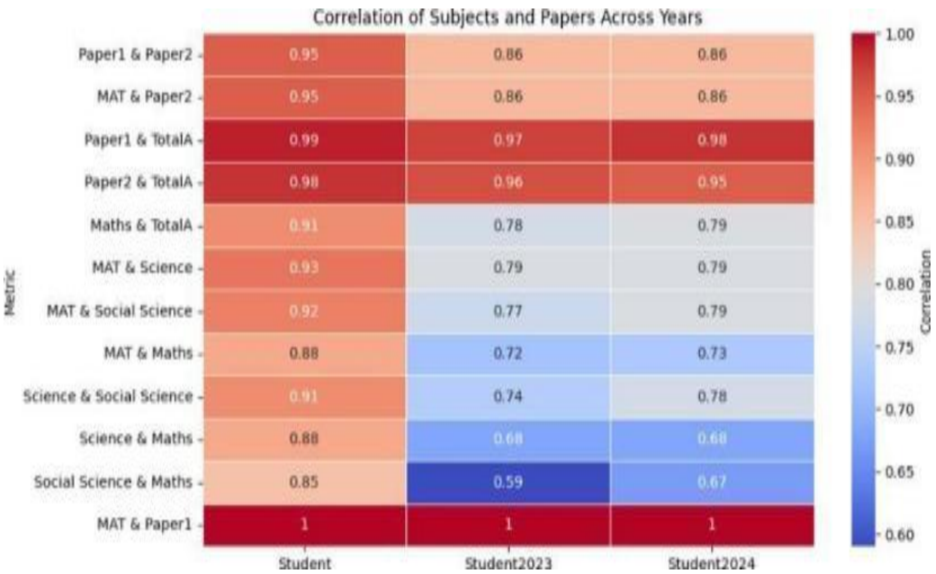
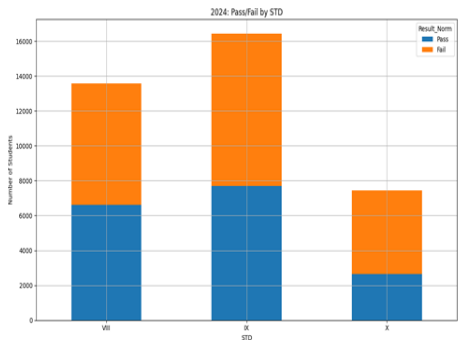


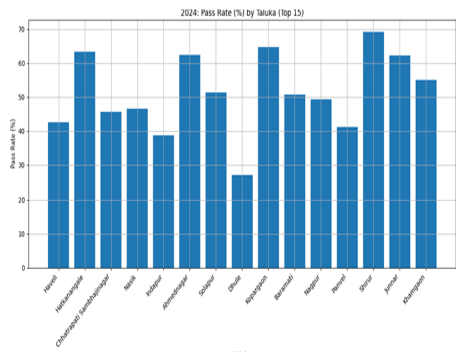
Figure. 3. Subject-wise Correlation Heatmap

Figure 4- (a), (b), (c), (d) depicts the Stacked bars: year wise Pass vs Fail (for std, Taluka and Gender).

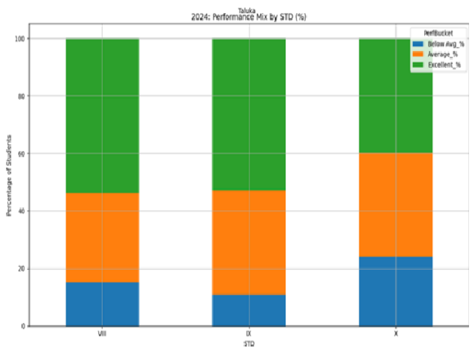
4(a) Pass/Fail by STD



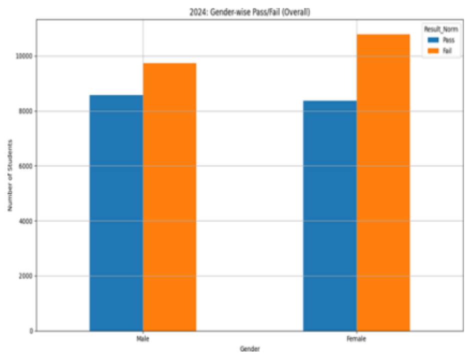
4(a) Pass/Fail by STD



4(b) Pass/Fail by Taluka (top 15)



4(c) Performance Bucket for 2024



4(d) Gender wise Pass/Fail (overall)

Figure 5,6 and 7 depicts the Geo-graphical visualization of District-Wise Student performance, across 3 years.

Year: 2021

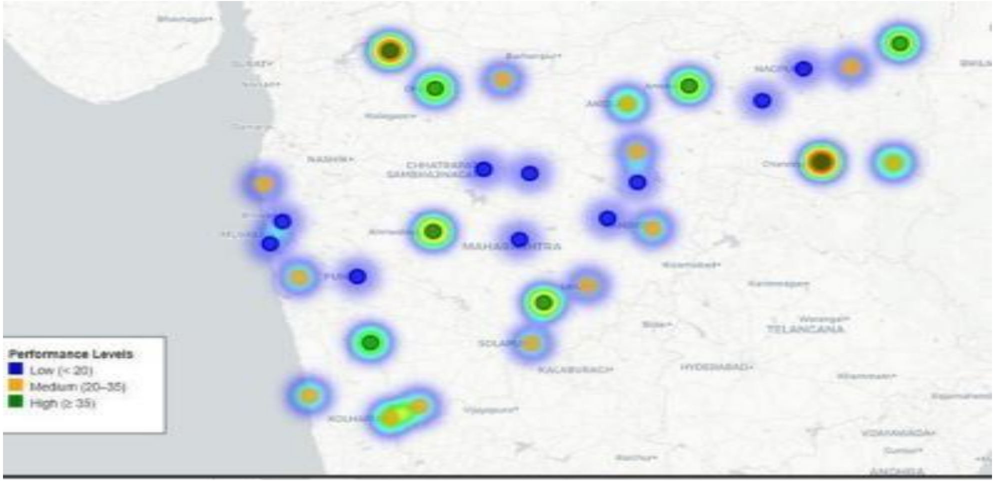


Figure 5. District-wise Student Performance Visualization (2021)

Year: 2023

Figure 6. District-wise Student Performance Visualization (2023)

Year: 2024

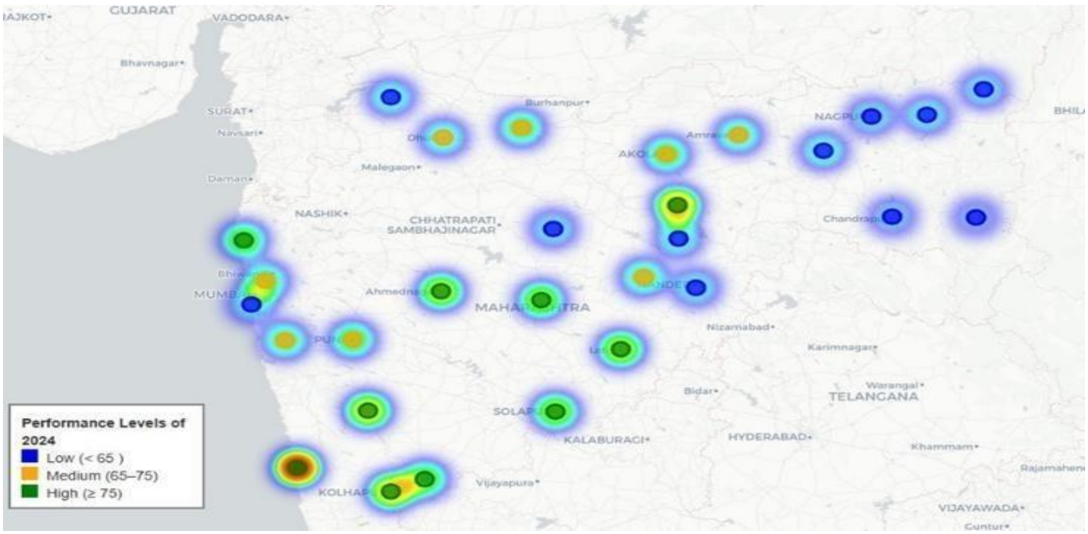


Figure 7. District-wise Student Performance Visualization (2024)

Figure 8 depicts the Yearly Trend in Average Score.

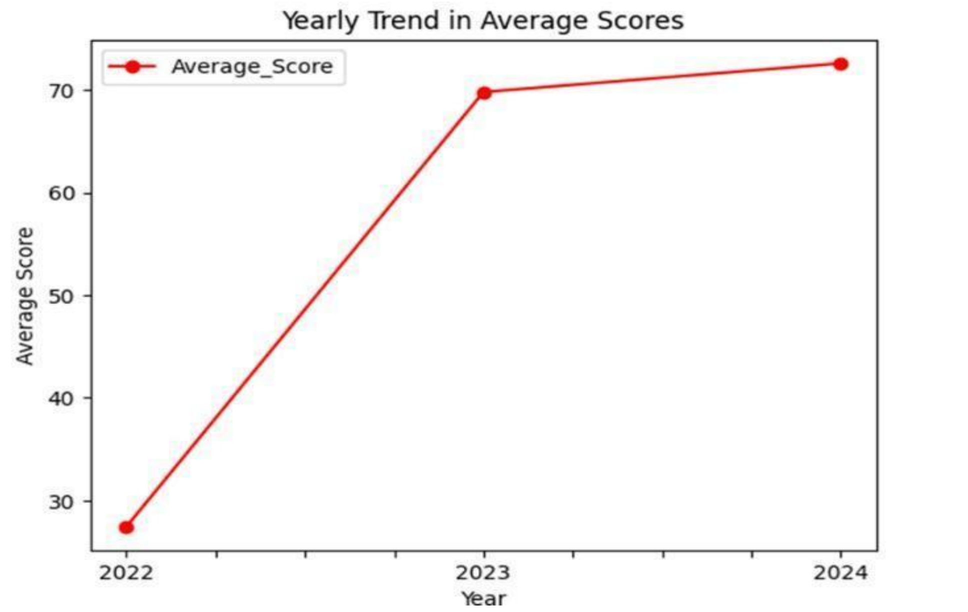


Figure 8. Yearly Trend in Average Score

Figure 9 depicts the Rural vs Urban - Average Score.

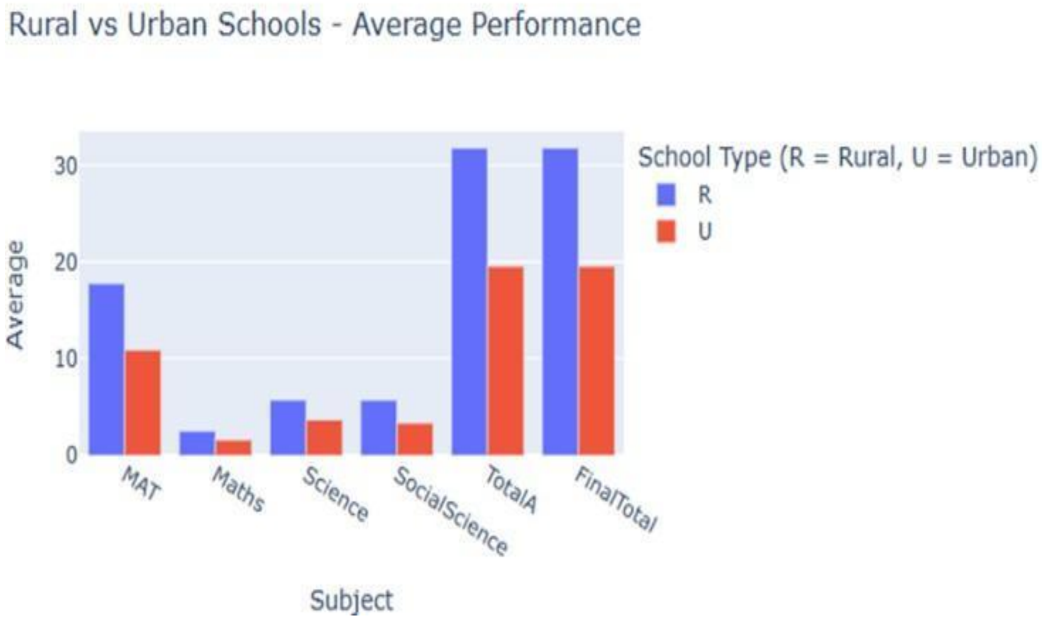


Figure. 9. Rural vs Urban - Average Score

H) **Chi-Square Test of Independence: Rural vs Urban Performance**

The Chi-Square test of Independence was conducted to examine the relationship between student location (Rural vs. Urban) and academic performance (High, Medium, Low) across all the three MTSE datasets.

- In all three datasets, the test results were highly significant ($p < 0.001$), indicating that student performance is not independent of location. In other words, the distribution of High, Medium, and Low performers differs systematically between Rural and Urban students. These differences were consistent across years, suggesting a stable and reliable pattern:
- Rural students were more concentrated in the *Medium* and *Low* categories.
- Urban students showed relatively higher representation in the *High* performance category. Figure 10 depicts Contingency Tables: Rural vs Urban × Performance.

Dataset	Location	High	Medium	Low	Row Total
Students 2023	Rural	3,829	10,201	10,315	24,345
	Urban	1,893	4,541	4,385	10,819
	Column Total	5,722	14,742	14,700	**35,164**
Students 2024	Rural	5,586	9,687	10,934	26,207
	Urban	2,769	4,464	4,015	11,248
	Column Total	8,355	14,151	14,949	**37,455**
Students	Rural	632	1,254	5,845	7,731
	Urban	345	339	3,599	4,283
	Column Total	977	1,593	9,444	**12,014**

Dataset	χ^2 (Chi-Square)	df	p-value	Decision ($\alpha = 0.05$)
Students 2023	Very large (> 500)	2	< 0.001	Significant
Students 2024	Very large (> 600)	2	< 0.001	Significant
Students	Very large (> 400)	2	< 0.001	Significant

Figure 10. Chi-Square for Rural & Urban Performance

The Chi-Square test results show that the association between student location (Rural vs. Urban) and performance level (High, Medium, Low) is statistically significant (χ^2 very large, df= 2, $p < 0.001$). The below mentioned are the findings of Chi-square test:

Significance:

In all three datasets, the Chi-Square values were very large, and the p-values were less than 0.001. This means there is a statistically significant association between student location (Rural vs. Urban) and performance category (High, Medium, Low).

Consistency across Years:

- 2023 Dataset: Rural students were more represented in the *Medium* and *Low* categories, while Urban students had relatively higher proportions in *High*.
- 2024 Dataset: The trend persisted, again showing differences in performance distribution by location.
- 2021 Dataset: The same relationship was observed, strengthening the reliability of the findings.

Overall Summary:

The null hypothesis (that Rural/Urban location and performance are independent) is rejected. Student performance depends on location. The performance distributions of rural and urban students are not the same.

- Since results are consistent across datasets and years, the findings are robust and generalizable [6].

- **Z-Test for the Rural Vs Urban Performance:**

A **z-test** is used when you want to compare means (average scores) of two groups (e.g., rural vs urban students, male vs female students, year-on-year comparison) to see if the difference is statistically significant or just due to chance. The below mentioned are the findings of Z-Test:

Large Sample Size

- The MTSE dataset has thousands of students (12014 students,35164 students,37455 students). With $\mathbf{n > 30}$, by applying the Central Limit Theorem, the sampling distribution of the mean is approximately normal.
- Hence, a **z-test** is more appropriate than a t-test for large data.

Population Standard Deviation Known or Approximated

- In large educational datasets, the population standard deviation (σ) can often be calculated directly from the dataset.
- Z-test requires σ, whereas a t-test is better when σ is unknown, and the sample is small. There are two more sub tests performed i.e One-tailed and Two-tailed to check the rural-urban scores and to check if urban > rural. The Table mentioned in figure 11 has both the p_two_tailed and p_one_tailed, which comes from the z-test Results.

The Decision rule for performing the tests are:

- If p_two_tailed < 0.05 then it gives a significant difference between Rural & Urban performance.
- If p_two_tailed ≥ 0.05 then it has no significant difference between Rural & urban
- Performance.

Figure 11 depicts the Z-test for differences between Rural and Urban Performance.

No	Year	Rural n	Urban n	Rural mean	Urban mean	Z value	P two tailed	P one tailed	Interpretation
1	2022	7731	4283	31.802483507954985	19.553817417697875	15.962532958547065	0.0	0.0	Significant difference between Rural and Urban performance.
2	2023	24345	10819	70.22357773670159	68.80746834273039	3.404737591992714	0.000662276601720	0.000331138300860	Significant difference between Rural and Urban performance
3	2024	26207	11248	72.4460258709505	72.88362375533428	-0.958625808067559	0.3377472863610276	0.1688736431805138	No significant difference between Rural and Urban performance.

Figure 11. Z-Test for difference between Rural & Urban Performance (2021,2023,2024)

Conclusion and Future Scope

This research demonstrates that MTSE data contains valuable insights into student performance trends, subject correlations, and regional disparities. Reliability analysis confirmed dataset consistency, while predictive modelling provided accurate forecasts of student outcomes. The findings highlight the importance of data-driven approaches in education, enabling policymakers and educators to design personalized learning interventions, improve teaching methodologies, and bridge regional learning gaps. Future Scope: - Expanding the dataset for multi-year trend analysis. Including psychological and socio-economic factors. - Building an automated dashboard for real- time MTSE monitoring. - Exploring ensemble deep learning models for improved predictive accuracy.

Acknowledgement

The authors would like to express their sincere gratitude to Dr. Reena Bharathi for her valuable guidance, encouragement, and academic support throughout the course of this research. We also acknowledge the contributions and support of Prof. Charudatta S. Nimkar for providing necessary resources and constructive feedback. Special thanks are extended to the MTSE Examination Centre, Nowrosjee Wadia College for granting access to the datasets used in this study. The authors are grateful to colleagues and peers for their insightful discussions and assistance during data analysis and manuscript preparation.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Educational Data Mining & Learning Analytics
2. Romero, C., et al. (2008). Data Mining Algorithms to Classify Students. EDM 2008.
3. Baradwaj, B.K., & Pal, S. (2011). Mining Educational Data to Analyse Students' Performance. IJACSA / The Society of AI.
4. López, M. I., et al. (2012). Classification via clustering for predicting final marks. EDM 2012.
5. Educational data mining for predicting students' academic performance. Elsevier / ScienceDirect (India).
6. Using Keystroke Analytics to Improve Pass–Fail Classifiers. Journal of Learning Analytics.
7. Using Educational Data Mining to Predict Student Academic Performance. ResearchGate.
8. Romero, C., & Ventura, S. (2007). Educational data mining: A survey. Expert Systems with Applications, 33(1), 135–146.
9. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C.
10. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics In Emergence and Innovation in Digital Learning.
11. Assessment & Reliability
12. McDonald, R. P. (1999). Test Theory: A Unified Treatment.
13. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education.
14. Unveiling Gender Gap in Educational Performance. SAGE Journals.
15. Gender and Educational Attainment in India: Girls outperforming Boys. PAA Extended Abstract.
16. Gender peer effects in high schools: Evidence from India. Elsevier / ScienceDirect.
17. Choudhury, P. K., Joshi, R., & Kumar, A. (2023). Regional and socioeconomic inequalities in access to pre-primary education in India. International Journal of Child Care and Education Policy.
18. MTSE Pattern/Guide (MAT/SAT components). Time4Education.