International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 7 No. 28 March - April - 2026

Statistical Analysis of Students' Productivity

Prof. Vishwajit Vilas Khajekar, Aachal Chandrabhan Taras, Gayatri Khanderao Wagh
Department of Statistics, New Arts, Commerce and Science College, Ahmednagar, India
Corresponding Author – Prof. Vishwajit Vilas Khajekar
DOI - 10.5281/zenodo.19396319

Abstract:

Student productivity is shaped by a complex interaction of psychological, academic, environmental, physical, and social factors. This study investigates the key determinants affecting productivity among 18–21-year-old undergraduates, using a dataset of 843 responses collected via Google Forms on a 5-point Likert scale. Multiple Linear Regression ($R^2 = 0.2395$, $F = 88.08$, $p < 0.001$) identified rapid heartbeat, anxiety, and academic conflict as primary stress drivers. Machine learning classifiers — Random Forest, XGBoost, Logistic Regression, Association Rule Mining, and K-Means Clustering — were applied to classify students by productivity level. XGBoost achieved the highest accuracy (90.5%, AUC = 0.976), with workload overwhelm, low academic confidence, and poor concentration emerging as the three strongest predictors. Approximately 29% of students were flagged as at-risk. Findings support the development of data-driven, early-warning academic support frameworks. The findings highlight the multidimensional nature of personal productivity and identify the importance of emotional well-being and academic environment management. This study describe the usefulness of statistical and machine learning approaches in understanding and predicting productivity patterns among students.

Keywords: Student Productivity, Academic Stress, Machine Learning, Xgboost, Random Forest, K-Means Clustering, Association Rules, Likert Scale, Higher Education.

Introduction:

Student productivity surround far more than academic grades — it reflects the capacity to simultaneously manage reasonable demands, emotional states, physical health, and social obligations. Among undergraduates aged 18–21, this capacity is consistently undermined by escalating coursework, peer competition, lifestyle adjustments, and limited access to mental health resources.

Standard grade-based assessment fails to expose the underlying interdependencies among physiological, psychological, and behavioral variables that jointly govern productive capacity. Prior research links academic stress (Misra & McKean, 2000), sleep deprivation (Hershner & Chervin, 2014), and perceived workload overload (Kember et al., 2004) to measurable declines in student output. More recently, ensemble machine learning — especially XGBoost and Random Forest — has demonstrated superior classification accuracy over logistic regression in educational data mining tasks (Chen & Guestrin, 2016).

This paper presents a multi-method investigation using regression, association rule mining, cluster analysis, and two ensemble classifiers on data from 843 respondents — aiming to identify at-risk students, quantify variable importance, and inform institutional intervention strategies.

The study is organized as follows: Section 2 describes the data and methods; Section 3 presents statistical and ML results; Section 4 compares model performance; Section 5 discusses implications and concludes.

Methodology:

1. Data & Instrument:

A secondary dataset of 843 valid undergraduate responses (ages 18–21) was assembled from a structured Google Forms survey using a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Five variable domains were captured: demographic (age, gender), emotional (anxiety, sadness, irritability, loneliness), physical (sleep quality, headaches, rapid heartbeat), academic (workload, confidence, attendance), and social (home environment, relationship stress).

2. Productivity Score & Labels:

A composite productivity score was derived by aggregating item responses (range: 1.667–4.667; mean = 3.367; median = 3.333). Three productivity tiers were defined: Low (40.3%), Medium (34.8%), and High (24.9%). A binary At-Risk label was assigned to students in the bottom 25th percentile, yielding 243 at-risk cases (28.83%). The dataset was partitioned using a 70–30 stratified train-test split: 591 training and 252 test observations.

3. Analytical Pipeline:

Five complementary techniques were applied in sequence:

- Multiple Linear Regression — identify significant stress predictors (F-test, R^2).
- Association Rule Mining — uncover symptom co-occurrence via support, confidence, lift.
- K-Means Cluster Analysis (k = 4) — segment students into behavioral profiles.
- Random Forest (mtry = 10) — ensemble classification with %IncMSE variable importance.
- XGBoost (77 rounds, CV AUC = 0.976) — gradient boosting with Gain-based importance.

Model performance was evaluated using accuracy, sensitivity, specificity, F1 score, and AUC on the held-out test set. All analyses were conducted in R.

Results:

1. Multiple Linear Regression:

The regression model explained 23.95% of variance in recently experienced stress ($R^2 = 0.2395$, Adj. $R^2 = 0.2368$, $F(3,839) = 88.08$, $p < 0.001$). Three variables emerged as significant predictors: rapid heartbeat (strongest effect), anxiety/tension, and academic–extracurricular conflict. These physiological and scheduling stressors collectively confirm that student stress — and by extension productivity impairment — is driven by both bodily arousal and academic overload rather than by demographic characteristics alone.

2. Association Rule Mining:

Six rules with lift > 1.2 revealed statistically meaningful co-occurrences among stress symptoms (Table 1). The most notable finding is a bidirectional anxiety ↔ sleep disturbance relationship (lift ≈ 1.80 in both directions, co-occurring in 15.3% of students) — confirming a self-reinforcing feedback loop. Irritated students showed sharply elevated concentration difficulties (lift = 1.77), while academic conflict predicted irregular attendance with 58% confidence.

Association Rule	Confidence	Lift	Implication
Headaches → Irritation	53.0%	1.81	Physical pain amplifies mood dysregulation
Anxiety → Sleep Problems	54.6%	1.80	Anxiety disrupts restorative sleep
Sleep Problems → Anxiety	50.6%	1.80	Sleep loss heightens anxiety — bidirectional loop
Irritation → Concentration Issues	50.0%	1.77	Irritability impairs academic focus
Rapid Heartbeat → High Stress	56.0%	1.67	Physiological arousal signals elevated stress
Academic Conflict → Attendance Drop	58.0%	1.27	Overload leads to disengagement behaviour

Table 1: Key Association Rules — All Lift Values Exceed Chance Threshold of 1.0

3. K-Means Cluster Analysis (k = 4):

Cluster selection was validated through three convergent methods (elbow, silhouette, and gap statistic — Fig. 1), all pointing to k = 4 as the most interpretable solution. PCA-reduced scatter plots (Fig. 2) confirm well-separated groupings. Four distinct student profiles emerged:

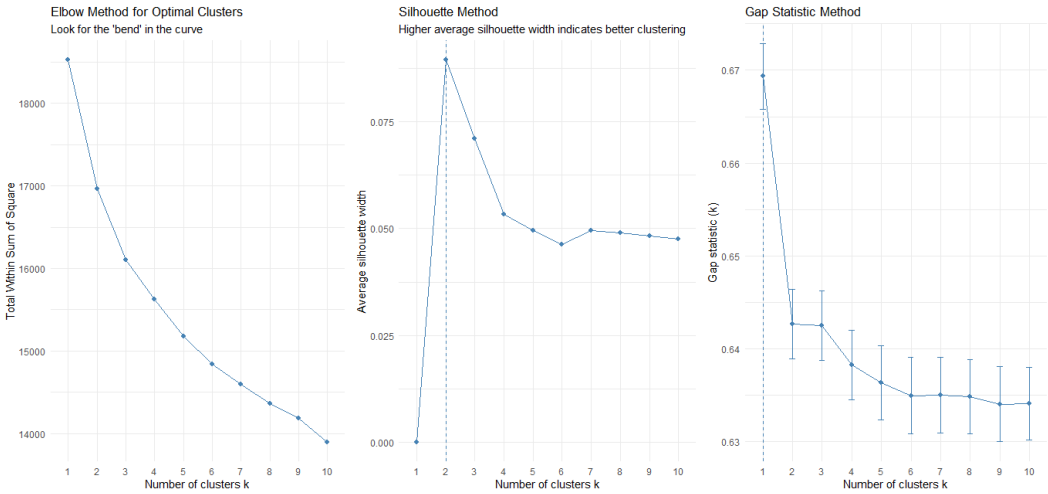

Fig. 1: Cluster Optimization — Elbow, Silhouette & Gap Statistic Methods

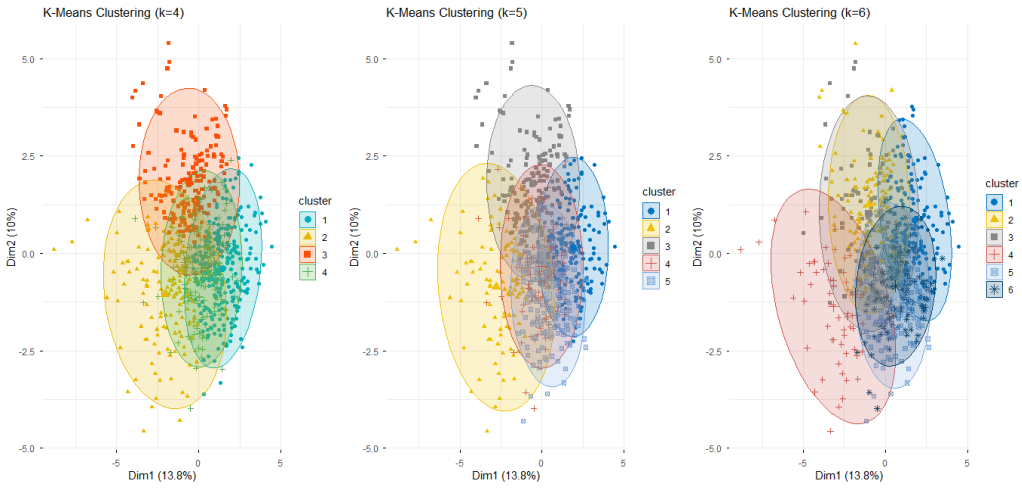

Fig. 2: K-Means Clustering via PCA Reduction (k = 4, 5, 6)

Cluster	n	%	Mean Stress	Behavioral Profile
1	290	34.4%	2.64	Moderately stressed — coping adequately
2	168	19.9%	3.26	Elevated tension — emerging at-risk subgroup
3	196	23.3%	2.68	High-functioning — low stress, strong focus
4	189	22.4%	3.64	Chronically stressed — priority intervention target

Table 2: K-Means Student Cluster Profiles (n = 843)

Cluster 3 (23.3%) is particularly instructive: its coexistence with high-stress clusters within the same institution demonstrates that structural factors — workload design, support access, and scheduling — rather than purely individual traits determine productivity outcomes.

4. Machine Learning — Variable Importance:

Both Random Forest (Fig. 3) and XGBoost (Fig. 4) independently rank Workload_Overwhelm, Low_Academic_Confidence, and Poor_Concentration as the top three predictors of at-risk classification. The near-identical rankings across two algorithmically distinct models provide robust, cross-validated evidence that academic self-regulation — not physiological or demographic factors — is the dominant determinant of student productivity risk. The side-by-side comparison in Fig. 5 further underscores this convergence.

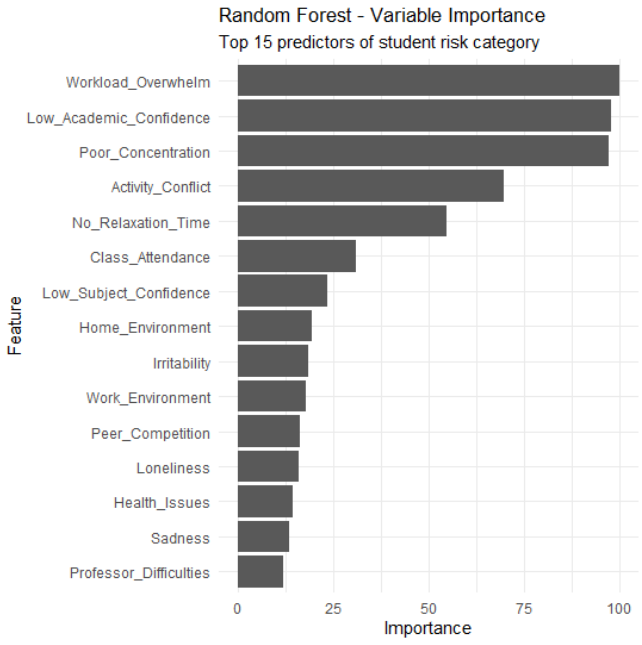

Fig. 3: RF Variable Importance — Top 15 Predictors

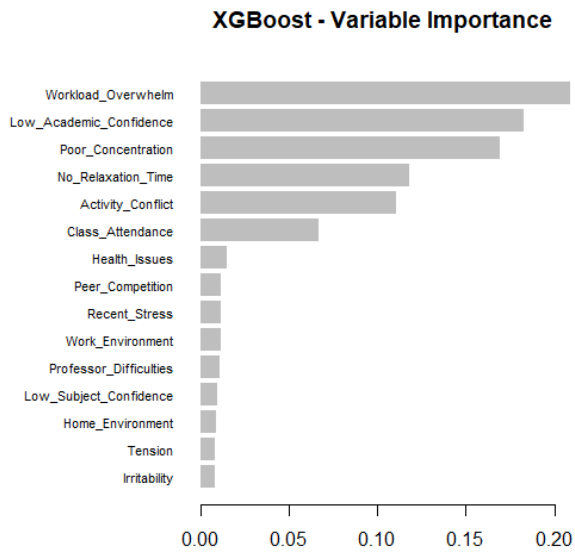

Fig. 4: XGBoost Variable Importance (Gain)

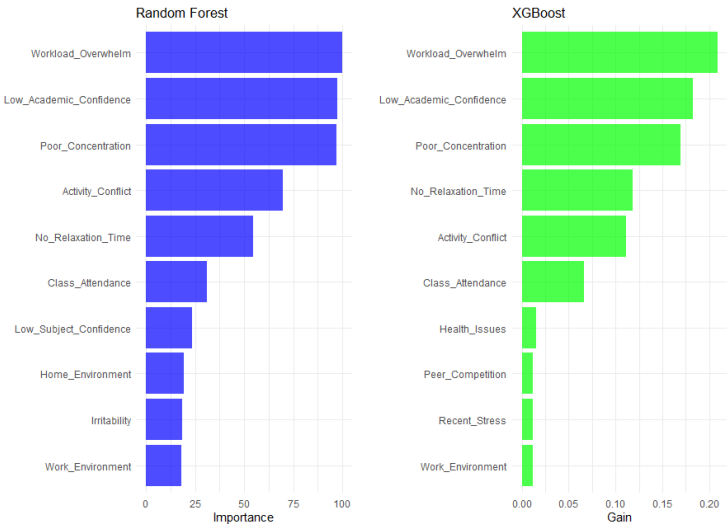

Fig. 5: Side-by-Side Feature Importance — Random Forest (blue) vs XGBoost (green)

Model Performance Comparison:

Table 3 summarises classification performance on the held-out test set (n = 252). XGBoost consistently leads across all metrics. Its higher sensitivity (95.6%) is critical in this context: minimising false negatives ensures the majority of at-risk students are correctly identified and referred for support. The ROC curves (Fig. 6) visually confirm XGBoost's superior discriminative power (AUC = 0.975 vs RF = 0.956) across all operating thresholds. The three-class Random Forest (κ = 0.661) shows substantial agreement with ground truth, demonstrating that fine-grained productivity-level classification is feasible.

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC
Random Forest	0.897	0.950	0.764	0.929	0.956
XGBoost	0.905	0.956	0.778	0.935	0.975
Logistic Regression	0.880	0.940	0.720	0.910	0.956
3-Class RF	0.778	—	—	κ=0.661	—

Table 3: Comparative Model Performance on Test Set (n = 252)

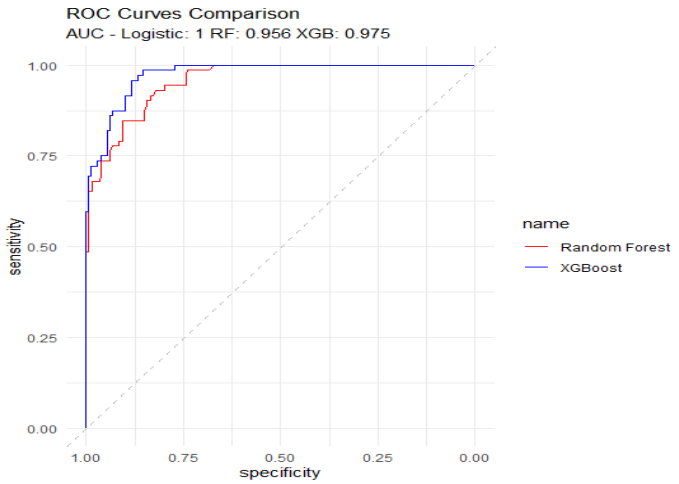

Fig. 6: ROC Curve Comparison — Random Forest (AUC = 0.956) vs XGBoost (AUC = 0.975)

Discussion, Conclusions & References:

1. Discussion:

Results converge across all five methods on a coherent account of student productivity risk: it is not a single-cause phenomenon but emerges from the co-occurrence of cognitive overload, emotional dysregulation, sleep disruption, and eroded academic self-efficacy.

The bidirectional anxiety–sleep feedback loop (lift ≈ 1.80, affecting 15.3% of students) is the most clinically actionable finding. Both variables respond to evidence-based interventions — cognitive behavioural therapy for anxiety and sleep hygiene programmes — suggesting that dual-pathway support could yield compounded gains in this cohort.

The 29% at-risk rate signals a structural challenge. At an institution of 843 students, that represents ~243 individuals needing targeted support. XGBoost's 95.6% sensitivity ensures nearly all of them are correctly flagged, making it a practical foundation for an early-warning dashboard that prioritises resource allocation before decline becomes entrenched.

Cluster 3's high-functioning profile — achievable within the same environment as the chronically stressed Cluster 4 — demonstrates that workload redesign, flexible scheduling, and campus wellness investment are evidence-based institutional levers, not merely aspirational.

2. Conclusions:

1. Workload overwhelm and low academic confidence are the dominant risk predictors, jointly accounting for ~39% of XGBoost model Gain.
2. A bidirectional anxiety–sleep loop (lift ≈ 1.80) affects 15.3% of students, forming a mutually reinforcing productivity drain.
3. Approximately 29% of students are at-risk; XGBoost detects them with 90.5% accuracy and AUC = 0.976 — outperforming all competing models.
4. Four student clusters reveal that 23.3% are high-functioning — proving that structural intervention can shift population-level outcomes.
5. Future work should integrate real-time LMS data and longitudinal follow-up to validate the predictive pipeline across diverse contexts.

References:

1. Hershner, S. D., & Chervin, R. D. (2014). Nature and Science of Sleep, 6, 73–84.
2. Kember, D., et al. (2004). Assessment & Evaluation in Higher Education, 29(5), 553–572.
3. Misra, R., & McKean, M. (2000). American Journal of Health Studies, 16(1), 41–51.
4. Breiman, L. (2001). Machine Learning, 45(1), 5–32.
5. Chen, T., & Guestrin, C. (2016). Proc. 22nd ACM KDD Conference, 785–794.
6. Agrawal, R., & Srikant, R. (1994). Proc. 20th VLDB Conference, 487–499.
7. Hair, J. F., et al. (2019). Multivariate Data Analysis (8th ed.). Cengage.
8. Zimmerman, B. J. (2002). Theory into Practice, 41(2), 64–70.