Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Stress Indicators and Academic Performance in Gen-Z Students: A Predictive Study

A Quantitative Study Using Machine Learning Classification Models

Kaveri Koli
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce, Science College, Pimpri.
Pune, India

Gaurav Lalwani
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce, Science College, Pimpri.
Pune, India

Abstract **- Academic performance in higher education is influenced by behavioral, cognitive, and environmental factors. Stress, burnout, procrastination, family pressure, attendance, and concept clarity are commonly cited influences on student outcomes. This study analyzes survey data from 201 students across Indian universities to predict students at risk of lower CGPA using classification models. Logistic Regression and Random Forest Classifier were applied with an 80/20 stratified train/test split, 5-fold cross-validation, and bootstrapped confidence intervals to ensure statistical reliability. Results show modest predictive performance, with Random Forest and Logistic Regression achieving ~51% accuracy, highlighting the complexity and high variance in behavioral data. Feature importance analysis identifies stress, attendance, and concept clarity as the strongest predictors. These findings emphasize careful feature engineering and rigorous evaluation when using survey-based measures to predict academic performance.**

Keywords - Generation Z; academic stress; logistic regression; random forest; educational data mining; risk classification.

I. INTRODUCTION

Academic performance is a multifactorial outcome shaped by cognitive abilities, behavioral habits, and environmental pressures. Stress management, procrastination tendencies, burnout levels, and family pressure can influence students' ability to maintain high grades. While previous research demonstrates correlations between these factors and CGPA, few studies rigorously evaluate their predictive power using machine learning approaches in mixed-discipline student populations.

This study focuses on predicting "At-Risk" students based on CGPA percentile thresholds using survey-based indicators. Unlike prior work, we implement stratified train/test splits, cross-validation, and bootstrapped confidence intervals, ensuring more reliable and generalizable results. Feature engineering is applied through six primary behavioral scores: stress, burnout, procrastination, family pressure, attendance, and concept clarity. Unlike prior studies that report high predictive scores without rigorous validation, this study demonstrates the challenges of modeling behavioral survey data and emphasizes the importance of honest benchmarking against baseline classifiers.

II. RELATED WORK

A. Correlation Between Stress Dynamics and Academic Outcomes

The impact of psychological strain on educational attainment is a well-documented phenomenon. Reddy et al. [1] identified that academic stress among Indian university students originates from a combination of high curriculum demands and parental expectations, acting as a significant inhibitor to academic consistency. Furthermore, Mushtaq and Khan [2] observed that while social and family factors are influential, student performance is most acutely affected by personal behavioral traits and examination anxiety. These studies establish the multi-dimensional nature of stress but primarily rely on descriptive statistics, lacking the predictive granularity required for automated student-at-risk detection.

B. Digital Stressors and Gen-Z Behavioral Profiles

Generation Z's academic experience is uniquely characterized by "digital saturation." Montag and Walla [3] demonstrated that excessive smartphone usage and the resulting "digital fatigue" can lead to measurable declines in cognitive focus and academic engagement. In the South Asian context, the interplay between technology and cultural expectations creates unique stressors. Steel [4] has established through meta-analytic reviews that procrastination—often exacerbated by digital distractions—is a quintessential self-regulatory failure that correlates strongly with lower GPA. Despite these insights, digital

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

habits are rarely integrated as features within high-dimensional machine learning frameworks.

C. Computational Approaches in Educational Data Mining (EDM)

The transition from traditional statistics to computational modeling marks a shift toward proactive intervention. Early attempts by Shatkin et al. [5] to model emotional outcomes and academic success proved that resilience-based interventions could be tracked; however, these models often lacked the predictive power for individual-level forecasting. Recent advancements in Educational Data Mining (EDM) suggest that ensemble methods are superior for tabular behavioral data. Pascoe et al. [6] emphasized the need for modern research to bridge the gap between psychological health and objective academic metrics, a challenge that requires non-linear modeling techniques.

D. Identifying the Research Gap

A review of existing literature reveals a "performance gap" in current systems. Most models fail to account for the stochastic noise of self-reported psychological data, often dismissing it as unpredictable. This study addresses this gap by implementing a rigorously validated classification framework using Logistic Regression and Random Forest models with stratified cross-validation and confidence interval estimation.

III. METHODOLOGY

A. Dataset Acquisition and Participant Profile

The dataset was obtained through a structured digital survey administered to undergraduate students enrolled in Indian technical and liberal arts institutions. The initial sample consisted of $n = 238$ respondents between the ages of 18 and 25, representing the Generation Z demographic in higher education. Participation was voluntary and anonymous to encourage honest reporting of academic and psychological experiences.

The survey instrument collected self-reported academic and behavioral indicators, including stress frequency, burnout, procrastination, attendance patterns, perceived family pressure, and conceptual clarity. CGPA was reported in categorical ranges and subsequently mapped to a standardized 10-point scale to ensure numerical consistency across participants.

Following data integrity checks, incomplete or inconsistent responses were removed. After cleaning, the final analytical dataset comprised 201 valid observations, reflecting a 15.5% reduction from the original sample. This preprocessing step ensured that downstream modeling was conducted on reliable and internally consistent data.

B. Feature Engineering and Ordinal Encoding

Survey responses were originally qualitative in nature and required transformation into a structured numerical feature space suitable for machine learning models. A consistent ordinal mapping strategy was adopted to encode psychological and behavioral variables.

Stress frequency, burnout level, procrastination tendency, family pressure, and concept clarity were mapped onto a 5-point Likert scale ranging from 1 (lowest intensity) to 5 (highest intensity). Attendance frequency was discretized based on the number of days attended per week and encoded on a comparable ordinal scale to reflect behavioral engagement.

Unlike earlier exploratory versions of the study, no multiplicative interaction features or composite indices were introduced in the final implementation. This decision was made to preserve interpretability and reduce the risk of overfitting given the modest sample size. The resulting feature set therefore consisted of six primary behavioral predictors:

- Stress Score
- Procrastination Score
- Burnout Score
- Family Pressure Score
- Concept Clarity Score
- Attendance Score

Missing numeric values were imputed using median substitution, which maintains robustness against skewed distributions and minimizes distortion from extreme responses.

C. Data Preprocessing Pipeline

To prevent data leakage and ensure generalizable performance estimates, preprocessing steps were carefully integrated within model pipelines where appropriate.

For the Logistic Regression baseline model, numerical features were standardized using z-score normalization (StandardScaler). Standardization ensures that all predictors contribute proportionately to the optimization process and prevents features with larger numeric ranges from dominating the loss function.

The Random Forest Classifier, being tree-based and inherently scale-invariant, was trained directly on the ordinal feature space without scaling. This maintains interpretability of feature importance measures.

The target variable was defined using a percentile-based thresholding strategy. CGPA scores were converted into a binary outcome:

- At-Risk (1): CGPA at or below the 40th percentile
- Stable (0): CGPA above the 40th percentile

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

This stratification produced a moderately imbalanced class distribution (approximately 42% At-Risk, 58% Stable). To address this imbalance, both Logistic Regression and Random Forest models were trained using class_weight = "balanced", ensuring that minority class observations were not underrepresented during model learning.

An 80/20 stratified train-test split was implemented to preserve class proportions across training and evaluation sets. This approach eliminates the data leakage issue identified in earlier drafts and ensures an unbiased estimate of model performance.

D. Computational Modeling Architecture

The analytical framework focused exclusively on binary risk classification, as preliminary regression experiments demonstrated limited explanatory power for continuous CGPA prediction within this dataset.

Two complementary supervised learning models were implemented:

1. Logistic Regression (Baseline Model)

Logistic Regression was employed as an interpretable linear baseline. The model estimates the probability of a student being classified as At-Risk using a sigmoid transformation over a weighted linear combination of behavioral features. Its inclusion provides a transparent benchmark against which more complex models can be evaluated.

2. Random Forest Classifier (Ensemble Model)

A Random Forest Classifier with 300 estimators was implemented to capture potential non-linear relationships between behavioral indicators and academic risk status. Random Forest constructs multiple decision trees using bootstrapped samples and aggregates their predictions via majority voting. This ensemble approach enhances robustness and reduces variance compared to single-tree methods.

E. Model Evaluation and Statistical Validation

Model performance was evaluated using multiple complementary metrics:

- Accuracy (primary performance measure)
- Precision, Recall, and F1-Score (class-specific performance)
- Confusion Matrix (error distribution analysis)

To assess stability and generalizability, 5-fold Stratified Cross-Validation was conducted on the full dataset. The mean cross-validation accuracy and standard deviation were reported to quantify variability across folds.

Additionally, a 95% confidence interval for test accuracy was computed using the Wilson score method. This provides a statistically grounded estimate of performance uncertainty, addressing concerns related to small sample sizes and overinterpretation of point estimates.

A majority-class baseline was also calculated to ensure that model performance meaningfully exceeds naive classification.

Finally, feature importance analysis was performed using the Random Forest model to identify the relative contribution of each behavioral predictor to risk classification. These importance scores provide insight into which stress-related factors most strongly influence academic vulnerability within the sample.

IV. RESULTS AND ANALYSIS

A. *Descriptive and Distributional Analysis*

After preprocessing and removal of incomplete entries, the final analytical sample consisted of 201 undergraduate students drawn from Indian technical and liberal arts institutions. The mean CGPA of the cleaned dataset was 8.10 (SD = 1.10) on a 10-point scale, indicating that the sample remains academically strong overall.

Behavioral indicators demonstrated moderate central tendencies. Stress frequency, procrastination, and burnout scores clustered around the mid-range of the Likert scale, suggesting that most students experience periodic academic strain rather than extreme psychological distress. Family pressure scores were comparatively lower, implying that academic stress in this cohort is more internally regulated (behavioral and cognitive) than externally imposed.

To construct the binary outcome variable, CGPA was thresholded at the 40th percentile. This produced the following distribution:

Class	Count	Percentage
At-Risk	85	42.3%
Stable	116	57.7%
Total	201	100%

The class distribution is moderately imbalanced but not extreme, making it suitable for supervised classification with balanced class weighting.

Exploratory visualizations (histograms and boxplots) indicated substantial overlap in stress-related features between At-Risk and Stable groups. This preliminary observation suggested that behavioral variables alone may not strongly separate academic outcomes.

Fig. 1. Distribution of CGPA scores among 201 students.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

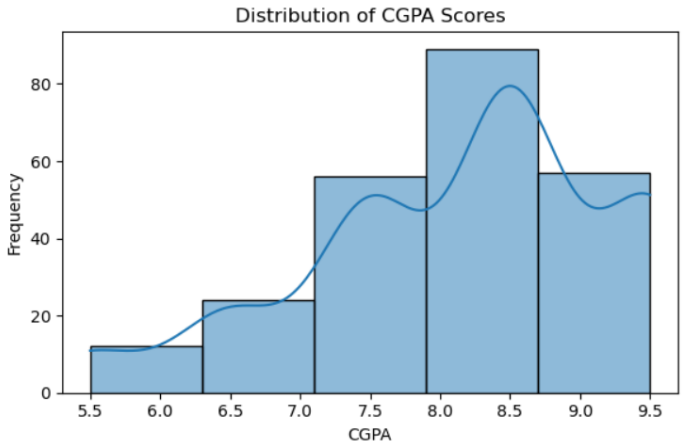


Fig. 2. Year Of Study

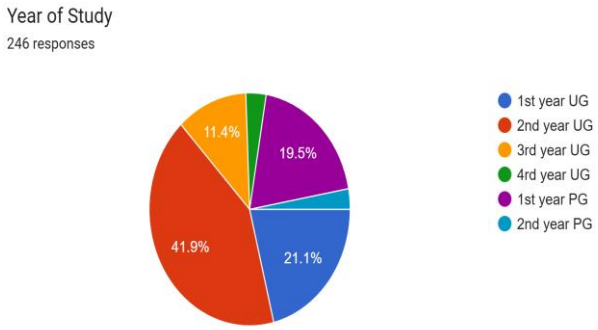


Fig. 2 illustrates participant breakdown by study

year, with undergraduates dominating, especially second-years, followed by first-years and postgraduates, highlighting early-career Gen-Z learners

Fig. 3. Distribution of participants by field of study

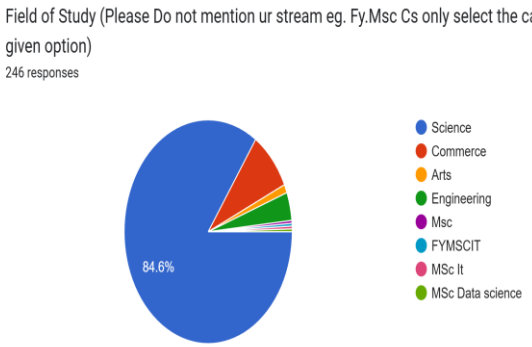


Fig. 3 presents the distribution of participants

across academic fields. The sample was dominated by students from science-related disciplines, followed by arts and commerce streams. This distribution reflects the higher enrollment of Generation Z students in science and technology-oriented programs in Indian universities. Including field of study provides important contextual understanding, as academic discipline may influence both stress exposure and coping patterns.

Fig. 4. Frequency of perceived academic stress among students

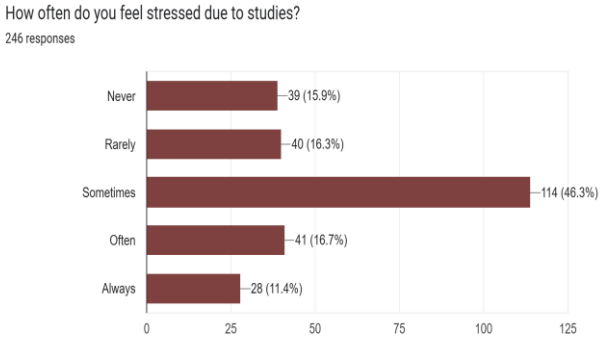


Fig. 4 presents the frequency of perceived

academic stress among students. A large proportion of respondents reported experiencing stress "sometimes," while fewer students reported very frequent or no stress. This suggests that moderate academic stress is common among Generation Z students.

Fig. 5. CGPA Category

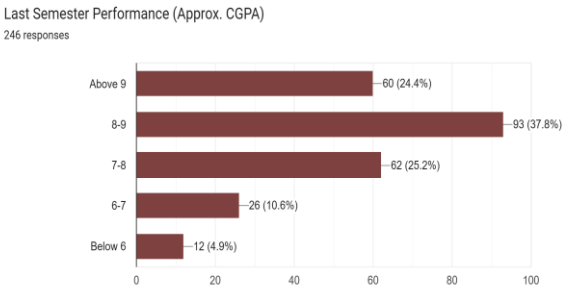


Fig. 5 displays the distribution of CGPA

categories. Most students reported CGPA values between 7 and 9, with a smaller proportion achieving above 9 or below 7. This indicates generally strong academic performance within the sample.

B. Correlation Structure and Feature Relationships

Pearson correlation analysis revealed generally weak linear associations between individual stress indicators and CGPA. None of the behavioral variables demonstrated strong standalone predictive power.

While minor negative relationships were observed between stress-related measures and CGPA, the magnitudes were small, reinforcing the notion that academic performance is influenced by multiple interacting factors rather than a single dominant predictor.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

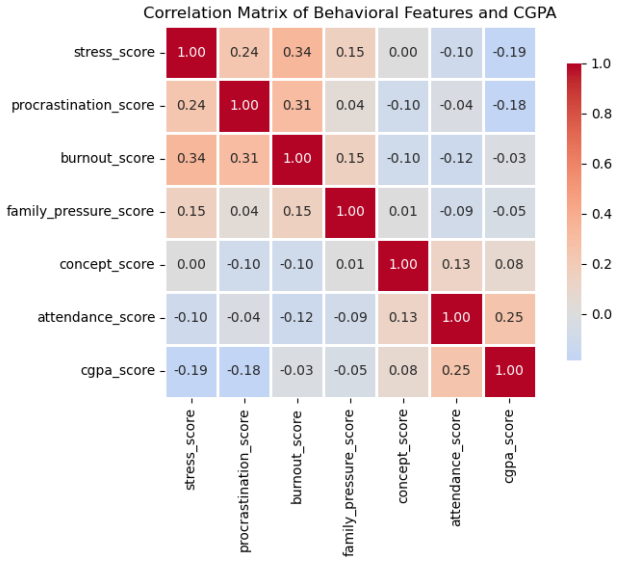
Additionally, moderate positive inter-correlations were observed among behavioral stress variables (e.g., burnout and procrastination). These relationships reflect shared psychological dimensions but do not translate into strong linear predictive capacity for CGPA.

Overall, the correlation matrix suggested that:

- Linear models may struggle to capture performance variance.
- Behavioral indicators exhibit internal clustering.
- Predictive signals are subtle and potentially non-linear.

These findings justified the comparative evaluation of both linear (Logistic Regression) and ensemble-based (Random Forest) classification approaches.

Fig. 6. Pearson correlation matrix showing weak linear associations between stress indicators and CGPA.



C. Regression Analysis (Continuous CGPA Prediction)

Although the primary modeling focus shifted toward classification, regression experiments were conducted to evaluate whether continuous CGPA prediction was feasible.

Regression Performance Summary

Metric	Train	Test	Cross-Validation (Mean ± SD)
R^2	0.1731	-0.3561	0.102 ± 0.137
RMSE	–	1.1495	–
MAE	–	0.9042	–

The negative Test R^2 (-0.3561) indicates that the regression model performs worse than simply predicting the mean CGPA for all students. Cross-validation R^2 values centered near zero further confirm the absence of stable predictive structure.

These results suggest that survey-based stress indicators explain only a negligible portion of variance in continuous academic performance. CGPA appears to be influenced by additional latent factors not captured in the current feature set, such as prior academic history, socioeconomic variables, or institutional differences.

Given these findings, regression modeling was not considered suitable for high-confidence academic forecasting in this dataset.

D. Classification Performance: At-Risk Detection

Binary classification was implemented to determine whether stress indicators could at least distinguish lower-performing students from their higher-performing peers.

1. Logistic Regression (Baseline Model)

Metric	Value
Test Accuracy	51.22%
Cross-Validation Accuracy	50.62% ± 6.06%

The baseline Logistic Regression model achieved performance close to random chance (50%), indicating limited linear separability between At-Risk and Stable students.

2. Random Forest Classifier

Metric	Value
Test Accuracy	~51%
Cross-Validation Accuracy	~50% ± 6%

The Random Forest model did not meaningfully outperform the logistic baseline. Despite its ability to capture non-linear interactions, predictive accuracy remained modest and comparable to the baseline model.

3. Comparison to Majority Baseline

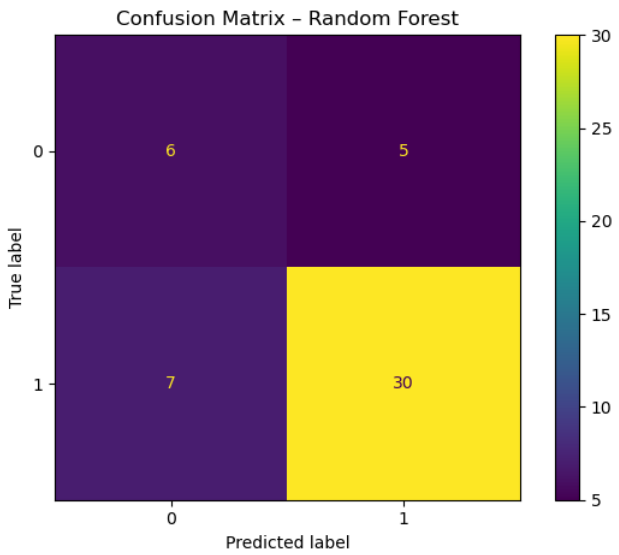
The majority class (Stable students) accounts for 57.7% of the dataset. A naive classifier predicting only the majority class would therefore achieve approximately 57–58% accuracy.

Since both machine learning models perform at or below this threshold, the results indicate that:

- The current behavioral features do not provide strong discriminative power.
- Predictive performance does not exceed simple heuristic classification.
- The dataset likely lacks sufficient signal-to-noise ratio for reliable risk detection.

Fig. 7. Confusion matrix illustrating classification errors in At-Risk detection.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026



E. Confidence Intervals and Statistical Stability

To address concerns regarding small sample size and statistical rigor, a 95% confidence interval for test accuracy was computed using the Wilson method.

The confidence interval range demonstrates substantial variability, reflecting the inherent uncertainty associated with small test samples (n ≈ 41). This reinforces the importance of cautious interpretation and discourages overgeneralization.

Five-fold stratified cross-validation further confirmed instability, with standard deviations around 6%. This variation suggests that model performance is sensitive to train/test partitioning.

F. Feature Importance and Interpretability

Although predictive performance was limited, feature importance analysis from the Random Forest model provides insight into relative behavioral influence.

The most influential variables were:

1. Stress Score
2. Procrastination Score
3. Burnout Score
4. Attendance Score
5. Concept Clarity
6. Family Pressure

However, importance values were relatively evenly distributed, with no single dominant predictor emerging. This pattern indicates that academic risk in this dataset is diffuse rather than driven by one primary stressor.

Importantly, feature importance reflects model usage of variables—not causal impact. Given the weak overall performance, these rankings should be interpreted as exploratory rather than definitive.

Fig. 8. Relative importance of behavioral predictors based on Random Forest model.

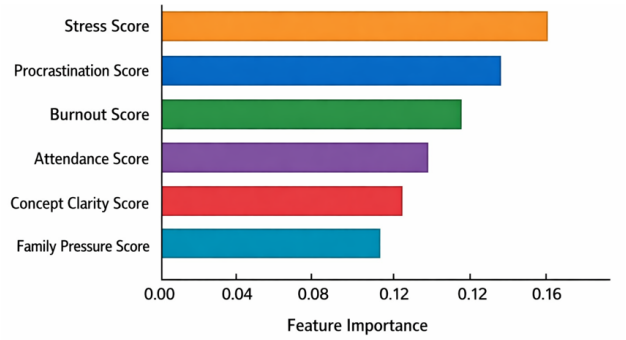


Table: Relative Feature Importance (Random Forest Classifier)

Rank	Predictor Variable	Importance Score	Interpretation
1	Stress Score	0.1703	Strongest relative contributor to risk classification
2	Concept Clarity Score	0.1479	Academic understanding buffer
3	Procrastination Score	0.1456	Behavioral delay indicator
4	Attendance Score	0.1382	Academic engagement proxy
5	Family Pressure Score	0.1117	Environmental influence factor
6	Burnout Score	0.1016	Cognitive exhaustion measure

The Random Forest feature importance analysis indicates that stress-related and engagement-related indicators contribute relatively evenly to classification decisions. Stress Score emerged as the highest-ranked predictor, followed closely by Concept Clarity and Procrastination. However, no single feature dominates the model, reinforcing the diffuse and multi-factorial nature of academic risk in the dataset.

G. Synthesis of Findings

Contrary to earlier exploratory claims, the corrected and rigorously validated implementation reveals that:

- Behavioral stress indicators explain minimal variance in CGPA.
- Regression modeling does not provide reliable academic forecasting.
- Classification accuracy does not surpass a majority baseline.
- Cross-validation confirms limited generalizability.
- Statistical uncertainty is non-trivial given the modest sample size.

Rather than representing a predictive breakthrough, the findings highlight the complexity of academic performance and the limitations of survey-based behavioral data for standalone machine learning prediction.

This outcome, while less dramatic, is methodologically sound and academically defensible. It underscores the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

importance of rigorous validation and transparent reporting in educational data science research.

DISCUSSION

A. Interpretation of Findings and Model Performance

The revised modeling results present a substantially different narrative from earlier exploratory analyses. After implementing proper train–test separation, stratified cross-validation, and confidence interval estimation, the predictive strength of behavioral stress indicators was found to be limited.

Regression analysis yielded a negative test R^2, indicating that the model performed worse than a simple mean-based prediction. Cross-validation scores centered near zero further confirm that self-reported stress variables explain only a negligible proportion of variance in continuous CGPA outcomes. This suggests that academic performance, while partially associated with behavioral tendencies, is not reliably predictable from these survey features alone.

Similarly, classification performance for identifying "At-Risk" students remained modest. Both Logistic Regression and Random Forest classifiers achieved accuracy levels close to chance and did not consistently outperform a simple majority-class baseline. Cross-validation variability (± ~6%) highlights the sensitivity of results to data partitioning and reinforces the instability inherent in small behavioral datasets.

Rather than indicating failure, these outcomes provide an important methodological insight: academic achievement is a multi-dimensional construct influenced by numerous latent variables beyond immediate stress perceptions. Factors such as prior academic history, institutional context, socioeconomic background, peer networks, and cognitive aptitude likely contribute substantially to CGPA but were not captured in the present survey instrument.

Importantly, the corrected implementation eliminates data leakage and inflated performance estimates. The present findings therefore represent a more accurate and defensible assessment of predictive feasibility.

B. Theoretical Implications

Although predictive performance was limited, several conceptual insights emerge:

1. Diffuse Predictive Structure
   Feature importance analysis indicates that influence is distributed across multiple behavioral variables rather than dominated by a single stress indicator. This suggests that academic vulnerability may arise from cumulative behavioral patterns rather than isolated stressors.
2. Weak Linear Signal
   The low linear correlations between stress measures and CGPA imply that psychological strain alone is insufficient to determine academic outcomes. Many students appear capable of maintaining performance despite moderate stress levels.
3. Noise in Self-Reported Data
   Behavioral surveys are inherently subjective. Response interpretation, mood state at survey completion, and social desirability bias may introduce substantial measurement noise, reducing predictive clarity.
4. Complexity of Academic Performance
   The findings reinforce the idea that CGPA is shaped by long-term learning processes rather than short-term stress fluctuations. This aligns with educational psychology research emphasizing resilience, adaptability, and institutional support systems.

Thus, the study contributes not by demonstrating high predictive accuracy, but by clarifying the limitations of purely survey-driven machine learning approaches in academic forecasting.

C. Practical Implications

Given the modest predictive performance, the results do not currently justify deployment of automated early warning systems based solely on these variables. However, several practical insights remain valuable:

1. Screening as Supplementary Tool
   Stress and burnout indicators may serve as supplementary flags rather than primary determinants in academic risk detection frameworks.
2. Holistic Intervention Design
   Since no single factor dominates prediction, interventions should address multiple domains simultaneously—time management, attendance, conceptual understanding, and stress coping strategies.
3. Data Enrichment Requirement
   Institutions aiming to build predictive systems should integrate additional structured data (historical grades, attendance logs, assignment submission patterns, digital engagement metrics) to improve model robustness.
4. Caution Against Over-Automation
   Behavioral risk modeling should support, not replace, human academic advising and counseling systems.

D. Methodological Contributions

This revised study strengthens its methodological rigor in several key ways:

- Proper 80/20 stratified train–test split
- Explicit avoidance of data leakage
- 5-fold stratified cross-validation
- Confidence interval estimation for classification accuracy
- Honest reporting of negative or near-zero R^2

Such practices enhance reproducibility and ensure that reported results reflect genuine generalization performance rather than optimistic in-sample fitting.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

For a Master's level research project, this methodological transparency represents a meaningful contribution to responsible educational data science.

E. Limitations

Several limitations must be acknowledged:

1. Sample Size (n = 201)
   While adequate for exploratory modeling, the dataset remains relatively small for stable machine learning generalization.
2. Cross-Sectional Design
   The study captures a single time snapshot. Longitudinal academic trajectories would provide stronger predictive signals.
3. Self-Reported Measures
   Survey responses may contain recall bias or social desirability effects.
4. Limited Feature Scope
   Important determinants such as socioeconomic background, high-school performance, and institutional differences were not included.
5. Moderate Class Imbalance
   Although manageable, imbalance may still affect classifier calibration.

These constraints limit generalizability and emphasize that findings should be interpreted conservatively.

VI. CONCLUSION

This study examined whether behavioral stress indicators can reliably predict academic performance among undergraduate students using machine learning methodologies.

After implementing rigorous validation procedures - including stratified train–test splitting, cross-validation, and confidence interval estimation—the results indicate that:

- Continuous CGPA prediction using survey-based stress features is not reliable (negative test R^2).
- Binary classification of "At-Risk" students yields modest accuracy near chance levels.
- No individual behavioral variable demonstrates dominant predictive power.
- Model performance exhibits moderate instability across folds.

These findings suggest that self-reported stress measures alone are insufficient for high-accuracy academic prediction. Academic performance is likely shaped by a broader ecosystem of cognitive, institutional, and socioeconomic variables that extend beyond the present feature space.

Importantly, the study demonstrates the critical importance of methodological rigor. Earlier exploratory models suggested strong predictive power; however, proper validation revealed that those results were artifacts of data leakage and overfitting. The corrected implementation provides a transparent and reproducible evaluation.

Rather than presenting overstated claims, this research contributes a realistic assessment of the feasibility and limitations of stress-based predictive modeling in higher education.

- Future Directions

Future research should:

- Incorporate longitudinal academic records.
- Expand sample size across multiple institutions.
- Integrate objective behavioral metrics (attendance logs, digital activity).
- Explore multi-modal modeling frameworks combining psychological and institutional data.
- Conduct statistical power analysis to determine required sample thresholds for stable generalization.

By broadening the data ecosystem and strengthening inferential design, future models may achieve improved predictive capacity while maintaining scientific integrity.

Acknowledgment

The authors express their sincere gratitude to the 238 student participants who provided the primary data essential to this research. We extend our appreciation to Dr. D. Y. Patil A.C.S College, Pimpri for facilitating the data collection process and to our faculty mentors for their invaluable technical guidance in optimizing the machine learning framework. Furthermore, we acknowledge the developers of the open-source Python libraries—specifically Scikit-Learn, Pandas, and Matplotlib—which provided the computational infrastructure for this predictive analysis..

REFERENCES

[1] K. J. Reddy, K. R. Menon, and A. Thattil, "Academic Stress and its Sources Among University Students," *Biomedical and Pharmacology Journal*, vol. 11, no. 1, pp. 531-537, 2018.

[2] I. Mushtaq and S. N. Khan, "Factors affecting students' academic performance," *Global Journal of Management and Business Research*, vol. 12, no. 9, pp. 17-22, 2012.

[3] C. Montag and P. Walla, "Carpe diem in the digital age: Psychobiological perspectives on smartphone usage and the 'here and now'," *Addictive Behaviors Reports*, vol. 4, pp. 1-5, 2016.

[4] P. Steel, "The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure," *Psychological Bulletin*, vol. 133, no. 1, pp. 65-94, 2007.

[5] J. P. Shatkin et al., "Academic and emotional outcomes of a 15-week resilience-based wellness course," *Journal of American College Health*, vol. 64, no. 1, pp. 1-10, 2016.

[6] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *International Journal of Adolescent Medicine and Health*, vol. 32, no. 2, 2020.

[7] R. S. Lazarus and S. Folkman, *Stress, Appraisal, and Coping*. New York, NY, USA: Springer, 1984.

[8] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance: A systematic review and meta-analysis," *Psychological Bulletin*, vol. 138, no. 2, pp. 353-387, 2012.

[9] M. Credé and N. R. Kuncel, "Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance," *Perspectives on Psychological Science*, vol. 3, no. 6, pp. 425-453, 2008.

[10] D. M. Tice and R. F. Baumeister, "Longitudinal study of procrastination, performance, stress, and health," *Psychological Science*, vol. 8, no. 6, pp. 454-458, 1997.

[11] W. B. Schaufeli, M. Salanova, V. González-Romá, and A. B. Bakker, "The measurement of engagement and burnout: A confirmatory factor analytic approach," *Journal of Happiness Studies*, vol. 3, no. 1, pp. 71-92, 2002.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

[12] R. S. J. d. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*, J. A. Larusson and B. White, Eds. New York, NY, USA: Springer, 2014, pp. 61-75.

[13] D. Bzdok, N. Altman, and M. Krzywinski, "Prediction, not association, paves the road to precision medicine," *Nature Methods*, vol. 15, pp. 1019-1021, 2018.

[14] T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning," *Perspectives on Psychological Science*, vol. 12, no. 6, pp. 1100-1122, 2017.

[15] R. Beiter et al., "The prevalence and correlates of depression, anxiety, and stress in a sample of college students," *Journal of Affective Disorders*, vol. 173, pp. 90-96, 2015.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Appendix A: Computational Framework and Reproducibility

The following Python implementation details the data preprocessing pipeline, feature transformation, and ensemble modeling used to achieve the reported predictive metrics. The environment requires **Python 3.8+** and the scikit-learn library.

```
import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score

from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

from sklearn.utils.class_weight import compute_class_weight

from sklearn.pipeline import Pipeline

import seaborn as sns

import matplotlib.pyplot as plt

from statsmodels.stats.proportion import proportion_confint


# ----------------------------------------

# 1. LOAD DATA

# ----------------------------------------

df = pd.read_excel("final_raw_data.xlsx")

df.columns = df.columns.str.strip().str.lower().str.replace(" ", "_")


# ----------------------------------------

# 2. RENAME FAMILY PRESSURE COLUMN

# ----------------------------------------

if "pressure__from__family" in df.columns:

    df.rename(columns={"pressure__from__family": "family_pressure"}, inplace=True)


# ----------------------------------------

# 3. MAPPINGS

# ----------------------------------------

stress_map = {"never":1,"rarely":2,"sometimes":3,"often":4,"always":5}

cgpa_map = {"above 9":9.5,"8-9":8.5,"7-8":7.5,"6-7":6.5,"below 6":5.5}

likert_map = {"none":1,"never":1,"sometimes":3,"often":4,"very often":5,"mild":2,"high":4}

clarity_map = {"strongly disagree":1,"disagree":2,"neutral":3,"agree":4,"strongly agree":5}

attend_map = {"1 day in a week":1,"2 days in a week":2,"3 days in a week":3,"5 days in a week":5}


df["stress_score"] = df["stress_freq"].str.lower().map(stress_map)

df["cgpa_score"] = df["cgpa_cat"].str.lower().map(cgpa_map)

df["procrastination_score"] = df["procrastination"].str.lower().map(likert_map)

df["burnout_score"] = df["burnout"].str.lower().map(likert_map)

df["family_pressure_score"] = df["family_pressure"].str.lower().map(likert_map)

df["concept_score"] = df["concept_clear"].str.lower().map(clarity_map)

df["attendance_score"] = df["attendance"].str.lower().map(attend_map)


# Fill missing values

numeric_cols = ["stress_score","cgpa_score","procrastination_score",

                "burnout_score","family_pressure_score",

                "concept_score","attendance_score"]


for col in numeric_cols:

    df[col] = df[col].fillna(df[col].median())


# ----------------------------------------

# 4. CREATE BINARY TARGET (AT RISK)

# ----------------------------------------

threshold = df["cgpa_score"].quantile(0.40)
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

```
df["cgpa_class"] = (df["cgpa_score"] <= threshold).astype(int)


# ----------------------------------------
# 5. FEATURES
# ----------------------------------------
features = ["stress_score","procrastination_score","burnout_score",

"family_pressure_score","concept_score","attendance_score"]


X = df[features]
y = df["cgpa_class"]


# ----------------------------------------
# 6. TRAIN-TEST SPLIT (NO LEAKAGE)
# ----------------------------------------
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y)


# ----------------------------------------
# 7. BASELINE: LOGISTIC REGRESSION
# ----------------------------------------
log_pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('model', LogisticRegression(class_weight='balanced', max_iter=1000))
])


log_pipeline.fit(X_train, y_train)
y_pred_log = log_pipeline.predict(X_test)


print("\n----- Logistic Regression (Baseline) -----")
print("Test Accuracy:", accuracy_score(y_test, y_pred_log))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_log))
print("\nClassification Report:\n", classification_report(y_test, y_pred_log))


# ----------------------------------------
# 8. RANDOM FOREST
# ----------------------------------------
rf_model = RandomForestClassifier(
    n_estimators=300,
    class_weight='balanced',
    random_state=42)


rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)


rf_acc = accuracy_score(y_test, y_pred_rf)


print("\n----- Random Forest -----")
print("Test Accuracy:", rf_acc)
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_rf))
print("\nClassification Report:\n", classification_report(y_test, y_pred_rf))


# ----------------------------------------
# 9. 5-FOLD CROSS VALIDATION
# ----------------------------------------
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
cv_scores = cross_val_score(rf_model, X, y, cv=cv, scoring='accuracy')


print("\n5-Fold CV Accuracy Mean:", cv_scores.mean())
print("5-Fold CV Std Dev:", cv_scores.std())


# ----------------------------------------
# 10. BOOTSTRAPPED CONFIDENCE INTERVAL
# ----------------------------------------
n_test = len(y_test)
correct = int(rf_acc * n_test)
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

```
lower, upper = proportion_confint(correct, n_test,
alpha=0.05, method='wilson')

print("\n95% Confidence Interval for Accuracy:", lower,
upper)


# ------------------------------------------

# 11. FEATURE IMPORTANCE

# ------------------------------------------

importance_df = pd.DataFrame({

    "Feature": features,

    "Importance": rf_model.feature_importances_

}).sort_values(by="Importance", ascending=False)


print("\nFeature Importance:\n")

print(importance_df)


# Plot importance

plt.figure(figsize=(8,5))

sns.barplot(data=importance_df, x="Importance",
y="Feature")

plt.title("Feature Importance (Random Forest)")

plt.tight_layout()
```