International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Student Performance Prediction with Machine Learning: An Exploratory Regression Experiment

[1]Suraj Kondekar, [2]Sohan Chaudhary, [3]Sumit Dubey

Abstract - *Early identification of students who are likely to score poorly can help institutions offer timely support. This paper reports an exploratory regression experiment on how well a student's most recent examination percentage can be predicted from self-reported academic, behavioural and psychological factors. An online questionnaire produced 42 complete responses on 18 predictors, which were encoded and analysed with correlation tests and eight regression models (Linear, Ridge, Lasso, Decision Tree, k-NN, Support Vector Regression, Random Forest and Gradient Boosting) compared with a mean-only baseline using 5×5 repeated cross-validation. Only self-rated academic performance ($r = 0.41$, $p = 0.007$) and previous semester score ($r = 0.32$, $p = 0.040$) were significantly correlated with the recent score, and neither survives a multiple-comparison correction. Linear models did not beat the baseline (cross-validated R^2 from −1.70 to −0.15). The best model, Support Vector Regression, reached $R^2 = 0.24$, MAE = 5.72 and RMSE = 7.71 percentage points, against a baseline MAE of 6.73 and RMSE of 9.06. Without the previous score, no model beat the baseline. The evidence points to modest predictive signal in self-reported data at this sample size; it does not show that any habit causes better or worse results.*

Keywords - Student performance prediction, educational data mining, machine learning, regression, support vector regression, cross-validation, self-reported study habits

I. INTRODUCTION

Academic performance influences scholarships, placements, further study and the standing of institutions. Teachers often learn that a student is struggling only after an examination has been marked. Educational data mining aims to change this by predicting outcomes early enough for an intervention to matter [1], [2].

Most published work uses institutional records or learning-platform logs. Far fewer studies use what students themselves report about how they study: hours of study, sleep, mobile and social-media use, revision habits, concentration and confidence. These answers can be collected cheaply with a short questionnaire, and many of them are habits a student can change. This paper describes how well several standard regression models predict the most recent examination percentage from such a questionnaire.

II. PROBLEM STATEMENT

Predicting a student's examination result early would let teachers offer support before marks are published. Existing prediction systems usually depend on institutional records or learning-platform logs [1], [3], [4], which an individual college may not hold in usable form. A short self-reported questionnaire is far cheaper to collect, but it is not known whether it holds enough information to predict a student's examination percentage better than simply guessing the class average.

The problem addressed here is therefore: can 18 self-reported academic, behavioural and psychological factors predict a student's most recent examination percentage more accurately than a mean-only baseline, and which of them carry the signal? The study examines this on primary data from 42 students.

III. LITERATURE REVIEW

Romero and Ventura [1] surveyed educational data mining and identified performance prediction as one of its most common applications; Baker and Yacef [2] described the field as a maturing discipline using classification, regression, clustering and relationship mining. Cortez and Silva [3] predicted the grades of Portuguese secondary-school students with decision trees, random forests, neural networks and support vector machines, and found that earlier grades were by far the strongest predictors. Kuzilek et al. [4] released the Open University Learning Analytics Dataset, which combines demographics, assessments and click-stream data.

Evidence on behavioural factors is mixed. Meta-analytic work links class attendance to grades [5] and study habits and skills to college GPA [6], but Plant et al. [7] showed that the amount of study time alone is a weak predictor, because the quality of study matters. On the methods side, Ridge [8] and Lasso [9] regularise linear models when predictors are many relative to observations, while Random Forests [10] and kernel methods such as support vector regression handle non-linear patterns and are often reported to work well on small tabular data [11]. With little data, repeated cross-validation is preferred to a single split, although estimates stay noisy [12].

IV. RESEARCH GAP

Most earlier studies rely on institutional records or

platform logs, and fewer use short self-reported questionnaires that any college could run. Many studies also report a single favoured model rather than comparing linear, regularised, tree-based, kernel and ensemble regressors against a naive baseline under the same validation scheme. Finally, behaviour at very small sample sizes, which is what an individual college will usually have, is under-reported.RESEARCH QUESTIONS

- **RQ1.** Which self-reported factors are most strongly associated with the most recent examination percentage?
- **RQ2.** How accurately can regression models predict that percentage compared with a naive baseline?
- **RQ3.** Do non-linear models outperform linear models on this dataset?
- **RQ4.** How much of the predictive signal comes from the previous semester score rather than behavioural factors?

V. RESEARCH OBJECTIVES

- To collect primary data on study habits, lifestyle and academic outcomes through a structured questionnaire.
- To clean, harmonise and encode the responses, including CGPA-to-percentage conversion.
- To measure the association between each factor and the recent examination percentage.
- To compare eight regression algorithms with repeated cross-validation using RZ, MAE and RMSE.
- To rank features by importance and test sensitivity to the previous score, an extreme record and the CGPA conversion factor.

VI. HYPOTHESES

- **H1.** Previous semester score is positively correlated with the recent score.
- **H2.** Self-rated academic performance is positively correlated with the recent score.
- **H3.** Students who study longer score higher.
- **H4.** Students with more social-media time score lower.
- **H5.** Students with higher attendance score higher.
- **H6.** At least one non-linear model achieves a positive cross-validated RZ.

All tests use a 5% significance level.

VII. RESEARCH METHODOLOGY

A. Design and data collection

A quantitative, cross-sectional, exploratory design was adopted. An online questionnaire was circulated to students from 7 October 2026 using convenience sampling and participation was voluntary. The final dataset holds 42 complete responses with no missing values.

Methodological note: this is an exploratory study. With 42 respondents, its results describe this sample and must not be read as evidence that any habit causes a change in examination results.

B. Variables and pre-processing

The target was the most recent examination percentage. The

18 predictors were previous semester score, attendance band, daily study hours, assignment punctuality, class-attendance frequency, class participation, sleep hours,

social-media hours, revision frequency, practice of past papers, use of online resources, weekly hours on additional courses, study environment, concentration, distraction frequency, confidence, help-seeking and self-rated performance. Age, gender, level, year and stream were not modelled because their categories are too sparse for 42 rows.

Some respondents entered a CGPA instead of a percentage (14 previous and 12 recent values were 10 or below). These were multiplied by 9.5, a conversion commonly used by Indian universities; a ×10 alternative was tested. Banded answers were replaced by band mid-points (for example “1–2 hours” became 1.5), frequency scales were coded 1 to 5, and standardisation was applied inside each model pipeline so that no information leaked from validation folds.

C. Models and validation

Eight models were trained: Linear Regression; Ridge (α = 10); Lasso (α = 1); Decision Tree (depth 3); k-NN (k = 5); SVR (RBF kernel, C = 50, ε = 1); Random Forest (300 trees, depth 4); and Gradient Boosting (100 trees, depth 2, learning rate 0.05). A mean-only regressor served as baseline. Hyper-parameters were fixed in advance and not tuned. Evaluation used five-fold cross-validation repeated with five shuffles, reporting mean RZ, MAE and RMSE in percentage points; leave-one-out validation was a further check. A negative RZ means a model predicts worse than always guessing the sample mean.

VIII. DATA ANALYSIS: TOOLS AND TECHNOLOGY

The analysis used Python 3 with pandas and NumPy for data handling, SciPy for correlation, t-tests and ANOVA, and scikit-learn [13] for pipelines, models, cross-validation and feature importance. Fixed random seeds were used throughout. Data came from a CSV export of the questionnaire responses (42 rows × 25 columns).

IX. MAJOR FINDINGS

1. Only self-rated performance (r = 0.41) and previous score (r = 0.32) were significantly correlated with the recent score; all other predictors had $|r| \le 0.22$ and $p > 0.05$.
2. Linear models, including regularised ones, did not beat the baseline.

3. Support Vector Regression was best (RZ = 0.24, RMSE 7.71), cutting RMSE by about 15% relative to the baseline.
4. Previous score carried 47% of Random Forest importance and self-rated performance 13%.
5. Without the previous score, no model beat the baseline.
6. Study hours, social-media use and attendance showed no significant relationship with the recent score.
7. Results were sensitive to one extreme record.

X. RESULTS AND ANALYSIS

A. Respondent profile and descriptive statistics

Of the 42 respondents, 25 (59.5%) were male, 16 (38.1%) female and one preferred not to say. Most were aged 21–23

(25, or 59.5%), as shown in Fig. 1. Postgraduate students formed 64.3% of the sample and Computer Science/IT students 59.5%, so the sample is not representative of all students.

TABLE I. SAMPLE PROFILE (N = 42)

Category	Respondents	Percentage
Gender: Male	25	59.5%
Gender: Female	16	38.1%
Gender: Prefer not to say	1	2.4%
Level: Postgraduate	27	64.3%
Level: Undergraduate	13	31.0%
Level: Diploma	2	4.8%
Stream: Computer Science / IT	25	59.5%
Stream: Science	6	14.3%
Stream: Commerce	5	11.9%
Stream: Management	3	7.1%
Stream: Other	2	4.8%
Stream: Engineering	1	2.4%

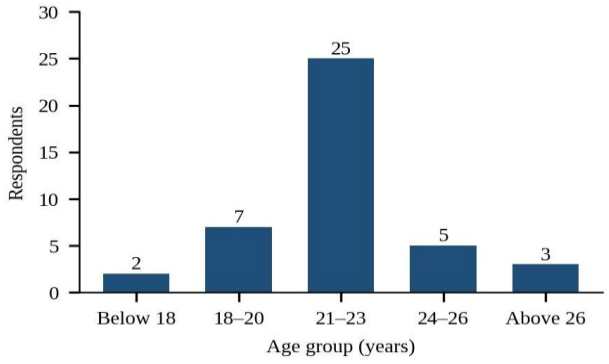

Fig. 1. Respondents by age group (n = 42).

The mean recent score was 77.02% (SD 8.99) and the mean previous score 75.12% (SD 8.31). The mean change of

+1.90 points was not significant (paired $t(41) = -1.22$, $p = 0.230$), and the recent scores did not depart from normality (Shapiro–Wilk $p = 0.419$).

TABLE II. DESCRIPTIVE STATISTICS

Variable	Mean	SD	Min	Max
Recent exam (%)	77.02	8.99	59.00	100.00
Previous semester (%)	75.12	8.31	50.00	89.32
Attendance (%)	75.24	15.22	45	95
Study (h/day)	1.86	1.14	0.5	4.5
Sleep (h/day)	6.48	1.00	4.5	8.5
Social media (h/day)	2.71	1.30	0.5	4.5
Self-rated performance (1–5)	3.21	1.07	1	5

B. Association with the recent score (RQ1)

Fig. 2 shows the correlation of each predictor with the recent score and Table III lists the strongest. Only self-rated

performance and previous score were significant at 5%. Because 18 correlations were examined without correction, even these are tentative: after a Bonferroni adjustment ($p \times 18$) neither remains below 0.05. The small positive coefficient for distraction, coded from Never (1) to Very often (5), is not significant and should not be read as distraction helping performance.

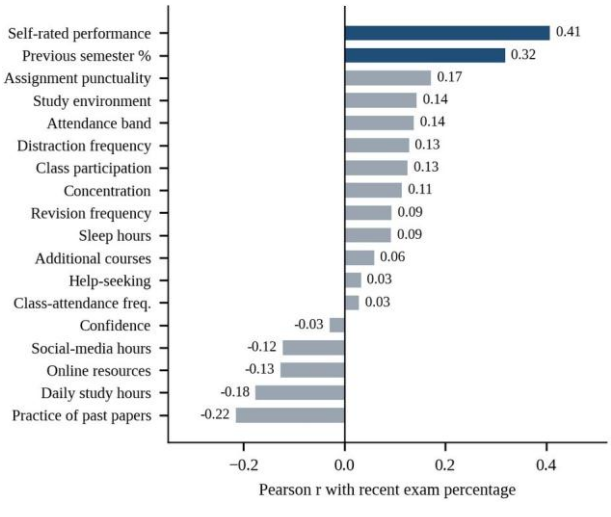

Fig. 2. Pearson correlation of each predictor with the recent exam percentage (dark bars: $p < 0.05$).

TABLE III. STRONGEST CORRELATIONS WITH RECENT SCORE (R: PEARSON, P: SPEARMAN)

Predictor	r	p	ρ	p
Self-rated performance	0.408	0.007	0.527	<0.001
Previous semester score	0.318	0.040	0.507	0.001
Practice of past papers	−0.216	0.170	−0.228	0.147
Daily study hours	−0.176	0.264	−0.288	0.065
Assignment punctuality	0.172	0.277	0.269	0.085
Study environment	0.143	0.366	0.260	0.097

C. Hypothesis tests

TABLE IV. SUMMARY OF HYPOTHESIS TESTS

ID	Result	Decision
H1	r = 0.318, p = 0.040	Supported
H2	r = 0.408, p = 0.007	Supported
H3	≥2.5 h/day (n = 13) mean 75.2 vs. 77.9 (n = 29); Welch p = 0.395; opposite direction	Not supported
H4	≥3.5 h/day (n = 16) mean 75.1 vs. 78.2 (n = 26); Welch p = 0.244	Not supported
H5	≥75% (n = 31) mean 77.4 vs. 75.9 (n = 11); Welch p = 0.683	Not supported
H6	SVR RZ = 0.24, RF 0.12, GB 0.05, k-NN 0.04	Supported

Students who studied under one hour a day (n = 9) averaged 83.3%, which probably reflects already-strong students who need less time, or differences in how the question was interpreted. Group sizes of 2–9 make these means unstable. An ANOVA across the five self-rating categories gave F = 2.33, p = 0.074.

D. Model comparison (RQ2 and RQ3)

TABLE V. CROSS-VALIDATED MODEL PERFORMANCE (5×5-FOLD)

Model	R^2	MAE	RMSE
Baseline (mean)	−0.042	6.73	9.06
Previous score only	−0.154	6.29	9.51
Linear Regression	−1.700	10.86	14.54
Ridge	−0.452	7.46	10.69
Lasso	−0.213	6.69	9.76
Decision Tree	−0.396	7.03	10.45
k-NN	0.039	6.36	8.69
SVR	0.244	5.72	7.71
Random Forest	0.117	5.67	8.34
Gradient Boosting	0.045	5.86	8.66

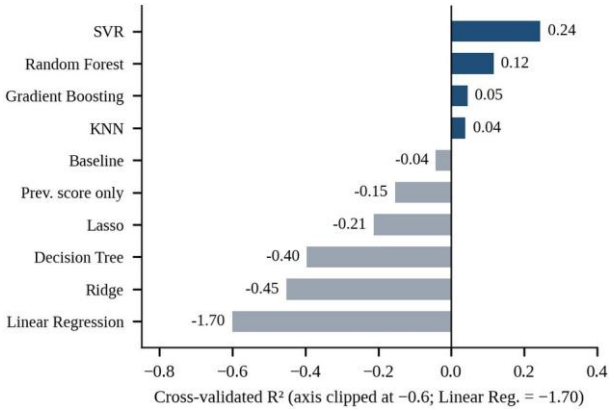

Fig. 3. Cross-validated RZ of each model (blue: better than baseline).

Unregularised Linear Regression performed worst because 18 predictors with 42 observations cause severe over-fitting. Ridge and Lasso were much better but still did not beat the baseline. SVR was best on RZ and RMSE, and Random Forest was marginally best on MAE. Leave-one-out validation agreed: SVR RZ = 0.34, Random Forest 0.13, Ridge −0.54 and baseline −0.05. The best average error of about 5.7 points is still large against a score standard deviation of 9 points.

E. Role of the previous score (RQ4)

TABLE VI. CROSS-VALIDATED RZ WITH AND WITHOUT PREVIOUS SCORE

Model	With	Without
Linear Regression	−1.700	−2.164
Ridge	−0.452	−0.519
Random Forest	0.117	−0.171
Gradient Boosting	0.045	−0.283

Without the previous score no model beat the baseline, so behavioural and psychological self-reports alone carried essentially no usable signal here. Fig. 4 confirms this: previous score (0.468) and self-rated performance (0.128) dominate Random Forest importance. Self-rating is plausibly a proxy for earlier results rather than an independent factor.

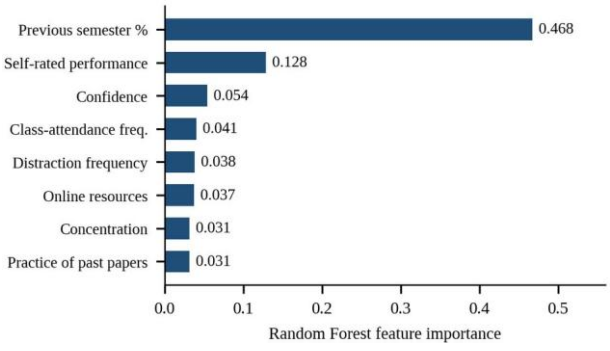

Fig. 4. Random Forest feature importance (top eight of 18).

F. Sensitivity analyses

TABLE VII. SENSITIVITY OF CROSS-VALIDATED RZ

Scenario	Ridge	RF	SVR
Main analysis (n = 42)	−0.452	0.117	0.244
Reduced 8 features	−0.085	0.194	0.208
Excluding extreme record (n = 41)	0.167	0.479	0.326
CGPA factor ×10 instead of ×9.5	—	0.186	0.230

One respondent had a previous score of 50% and a recent score of 100%. Excluding this record raised Random Forest RZ from 0.12 to 0.48 (baseline −0.03), so a single case can change how predictable performance appears. The record is not obviously invalid, so it was kept. The reduced feature set was chosen after viewing correlations on the full data, so its figures are mildly optimistic. Changing the CGPA conversion factor made little difference.

G. Discussion

Like Cortez and Silva [3], we find that the previous result is the most informative predictor. Our non-significant findings for attendance and study time differ from studies reporting strong effects [5], [6], probably because the sample is small, the predictors are coarse bands and self-reported hours are noisy; they are consistent with the caution in [7] that study time alone says little. The models could rank students roughly, but should not be used alone to make decisions about individual students.

XI. CONCLUSION

On the evidence of 42 survey responses, previous semester score and self-rated performance are the only factors associated with the recent examination percentage, and even these associations are tentative after correcting for multiple tests. Linear models failed to beat a mean-only baseline, while non-linear models, particularly SVR (RZ = 0.24, RMSE 7.71) and Random Forest (MAE 5.67), gave modest improvements. Without the previous score, no model beat the baseline. The results offer an exploratory snapshot and a reproducible template, not a deployable predictor. Testing these ideas properly will need larger samples and better measurement.

XII. LIMITATIONS

- Small sample (n = 42) with 18 predictors, so estimates are highly variable and non-significant results do not prove the absence of an effect.
- Convenience sampling, mostly postgraduate Computer Science/IT students from one region; some groups have only 1–2 students.
- Self-reported data are exposed to recall error and social-desirability bias.
- Mixed CGPA and percentage entries; the conversion factor is an assumption and the grading scheme was not recorded.
- Band mid-point encoding discards within-band variation and treats ordinal scales as interval.
- Habits were reported at survey time, so causal direction cannot be established.
- No hyper-parameter tuning, no nested cross-validation, and uncorrected multiple testing.

XIII. FUTURE SCOPE

- Collect a larger, more diverse sample across institutions and record the grading scheme.
- Use objective data such as attendance registers, LMS logs and submission times alongside self-reports.
- Collect longitudinal data (habits at the start of the semester, results at the end) to separate prediction from correlation.
- Reframe the task as classification of at-risk students and report precision and recall.
- Apply nested cross-validation with tuning, test further models and add explainability tools such as SHAP.
- Check fairness across gender, stream and background before any deployment.

REFERENCES

[1] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601–618, 2010.

[2] R. S. J. d. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” J. Educ. Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[3] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” in Proc. 5th Future Bus. Technol. Conf. (FUBUTEC), 2008, pp. 5–12.

[4] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open University Learning Analytics dataset,” Sci. Data, vol. 4, Art. no. 170171, 2017.

[5] M. Credé, S. G. Roch, and U. M. Kieszczynski, “Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics,” Rev. Educ. Res., vol. 80, no. 2, pp. 272–295, 2010.

[6] M. Credé and N. R. Kuncel, “Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance,” Perspect. Psychol. Sci., vol. 3, no. 6, pp. 425–453, 2008.

[7] 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.

[8] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,”

Technometrics, vol. 12, no. 1, pp. 55–67, 1970.

[9] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. R. Stat. Soc. B, vol. 58, no. 1, pp. 267–288, 1996.

[10] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[11] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

[12] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” BMC Bioinformatics, vol. 7, Art. no. 91, 2006.

[13] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.