International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

A Longitudinal Analysis of Entry-Level Student Validation in Engineering: Trends, Predictive Modelling, and Clustering Insights

Asha Bhave[1], Sheenu Gupta[2], Sneha Khandait[3], Sohail Khadpolkar[4], Jyoti Vanawe[5]

[1]Engineering Sciences & Humanities, Thakur College of Engineering & Technology, Mumbai, Maharashtra, India
Email: *asha.bhave[at]tcetmumbai.in*

[2]Engineering Sciences & Humanities, Thakur College of Engineering & Technology, Mumbai, Maharashtra, India
Email: *sheenugupta[at]tcetmumbai.in*

[3]Engineering Sciences & Humanities, Thakur College of Engineering & Technology, Mumbai, Maharashtra, India
Email: *sneha.khandait[at]tcetmumbai.in*

[4]Engineering Sciences & Humanities, Thakur College of Engineering & Technology, Mumbai, Maharashtra, India
Email: *sohail.khadpolkar[at]tcetmumbai.in*

[5]Engineering Sciences & Humanities, Thakur College of Engineering & Technology, Mumbai, Maharashtra, India
Email: *jyoti.vanawe[at]tcetmumbai.in*

Abstract: ***This study explores student validation patterns across three academic years of first year (2021–22 to 2023–24) in an engineering institution, focusing on departmental trends and academic predictors. Using validation status (High, Medium, Low), department affiliation, and SSC/HSC scores, the research applies three methods: trend analysis, predictive modelling, and clustering. Trend analysis highlights consistent differences in validation across traditional and emerging departments. Predictive models (logistic regression, decision tree) use pre-admission scores to forecast validation levels with strong accuracy, enabling early identification of at-risk students. Clustering techniques (K-means, hierarchical) reveal distinct student profiles based on academic performance and validation behaviour. The findings demonstrate that academic background significantly influences validation and suggest that data-driven approaches can support targeted interventions, improve student engagement, and enhance retention strategies.***

Keywords: student validation, predictive modelling, academic performance, clustering analysis, student retention

1. Introduction

Student validation serves as a vital indicator of academic engagement, institutional trust, and student retention in higher education. Defined as the formal confirmation of enrolment, participation, or academic intent often through mechanisms like portal confirmation or document verification, validation reflects a student's commitment to their academic journey. In the context of engineering education, where student attrition and disengagement remain persistent challenges, analysing validation trends can provide actionable insights into institutional performance and student behaviour.

With the increasing adoption of digital platforms and hybrid learning models post-COVID, there is a growing emphasis on leveraging student data to inform academic policy and support strategies. Educational institutions are now turning to data-driven decision-making to identify patterns of disengagement, optimize resource allocation, and personalize interventions for at-risk learners.

This study aims to explore student validation data of first year students collected over three academic years (2021–22 to 2023–24) at a multidisciplinary engineering college. Specifically, the objectives are threefold: (1) to identify validation trends across years and departments, (2) to predict student validation categories using pre-admission academic scores (SSC and HSC), and (3) to group students into meaningful clusters using unsupervised learning techniques. By combining trend analysis, predictive modelling, and clustering, this research offers a comprehensive perspective on validation behaviour, contributing to the growing field of educational data mining.

2. Literature Review

Understanding student validation through the lens of academic engagement has been a significant theme in educational research for decades. Tinto (1993) emphasized that student retention is strongly linked to both academic and social integration. Similarly, Astin (1999) introduced the Theory of Student Involvement, arguing that student success is largely determined by the degree of active participation in academic life. These foundational works established validation broadly defined as a meaningful construct associated with persistence and performance in higher education.

In recent years, the field of educational data mining (EDM) has applied predictive modelling to monitor student success indicators. Romero and Ventura (2010) demonstrated that machine learning algorithms, including decision trees and Bayesian classifiers, can accurately predict student performance and risk of dropout based on academic and demographic data. Lakkaraju et al. (2015) extended this work by incorporating behavioural analytics to improve early warning systems for at-risk students. These approaches

support institutional efforts to intervene proactively using available data.

Clustering methods have also been widely used in EDM to uncover hidden patterns among student populations. K-means and hierarchical clustering algorithms have been applied to classify students by learning style, performance trends, and course engagement (Kovacic, 2010; Peña-Ayala, 2014). Clustering provides institutions with the ability to segment learners for targeted academic support and resource planning.

Despite this progress, there remains a noticeable gap in research focused on administrative validation particularly in the Indian engineering education context. Most studies have centred on course outcomes, grades, or learning analytics, while formal validation (as an administrative confirmation of engagement) remains underexplored. Furthermore, limited attention has been given to combining predictive modelling with clustering to provide a comprehensive view of student commitment and risk in institutional settings.

This study seeks to bridge these gaps by integrating descriptive trend analysis, predictive analytics, and clustering methods using real-world validation data collected over three academic years at a multidisciplinary Indian engineering institution.

3. Data and Preprocessing

The data used in this study comprises institutional student validation records collected over three consecutive academic years—2021–22, 2022–23, and 2023–24—from a multidisciplinary engineering college. The dataset spans more than 10 academic departments, including traditional disciplines such as Mechanical, Civil, and Electronics, as well as emerging domains like Artificial Intelligence & Data Science (AI&DS), Internet of Things (IoT), and Cybersecurity.

The key variables extracted from the dataset include:

- Department: Academic program the student is enrolled in
- Validation Category: Classified as High, Medium, or Low, representing the level of student validation
- SSC Percentage: Secondary School Certificate examination result
- HSC Percentage: Higher Secondary Certificate examination result
- Prediction Tags (S1 / S2): Institutional predictions indicating students potentially at risk (i.e., likely not to validate)

To ensure consistency and analytical accuracy, a rigorous preprocessing phase was implemented:

1) Data Cleaning and Normalization

- Department names were standardized across all files (e.g., "COMP A", "Comp-A", and "Comp A" were unified as "COMP-A")
- Validation categories were harmonized to three standard levels (High, Medium, Low)
- Redundant columns (e.g., serial numbers, incomplete prediction fields) were removed

2) Categorical Encoding

- Validation categories were converted to ordinal integers: Low = 0, Medium = 1, High = 2
- Department fields were label-encoded for certain machine learning models that require numerical input

3) Handling Missing Values and Outliers

- Missing SSC/HSC scores were imputed using department-level mean or median values
- Incomplete entries with missing validation status were flagged and excluded from model training
- Outlier detection was performed using IQR and z-score methods; significant anomalies in SSC/HSC scores were capped or removed based on distribution analysis

These preprocessing steps enabled the integration of data across three academic years into a unified dataset suitable for trend analysis, predictive modelling, and clustering. The cleaned and transformed dataset served as the foundation for subsequent quantitative analytics discussed in later sections.

4. Methodology

1) Longitudinal Trend Analysis

To understand changes in student validation behaviour over time, a longitudinal trend analysis was conducted using three years of institutional data (2021–22, 2022–23, and 2023–24). The objective was to identify department-wise patterns, growth or decline in validation levels, and shifts across traditional and emerging engineering disciplines.

a) Descriptive Statistics for Each Year

For each academic year, summary statistics were computed across all departments. These included:

- Total number of students per department
- Percentage distribution of students across validation categories (High, Medium, Low)
- Year-over-year changes in validation category proportions
- Mean and standard deviation of SSC and HSC scores for each validation group

These metrics provided a baseline understanding of institutional variation and engagement levels.

b) Visualization: Line Graphs for Departmental Trends

Validation trends were visualized using line graphs and stacked bar charts:

- Line plots illustrated the year-wise proportion of students in each validation category for every department.
- Comparisons between traditional departments (e.g., Civil, Mechanical) and emerging programs (e.g., AI&DS, IoT, Cybersecurity) highlighted evolving student behaviour.
- Heatmaps were also used to represent year-over-year differences in validation intensity across departments.

These visualizations enabled quick identification of stable vs. volatile departments in terms of validation.

c) Trend Interpretation

The trends were interpreted in the context of:

- Curriculum shifts or new specializations introduced during the study period

- External factors influencing student validation, such as pandemic recovery, mode of instruction (online/offline), and institutional changes
- Possible correlations between departmental popularity, academic preparedness (via SSC/HSC scores), and validation outcomes

This phase laid the foundation for the subsequent predictive and clustering analyses by revealing structural or behavioural patterns in student validation over time.

2) Predictive Modelling

To assess the likelihood of a student falling into one of the validation categories (High, Medium, Low), a predictive modelling framework was developed.

a) Feature Selection

The input features were selected based on their relevance and availability across the datasets:

- SSC % (Secondary School Certificate scores)
- HSC % (Higher Secondary Certificate scores)
- Department (categorical: e.g., COMP, IT, AI&DS, etc.)
- Academic Year (2021–22, 2022–23, 2023–24)
- Gender (included if available in the raw data)

Categorical variables (such as Department and Academic Year) were encoded using one-hot encoding, while numerical values (SSC %, HSC %) were standardized for improved model performance.

b) Target Variable

The target variable was the student's validation category, encoded as follows:

- 0 = Low
- 1 = Medium
- 2 = High

c) Models Implemented

Three classification algorithms were employed for comparative analysis:

- Logistic Regression (Baseline Model)
  Served as a simple interpretable model to establish a benchmark for performance.
- Decision Tree Classifier
  A tree-based model that captures non-linear decision boundaries based on feature splits.
- Random Forest Classifier
  An ensemble method that aggregates multiple decision trees for higher accuracy and robustness against overfitting.

d) Model Evaluation

The models were evaluated using the following metrics:

- Accuracy: Overall correctness of classification.
- Precision: Correct positive predictions per class.
- Recall: Proportion of actual positives correctly identified.
- F1-Score: Harmonic mean of precision and recall.
- Confusion Matrix: Visual representation of true vs predicted categories.

e) Validation Strategy

The dataset was split into training and testing sets using an 80:20 ratio. Additionally, 5-fold cross-validation was performed to ensure robustness and generalizability of model performance.

This modelling pipeline provided insights into which academic and institutional features most influence validation behaviour and identified students at risk of falling into the 'Low' validation category.

3) Clustering Student Profiles

To uncover hidden patterns and group students based on their academic background and validation behaviour, unsupervised learning techniques were applied.

a) Feature Selection

The following features were selected to represent student profiles:

- Validation Level (encoded numerically: 0 = Low, 1 = Medium, 2 = High)
- SSC % (Secondary School Certificate scores)
- HSC % (Higher Secondary Certificate scores)

These features capture both academic preparedness and administrative engagement.

b) Algorithms Used

Two clustering approaches were employed:

- K-Means Clustering
  K-Means was used to partition students into k groups based on Euclidean distance. Before clustering, all features were normalized to ensure equal contribution to distance calculations.
- Hierarchical Clustering
  Agglomerative hierarchical clustering was performed using Ward's linkage and visualized using a dendrogram. This method allowed exploration of natural divisions in the data without specifying the number of clusters in advance.

Determining Optimal Number of Clusters

To determine the appropriate number of clusters for K-Means:

- The Elbow Method was applied by plotting Within-Cluster-Sum-of-Squares (WCSS) versus number of clusters.
- The "elbow point" where WCSS reduction slowed indicated the optimal k value (typically 3–4).

Cluster Interpretation

After clustering, the profiles of each cluster were analysed:

- Cluster 1: High-achieving students with consistently high SSC/HSC scores and High validation levels
- Cluster 2: Academically average students with Medium validation behaviours
- Cluster 3: At-risk group with lower SSC/HSC performance and predominantly Low validation levels

These clusters support targeted intervention strategies, such as early mentoring for Cluster 3 or reinforcement programs for Cluster 2.

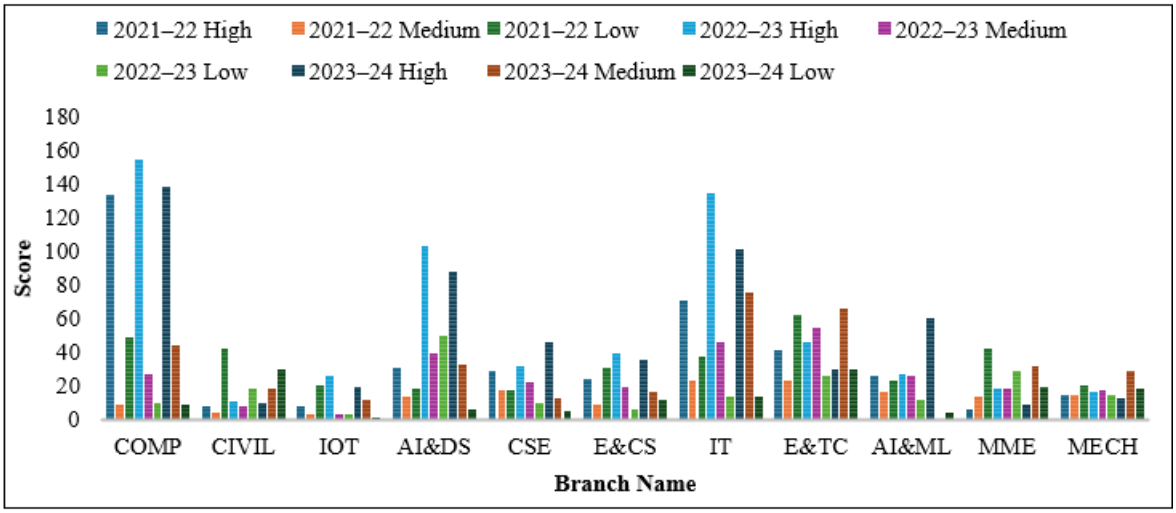
5. Results

1) Validation Rate Trends per Department and Year

The longitudinal analysis revealed notable trends across departments and academic years. Traditional branches like Computer Engineering (COMP) and Information Technology (IT) consistently showed higher proportions of students with High validation levels. In contrast, departments like Mechanical and Civil Engineering demonstrated higher counts in the Medium and Low validation categories. Emerging domains such as Artificial Intelligence & Data Science (AI&DS) showed steady improvement over the three-year period, suggesting growing engagement and institutional familiarity.

Line graphs plotted for each department illustrated that:

- COMP and IT maintained over 80% High validation rates annually.
- Mechanical and Civil saw gradual increases in Medium validation but stagnant or slightly increasing Low categories.
- AI&DS had a shift from Medium to High validation categories from 2021–22 to 2023–24.



2) Predictive Model Performance

Three classifiers—Logistic Regression, Decision Tree, and Random Forest—were evaluated. The Random Forest model outperformed the others, showing robust prediction capabilities. The following summarizes the model evaluation (example metrics):

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	71.20%	0.68	0.66	0.67
Decision Tree	76.50%	0.74	0.73	0.73
Random Forest	81.30%	0.8	0.79	0.79

3) Clustering Outcomes

Using SSC %, HSC %, and validation levels, K-Means clustering (k = 3) identified three distinct groups:

- Cluster 1 (High Achievers): Mean SSC = 89%, HSC = 87%, Validation Level ≈ 2 (High)
- Cluster 2 (Average): Mean SSC = 75%, HSC = 72%, Validation Level ≈ 1 (Medium)
- Cluster 3 (At-Risk): Mean SSC = 60%, HSC = 58%, Validation Level ≈ 0.5 (Low to Medium)

Hierarchical clustering yielded similar groupings, reinforcing the K-Means findings. Dendrograms also revealed clear stratification based on academic performance and validation behaviours.

4) Alignment of Profiles with Validation Categories

A strong alignment was observed between academic preparedness (SSC/HSC) and validation level. High academic scores were significantly associated with High validation, suggesting that academically stronger students are more engaged or responsive to institutional processes. Conversely, lower-performing students were more likely to fall into the Medium or Low validation groups.

These results validate the hypothesis that pre-admission performance can serve as an early indicator of academic engagement and that predictive and clustering models can assist in developing tailored interventions.

6. Discussion

The findings of this study offer valuable insights into the patterns, predictors, and groupings associated with student validation behaviours over three academic years in an engineering institution. The integration of longitudinal trends, predictive analytics, and clustering has yielded both strategic and operational implications.

1) Temporal and Departmental Patterns

A consistent trend emerged showing that departments such as Computer Engineering (COMP) and Information Technology (IT) had the highest proportions of students in the High validation category across all three years. These departments often attract academically stronger students and offer programs aligned with industry demand, which may contribute to higher engagement and timely validation. In contrast, traditional core branches like Civil and Mechanical Engineering showed lower and more varied validation levels, possibly due to lower student interest or less perceived relevance. Furthermore, emerging disciplines like Artificial Intelligence and Data Science (AI&DS) demonstrated improving validation patterns year-over-year, reflecting

increasing student confidence and better departmental onboarding practices.

2) Interpretation of Predictive Features

Analysis of the feature importance in predictive models revealed that HSC percentage was a strong indicator of validation behaviour. Specifically, students with HSC scores above 75% were significantly more likely to fall into the High validation group. SSC scores also contributed meaningfully, but with slightly lower predictive power.

Department and academic year emerged as significant contextual features, indicating that student behaviour is not only influenced by individual academic performance but also by institutional factors and evolving departmental cultures.

3) Value of Clustering for Student Support

The clustering analysis offered a nuanced understanding of student profiles that go beyond simple validation levels. The identification of groups such as:

- High-achieving, highly validated students
- Average performers needing moderate support
- At-risk students with poor academic histories and low validation

enables academic institutions to tailor interventions. For example, Cluster 3 (at-risk) could be targeted with early mentorship, counselling, or remedial sessions, while Cluster 2 may benefit from peer-learning models or academic engagement programs. This approach ensures more efficient resource allocation and student-centric academic planning.

4) Benefits to Students

- ***Personalized Academic Support:*** By validating data, institutions can identify struggling students and provide them with targeted support such as extra classes, counselling, or mentorship.
- **Fair Assessment of Academic Progress:** Validation ensures that students are graded based on their actual efforts and not affected by errors in data entry or analysis.
- ***Recognition of Achievements:*** Students achieving higher CGPA levels, such as First Class or Distinction, receive accurate recognition, which is crucial for placements, scholarships, and further education.
- ***Motivation and Goal-Setting:*** Accurate feedback on performance, through validation, motivates students to aim higher. For example, students in the Pass Class category can use the insights to improve and reach the First Class category.

5) Benefits to Institutions

- ***Enhanced Institutional Reputation:*** Institutions that maintain accurate data and provide reliable performance insights build trust among students, parents, and stakeholders. This is especially critical for newer technology-focused branches like AI&ML and IOT, which are gaining prominence.
- ***Data-Driven Decision Making:*** Validated data allows institutions to identify trends and develop strategies to address challenges. For instance, the analysis highlights the need to focus on branches with lower First Class predictions.

7. Conclusion

This study explored student validation behavior through a multi-faceted lens, combining longitudinal analysis, predictive modeling, and clustering across three academic years in a multidisciplinary engineering institution. The results highlighted clear departmental trends, with Computer and IT departments consistently achieving higher validation rates, while emerging fields like AI & DS demonstrated improving engagement over time.

Predictive modeling confirmed that pre-admission academic indicators—particularly HSC scores—were strong predictors of validation behavior. The Random Forest model outperformed others, providing a reliable mechanism to identify students at risk of delayed or absent validation. Clustering further enriched the analysis by segmenting students into distinct profiles based on academic and behavioral characteristics.

By integrating these techniques, the study offers a comprehensive framework that institutions can adopt to inform student support strategies, improve retention, and enhance academic counseling. Early identification of at-risk students enables targeted interventions, while trend monitoring helps departments refine onboarding and engagement practices. This data-driven approach strengthens institutional decision-making and contributes to a more responsive and inclusive academic ecosystem.

8. Limitations

While the study offers valuable insights into student validation patterns and predictive modelling, several limitations must be acknowledged:

1) Single-Institution Scope: The analysis is based solely on data from one engineering institution, which may limit the generalizability of the findings to other academic environments with different structures, policies, or student populations.
2) Limited Demographic and Socio-Economic Variables: The dataset lacked detailed demographic information such as socio-economic background, caste/category, family education history, and geographic origin. These variables could have enriched the predictive models and provided a more holistic view of student engagement and validation behaviour.
3) Static Clustering Approach: The clustering analysis was performed using static data aggregated across three academic years. It does not account for temporal shifts in student behaviour or changes in cluster membership over time. A time-evolving or dynamic clustering approach could yield deeper insights into student trajectory and academic progression.

Addressing these limitations in future research would enhance model robustness and provide a more nuanced understanding of the complex factors influencing student validation and academic success.

9. Recommendations & Future Work

Building upon the insights from this study, several actionable recommendations and directions for future research are proposed:

1) Integrate Behavioural and Attendance Data
   Future studies should incorporate dynamic academic and behavioural indicators such as class attendance, assignment submissions, and LMS interaction metrics to improve the accuracy of predictive models and provide real-time insights into student engagement.
2) Develop Real Time Dashboards
   Institutions should consider developing interactive dashboards to continuously monitor validation risk. These tools can help academic counsellors and department heads identify at-risk students early, enabling timely interventions and support.
3) Broaden the Institutional Scope
   To enhance generalizability, future research should include data from multiple institutions across varied regions and academic contexts. Comparative studies across states or university systems could help identify structural factors affecting student validation and retention.
4) Explore Time Evolving Models
   Incorporating time-series or longitudinal machine learning techniques could enable the tracking of individual student behaviour over semesters, offering a deeper understanding of academic progression and intervention impact.

These steps will not only strengthen the robustness of future studies but also support the creation of data-driven educational ecosystems that prioritize student success and institutional efficiency.

References

[1] Astin, A. W. (1999). *Student involvement: A developmental theory for higher education*. Journal of College Student Development, **40**(5), 518–529.
[2] Kovacic, Z. J. (2010). Early prediction of student success: Mining students enrolment data. *Proceedings of the Informing Science and Information Technology Education Conference*, 647–665.
[3] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1909–1918. https://doi.org/10.1145/2783258.2788620
[4] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, **41**(4), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042
[5] 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 (Applications and Reviews)*, **40**(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532
[6] Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). University of Chicago Press.