EXPLORING AND PREDICTING CGPA OF BANGLADESHI STUDENTS

Team 2

DOTAN KATZ, ORAN KEDEM, NOAM YEHOSHUA, BAR ELHAYANI

LINK TO CODE: GitHub

1. Introduction:

Predicting academic success lies at the heart of educational data-mining. Prior work shows that university records, especially past grades, are typically the strongest predictors, demographics add only marginal value when stronger features are present, and the impact of learning-behavior varies across studies, often depending on data quality and its link to learning effectiveness. To revisit these insights, we analyze a 2024 survey of 1,195 Computer-Science and Engineering undergraduates at a private Bangladeshi university and ask: how do academic, learning-behavioral, and demographic factors shape cumulative GPA (CGPA)?

Research questions: The general question is whether a student's CGPA can be predicted using features from different domains. To answer this, we analyze three feature groups separately:

- RQ1: Can CGPA be predicted from academic history?
- RQ2: Can CGPA be predicted from demographic background?
- RQ3: Can CGPA be predicted from learning-behavior patterns?

Examining demographic and academic variables tests known patterns on an additional population. Assessing the effect of learning habits challenges existing insights, as previous studies were able to predict academic performance using objective measures of study behavior or in combination with indicators of learning quality. In contrast, our data relies on self-reported features that may be biased, making this question difficult to answer. However, addressing it is important, as it could reinforce less-established insights and help universities identify at-risk students and provide targeted support based on information about their learning habits.

Hypotheses: Based on prior researches, we expect previous academic performance and institutional records to be the strongest predictors of CGPA. Demographic variables are likely to have low predictive value, and learning-habit measures are expected to have only a weak connection to CGPA, unless they prove to reflect learning quality.

Method overview: We will perform descriptive exploration, apply rigorous cleaning and feature engineering, build and compare interpretable predictive models, and use feature-importance techniques to isolate the most influential variables before interpreting the results against existing articles.

2. Data overview:

Our study uses the Academic Performance Dataset of Bangladeshi students, initially collected in 2021 by Shahariar Anwar through a self-administered Google Forms survey, and published in 2024. The dataset includes 1,195 undergraduate students in Computer Science and Engineering from a single private university, who began their studies between 2013 and 2023. It contains 31 self-reported variables, covering academic performance, learning habits, technology access, and demographic characteristics. Some fields are free-text.

Feature families for our research:

Academic records: CGPA (final cumulative GPA), SGPA (GPA from the most recent completed semester), attendance, scholarship, transport, probation/suspension.

Learning behavior: daily study hours and daily sessions, preferred learning mode, faculty consultation, co-curricular involvement.

Demographic and socio-economic: age, gender, family income, living arrangement, relationship status, health issues and disabilities.

The data collection method may introduces typical limitations of self-report: social desirability bias, memory inaccuracies, and inconsistent interpretation, which may distort student responses. Additionally, since the dataset is based on one institution and a narrow academic cohort from specific years, the findings limits both generalizability and comparison with other studies.

3. Methods and results:

This section presents the modelling process and findings: from cleaned and engineered data, we trained both linear and non-linear models to evaluate how academic, demographic, and learning-behavior features each contribute to predicting CGPA.

Data Processing: We performed data cleaning and preprocessing, including merging free-text values, removing outliers, applying transformations to skewed variables, and encoding CGPA both as a continuous variable and as grade categories. In addition, we engineered new features intended to reflect study effort and learning quality. Using this cleaned dataset, we built both linear and non-linear models to address the research questions.

Overall Model Benchmark: As a preliminary step, we combined all features from the three research questions and trained a Random Forest regressor, chosen for its robustness to outliers and ability to capture non-linear relationships. Compared to the linear baseline ($R^2 = 0.79$), the Random Forest achieved a higher R^2 of 0.87 on the test set and reduced RMSE by ~20%. Feature importance was dominated by SGPA, completed credits, and current semester. Attendance, study hours, and income had minor impact. These results serve as a benchmark for the upcoming per-group analysis, with similar CV and test results indicating no overfitting.

RQ-1: Pre-analysis of RQ1 confirmed prior findings: SGPA is the strongest correlate of final CGPA ($r \approx 0.65$). The non-linear random forest regressor outperformed its linear baseline (test $R^2 = 0.925$ vs. 0.87; RMSE = 0.216). SGPA contributed the highest feature importance (0.70), followed by completed credits (0.23) and current semester (0.11), while probation and scholarship had negligible impact. Prior studies similarly highlight the role of early academic performance, Yağcı (2022) notes that "Students' mid-term-exam grades are an important predictor to be used in predicting their final-exam grades" [1]. Academic history alone provides a strong predictive signal, reinforcing the value of systematic academic record tracking.

RQ-2: Pre-analysis showed that all demographic features had only weak correlations with CGPA ($|r| \leq 0.10$), with relationship status performing "best" at r = -0.10. As expected, models trained solely on demographic variables performed poorly: both linear regression and random forest reached $R^2 \approx 0.04$ and RMSE ≈ 0.77 on the 0–4 GPA scale, while an RF classifier applied to CGPA class grades (A–D) achieved just 33% accuracy, barely above random. Feature importance never exceeded 4%. English proficiency and relationship status emerged as the "strongest" predictors, yet had negligible impact. Prior work reports similar patterns: Asif et al. (2017) note that "only admission marks and final marks are used; no socio-economic or demographic features are considered." [2]. In short, as the literature also suggests, demographic data alone are insufficient for prediction and sometimes add little value when stronger academic indicators are available.

RQ-3: In this question, we encountered limitations with study-behavior features due to their clustered distributions. For example, 65% of students report attendance ≥ 90%, 80% study 1–4 hours per day, and 95% report studying 1–3 times daily. These narrow ranges leave little variance for the models to learn from. In addition, correlations with CGPA were weak (all $|r| \leq 0.12$).

Model performance was weak across both linear and non-linear approaches. The linear model outperformed the random forest, despite the non-linearity of the features, indicating overall poor predictive power. On the cleaned from outliers dataset, the best R^2 was only 0.03 (compared to 0.01–0.03 on the full set), with RMSE ≈ 0.47. An RF classifier applied to CGPA grade classes reached only 32% accuracy, barely above random. This pattern remained consistent even after removing outliers, suggesting that the main limitation lies in the lack of

feature variability rather than noise. Attendance and study hours consistently topped the RF importance list, but each contributed less than 15% of the total gain (Table 1).

Table 1: Learning-behavior model

Task & Target (scale)	Approach	Best Model	Test Metrics	Top-3 Predictors
CGPA (0–4)	Regression	Linear Regression	$R^2 = 0.03$ RMSE = 0.46	Attendance, study_hours, study_freq
CGPA bands (A–D)	Classification	Random-Forest Classifier	Accuracy ≈ 0.325 F1 ≈ 0.322	attendance, study_hours, study_freq

Sub-group tests: We applied the same modeling pipeline separately to demographic sub-groups, but the improvements were minimal. Even the best-performing group, male students, explained less than 4% of the variance in CGPA ($R^2 = 0.034$) and reached classification accuracy of only 30–33%. All other groups showed similarly low results, with classification performance only slightly better than random guessing (Figures 1 and 2).

Figure 1: Sub-group accuracy (CV)

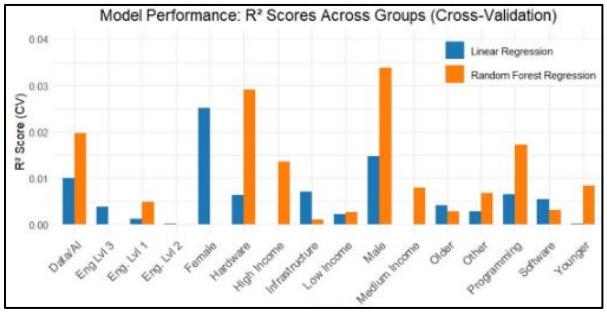
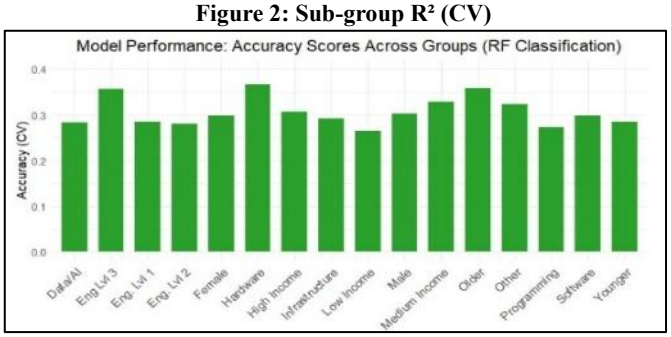


Figure 2: Sub-group R^2 (CV)



Prior findings show that predictive value of study-behavior often hinges on the use of validated inputs and their alignment with effective learning. Plant et al. (2005) likewise observed that "the amount of study only emerged as a significant predictor of cumulative GPA when the quality of study and previously attained performance were taken into consideration." [3]. High-performing studies that rely on objective LMS logs or verified attendance achieve classification accuracies of 0.80 – 0.90, whereas self-reports do not [4, 5].

Main Conclusion: Self-reported learning-habit data, even when cleaned and engineered, cannot meaningfully predict CGPA in this cohort. Without validated learning-behavior or quality indicators, such models cannot reliably predict academic performance.

Limitations and Future Work:

Key Limitations:

- **Self-report bias**: The data is entirely self-reported, with no external validation, and includes free-text fields prone to bias and interpretation errors. While it captures time spent on learning, it lacks verified indicators of learning quality and shows low variance across key features, limiting its reliability for modeling effort and outcomes.
- **Limited sample**: The dataset covers a concentrated group of students from one institution, specific disciplines, and a limited range of enrollment years, which restricts generalizability and limits cross-study comparisons.

One-month plan: Engineer richer features using existing variables (e.g., combining study hours with faculty consultation), and apply transformations to increase variance in clustered fields. Review similar studies with self-reported data or Bangladeshi cohorts to better contextualize limitations.

Three-month plan: Analyze local curriculum and learning context to validate feature plausibility. Run subgroup models to detect high-risk groups and refine support strategies and compare findings with external datasets to assess generalizability.

Appendix:

Link to our GitHub repository:

https://github.com/NoamYehoshua/cgpa-prediction-bangladesh

Link to the data:

https://data.mendeley.com/datasets/dc3797vf3t/1

Bibliography:

[1] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine-learning algorithms," Int. J. Educ. Technol. High. Educ., vol. n/a, no. n/a, pp. n/a, 2022.

Available: https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z

[2] R. Asif, A. Merceron, N. Ali, and Z. A. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177-194, 2017.

Available: https://www.sciencedirect.com/science/article/abs/pii/S0360131517301124

[3] E. A. Plant, K. A. Ericsson, L. Hill, and K. Asberg, "Why study time does not predict grade-point average across college students: Implications of deliberate practice for academic performance," Contemp. Educ. Psychol., vol. 30, no. 1, pp. 96-116, 2005.

Available: https://www.sciencedirect.com/science/article/abs/pii/S0361476X04000384?via%3Dihub

[4] K. Tao, Z. Liu, W. Zhang, and H. Zhang, "Deep neural network-based prediction and early warning of student grades and recommendations for similar learning approaches," Appl. Sci., vol. 12, no. 15, Art. no. 7733, 2022.

Available: https://www.mdpi.com/2076-3417/12/15/7733

[5] M. Riestra-González, M.-P. Paule-Ruíz, and F. Ortín, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," Comput. Educ., vol. 163, Art. no. 104108, 2021.

Available: https://doi.org/10.1016/j.compedu.2020.104108