

Non-Academic Interests and Competency Prediction: A Proof-of-Concept Multi-Reference ML Pipeline for Student Profiling

Pham Ngoc Thanh¹, Nguyen Khanh Toan², Trinh Thanh Do³

¹Nguyen Tat Thanh High School, Hanoi, Vietnam

Email: kukri123thanh@gmail.com

²VinUniversity, Hanoi, Vietnam

Email: 26ai.toannk3@vinuni.edu.vn

³Rikai Technology, Danang, Vietnam

Email: Dotrinh.cs@gmail.com

*Corresponding author

ABSTRACT: Existing models for academic prediction oversimplify non-academic interests and rely on single data sources. This proof-of-concept presents a multi-reference machine-learning pipeline predicting four student competencies using Likert-scale interest profiles from four distinct sources. Evaluated on a reproducible synthetic cohort ($n=390$), the protocol tests feature ablation, three source-fusion architectures, inter-source disagreement, and SHAP explanation stability. Results under this generative baseline were predominantly null: adding interest features, multiple sources, or disagreement metrics yielded no significant predictive improvement over an academic baseline. Furthermore, gated weighting proved the weakest fusion strategy, and SHAP rankings lacked stability in complex configurations. Ultimately, this study contributes a fully specified, end-to-end reproducible analysis protocol ready for empirical application, rather than a deployable predictive model.

Keywords: Student Competency Prediction; Non-Academic Interests; Multi-Source Machine Learning; Data Fusion; SHAP; Student Profiling; Proof Of Concept; Synthetic Cohort.

1. INTRODUCTION

Predicting student academic outcomes with machine-learning techniques has become a highly active line of research in Educational Data Mining. Recent work reaches very high accuracy when predicting grades from learning-behaviour data: a 2025 study in Scientific Reports reports a coefficient of determination R^2 of approximately 0.989 by combining ten regression models under an Ensemble Voting architecture on two datasets of 10,000 and 6,607 records [1]. Notably, the LIME- and SHAP-based explanation analysis in that same study identifies extracurricular participation as one of the most important predictors of academic success, alongside sleep hours and study preparation [1]. Similarly, a 2024 study in Frontiers in Psychology combining chi-square testing with XGBoost confirms the role of non-academic factors such as metacognitive awareness, learning motivation, and time management in academic performance [2].

A review of the literature nevertheless reveals three substantial research gaps. First, non-academic interests — sports, arts, strategy games, social activities — are almost always encoded as binary participation indicators, discarding information about the intensity and structure of a student's interest profile. Second, the target variable in most studies is a single academic score, whereas modern educational goals emphasise multi-dimensional competencies such as critical thinking, creativity, and communication. Third, input data typically come from a single source (student self-reports or a learning-management system), ignoring the reality that different reference sources — teachers, parents, observable behaviour — provide

complementary and sometimes conflicting views of the same student.

Motivated by these three gaps, this paper develops a multi-reference machine-learning framework for student profiling and takes it through a proof-of-concept evaluation. The aim at this stage is not to estimate how strongly interests predict competency in any real population, but to establish that a design of this complexity can be specified in full, executed end to end, and interrogated by its own diagnostic criteria before it is taken to an empirical cohort. The framework is organised around three research questions:

- (1) To what extent does adding non-academic interest features to an academic-only model improve competency prediction?
- (2) Which architecture for combining reference sources is optimal, and does the degree of disagreement between sources itself carry predictive information?
- (3) Which interests contribute to which competencies, and are the resulting explanations stable and consistent with existing educational theory?

The main contributions of this proof-of-concept study are: (i) a four-configuration ablation design that quantifies the incremental value of non-academic interests for multi-dimensional competency prediction; (ii) a controlled comparison of three fusion architectures together with an explicit test of whether inter-source disagreement carries incremental predictive information, a question that to the best of our knowledge has not previously been posed directly in educational data mining; and (iii) a SHAP-based interest-competency contribution matrix subjected to cross-fold stability checks, with interpretation restricted to the configurations that

pass them; and (iv) a released generator, dataset and analysis script that make every reported number reproducible from a single seed, so that what is contributed is an executable protocol rather than a description of one.

2. RELATED WORK

2.1. Machine learning for academic performance prediction

Gradient-boosted decision-tree models such as XGBoost [3] and CatBoost [4] are currently the dominant choice for tabular educational data thanks to their handling of mixed feature types and strong resistance to overfitting. Architectures that combine multiple base models under the stacked-generalization principle [5] are also widely applied: study [1] shows that an Ensemble Voting Regressor combining the five best base models achieves the lowest error (MAE 0.0837; RMSE 0.1050) on a 10,000-record dataset. Beyond the choice of algorithm, the systematic review of Alyahyan and Düşteğör [12] frames educational prediction as an end-to-end process spanning data collection, preparation, preprocessing, modelling and evaluation, and shows that decisions taken along that pipeline account for much of the variation in published accuracy. Prediction targets also extend beyond grades: Xing and Du [13] used a temporal deep-learning model to generate individual dropout probabilities in MOOCs for prioritised intervention. These studies, however, focus on optimising grade-prediction accuracy and do not address multi-dimensional competencies.

2.2. Non-academic factors in educational prediction

Evidence for the role of non-academic factors continues to accumulate. Beyond the extracurricular-participation finding in [1], study [2] identifies six groups of non-academic factors with significant effects on academic performance, including motivation, engagement, and mental health. Meta-analytic evidence points in the same direction for adjacent constructs. Poropat [15], pooling cumulative samples in excess of 70,000, found Agreeableness, Conscientiousness and Openness to be significantly associated with academic performance, with age and educational level acting as moderators. Ergen and Kanadlı [16] reported a large random-effects association between self-regulated learning strategies and achievement ($d = 0.859$) across the Turkish studies included in their meta-analysis. Ji et al. [14] further found visual-spatial skills useful in classifying performance among science and engineering students. Interests themselves have their own evidence base: the meta-analysis of Nye et al. [23], covering more than sixty years of research, found that vocational interests predict performance in work and educational settings, and that congruence between a person and their environment predicts more strongly than interest scores considered alone. That last result matters for the present design, because it implies that an interest score carries predictive weight partly through its fit with context rather than as a standalone trait. These constructs are related to interests but not interchangeable with them, which is why the present study measures interest intensity directly rather than treating it as a proxy for personality, self-regulation or cognitive skill. On the theoretical side, Gardner's theory of multiple intelligences [6]

and Deci and Ryan's self-determination theory [7] both suggest links between out-of-class activities and the development of specific competencies, while Zimmerman's model of self-regulated learning [8] provides a measurement framework for self-directed learning. No quantitative study, however, has yet modelled interests at a fine-grained level (type $\times$ intensity) as a primary predictive signal.

2.3. Multi-source learning and model explanation

In multi-view/multi-source learning, the two dominant architecture families are early fusion — concatenating the features of all sources into a single vector — and late fusion — training a separate model per source and combining the outputs with a meta-learner [5]. Baltrušaitis et al. [18] organise this field more finely around representation, translation, alignment, fusion and co-learning, so that the early-versus-late contrast is only one axis of a broader taxonomy. In education, Ochoa and Worsley [17] argue that video, audio and interaction traces can capture learning activity that conventional digital logs miss. A separate literature concerns the quality of the human ratings themselves: Uto and Ueno [19] show that assessment reliability depends on rater severity and consistency, which means disagreement between reference sources may encode rater behaviour as well as genuine differences between students. Disagreement between informants is in fact a long-standing object of study in its own right outside educational data mining: reviewing several decades of multi-informant assessment, De Los Reyes and Kazdin [24] report that correlations between informants are typically modest, and propose that discrepancies be modelled as meaningful data reflecting the setting in which each informant observes the child, rather than treated as error to be averaged away. The present study adopts that stance operationally while testing it predictively. For explanation, SHAP [9] and LIME [10] are the standard methods; TreeSHAP computes exact Shapley values on tree ensembles in polynomial time [9]. The limits of these methods are equally well documented. Ghassemi et al. [22] argue that current post-hoc explanations should support debugging and audit rather than substitute for rigorous internal and external validation. For Shapley-based methods specifically, Kumar et al. [29] show that the axioms which make Shapley values attractive in cooperative game theory do not by themselves license a feature-importance reading, and that correlated predictors make the resulting attributions sensitive to modelling choices. In education specifically, the XAI-ED framework of Khosravi et al. [20] ties explanation design to stakeholders, benefits, presentation approaches and pitfalls, while Conati et al. [21] find that the effect of explanations on trust and perceived usefulness varies with user characteristics. What remains open is the application of fusion architectures to human-assessment data from multiple reference sources in education, and in particular the question of whether disagreement between sources carries predictive information — the question this paper directly addresses.

3. METHODOLOGY

3.1. Overall framework

The research framework consists of five stages: (1) multi-source data collection; (2) feature construction, including inter-source disagreement features; (3) source combination under three fusion architectures; (4) prediction of the four target competencies; (5) SHAP-based model explanation. The three research questions are three views over the same pipeline: RQ1 uses stages 2–4 (feature-set ablation), RQ2 uses stages 1–3 (fusion comparison), and RQ3 uses stage 5.

3.2. Data and features

Consistent with the proof-of-concept aim, the framework is exercised on a synthetic cohort of 390 students rather than on an empirical one. No human participants were involved and no student records were accessed, so the study required no ethics approval; the cost of that choice is that nothing reported here is evidence about real students. The generator (NumPy PCG64, seed 20260911) draws correlated latent interests, converts them to integer Likert ratings with source-specific noise and severity, and produces bounded 0–100 competency scores from academic and selected interest signals plus Gaussian noise. Its parameters were calibrated so that the academic-only baseline falls in the range of explained variance typical of this literature. The dataset and the generating script are released with the paper so that every number below can be reproduced exactly. Three feature groups are constructed:

Group A: grade point averages, entrance scores, attendance rate, and number of late submissions.

Group B: Non-academic interest seven interest categories (sports, arts, games, reading, social activities, technology, music), each measured on a five-point Likert scale of participation and passion, replacing the binary encoding used in prior work.

Group C: Reference sources: the same interest questionnaire answered independently by four sources: student self-report, homeroom teacher, parent, and objective behavioural records (club participation hours, library borrowing history).

The target variables are four competencies: logical reasoning (a standardised reasoning assessment), creativity (a teacher-panel rubric on project work), communication (an observational rubric for presentations and teamwork), and self-directed learning (a self-report scale of self-regulated learning strategies, constructed on the cognitive, metacognitive, motivational and resource-management dimensions set out by Zimmerman [8] rather than adopted from a single published instrument). To control information leakage, the raters of the target variables are fully independent of the assessors providing Group C features; otherwise the model risks learning rater bias (halo effect) rather than genuine interest–competency relationships.

3.3. Inter-source disagreement features

For each interest i of student s , let $L(s, i, k)$ denote the Likert score provided by source k . The disagreement feature is defined as the standard deviation of $L(s, i, k)$ over the set of sources k , denoted $d(s, i)$. Research hypothesis H3 states: the degree of disagreement between reference sources — in particular between student self-reports and teacher assessments — itself carries predictive information about competency, with the intuition that self-assessment gaps reflect a student's metacognitive ability. That intuition should be held loosely: the metasynthesis of Zell and Krizan [25] finds self-evaluations of ability to be only moderately related to objective performance, with the strength of that relation itself varying by domain and by the quality of feedback available, so a self–other gap is not a clean index of metacognitive skill. This is not the only available interpretation: the same disagreement may reflect rater severity, unequal opportunity to observe, or inconsistent use of Likert categories [19], so the ablation tests only whether disagreement is predictive, not why. H3 is tested by comparing configurations with and without the $d(s, i)$ feature set, combined with SHAP analysis to confirm the position of these features in the importance ranking.

3.4. Ablation design (RQ1)

Table 1. The five ablation configurations and their dimensionality

Configuration	Feature set	Role	Features
M0	Group A only (academic)	Traditional baseline	4
M1	Group B only (interests, self-report)	Standalone predictive power of interests	7
M2	A + B (single source: self-report)	vs. M0: incremental value of interests	11
M3	A + B (four reference sources)	vs. M2: value of multi-reference data	32
M3+	A + B (four sources) + inter-source disagreement features	vs. M3: value of disagreement as a signal (H3)	39

The five configurations differ substantially in dimensionality, and this is central to interpreting the results that follow. With 390 students the ratio of observations to features falls from about 98 under M0 to about 35 under M2, about 12 under M3 and about 10 under M3+. Any incremental signal carried by the additional sources must therefore overcome a corresponding

rise in estimation variance before it can appear as lower held-out error.

Three hypotheses are tested on this design: H1, that adding self-reported interests improves on the academic-only baseline (M2 versus M0); H2, that adding the three further reference sources

improves on self-report alone (M3 versus M2); and H3, that adding inter-source disagreement features improves on the four-source model (M3+ versus M3). Each pair of configurations is compared by a paired comparison of per-fold errors across the five outer folds of the nested design described in Section 3.7, using the $t(4)$ quantile for interval estimation and reporting standardised effect sizes alongside every difference; conclusions are never drawn from single absolute differences. The 5×2cv procedure of Dietterich [11] was considered and rejected because it presumes five replications of two-fold resampling and is therefore incompatible with the nested five-fold outer loop adopted here. The entire procedure is repeated separately for each target competency.

3.5. Three source-fusion architectures (RQ2)

F1 — Early fusion: features from all sources are concatenated into a single vector and fed to one XGBoost/CatBoost model. Simple and effective when sources are consistent, but sensitive to noise when they conflict.

F2 — Late fusion (stacking): one base model per source; a meta-learner (Ridge regression) combines their outputs following the stacked-generalization principle [5]. The meta-learner captures the average reliability of each source.

F3 — Gated weighting: the meta-learner learns per-student reliability weights for each source through a simple gating mechanism, allowing source weights to vary by individual (e.g., for boarding students, teacher assessments may be more reliable than parent assessments).

All three architectures are trained on the same M3 data configuration under the same cross-validation protocol to guarantee a fair comparison; the best-performing architecture is selected as the main model for the explanation analysis.

3.6. Model explanation (RQ3)

TreeSHAP [9] is applied to the best model, run separately for each target competency to construct the interest–competency contribution matrix. Explanation stability is assessed with Spearman rank correlation between SHAP importance rankings

across folds; explanations are reported only when acceptable stability is reached. Explanations are treated throughout as evidence about model behaviour rather than about causal mechanism [22], and are tailored to their intended audience — research audit, counsellor review, or student conversation — following the stakeholder emphasis of XAI-ED [20]. Results are contrasted with the theory of multiple intelligences [6] and self-determination theory [7] in the discussion.

3.7. Evaluation protocol

The study uses nested cross-validation: a 5-fold outer loop provides unbiased performance estimation, while a 3-fold inner loop tunes gradient-boosted tree hyperparameters (maximum depth, number of trees, learning rate) by grid search. Folds are defined over individual students; because the synthetic cohort carries no school or class identifiers, grouped validation could not be applied, and this is stated as a limitation rather than presented as a design choice. Metrics are MAE, RMSE and held-out R^2 , reported as mean and standard deviation across the five outer folds. Paired differences between configurations use the $t(4)$ quantile of 2.776; the Holm procedure corrects within the twelve-test family formed by H1 to H3 across the four outcomes. Calibration of out-of-fold predictions and learning curves are reported in Section 4.4. This design follows the finding of Cawley and Talbot [26] that selecting hyperparameters on the same data used to report performance produces optimistically biased estimates, and the result of Vabalas et al. [27] that plain K-fold validation remains biased at small sample sizes whereas nested cross-validation does not. Reporting follows the principles of the TRIPOD statement [28]: the predictors, the outcome definition, the resampling scheme, the measures of dispersion around every performance estimate, and the handling of missing data are stated explicitly so that the model can be appraised and replicated.

4. EXPERIMENTAL RESULTS

4.1. Ablation results (RQ1)

Table 2. Mean absolute error (MAE) by configuration

Configuration	Logical reasoning	Creativity	Communication	Self-directed learning
M0	9.07 (0.94)	9.50 (0.53)	9.18 (0.70)	7.64 (0.55)
M1	10.04 (0.59)	9.64 (1.24)	9.76 (0.42)	9.24 (0.97)
M2	9.02 (0.84)	8.63 (0.69)	8.94 (0.62)	7.65 (0.76)
M3	9.19 (1.10)	8.87 (0.72)	8.91 (0.63)	7.56 (0.83)
M3+ (with disagreement)	9.05 (0.95)	9.00 (0.88)	8.97 (0.53)	7.63 (0.92)

Outcomes are scored on a 0 to 100 scale, with standard deviations of 12.0 for logical reasoning, 11.8 for creativity, 12.1 for communication and 11.0 for self-directed learning; these are reported so that every difference quoted below can be converted to a standardised effect size. Lower values indicate better

performance; the value in parentheses is the standard deviation across the five outer folds. On mean absolute error the best configuration is M2 for logical reasoning and creativity, and M3 for communication and self-directed learning. M1, which uses interests alone, is the worst configuration for every outcome.

Table 3. Root mean squared error (RMSE) by configuration

Configuration	Logical reasoning	Creativity	Communication	Self-directed learning
M0	11.17 (1.12)	11.69 (0.47)	11.29 (0.72)	9.62 (0.68)
M1	12.72 (0.87)	11.85 (1.28)	12.18 (0.56)	11.60 (1.24)
M2	11.03 (0.81)	10.82 (0.77)	10.99 (0.84)	9.67 (0.81)
M3	11.24 (1.11)	11.05 (0.81)	11.01 (0.72)	9.54 (0.95)
M3+ (with disagreement)	11.16 (0.95)	11.26 (1.00)	11.00 (0.67)	9.58 (1.05)

RMSE penalises large individual errors more heavily than MAE. The two metrics do not rank the configurations identically. For communication, M3 attains the lower mean absolute error (8.905 against 8.936) while M2 attains the lower root mean squared error (10.988 against 11.007) and the higher

held-out R^2 (0.178 against 0.175); M3 therefore achieves its slightly lower average error at the cost of a heavier tail. For the remaining three outcomes the ordering agrees. Because MAE alone conceals this pattern, both metrics are reported and no configuration is ranked on MAE in isolation.

Table 4. Held-out R^2 of the five ablation configurations per competency (mean over 5 outer folds)

Configuration	Logical reasoning (R^2)	Creativity (R^2)	Communication (R^2)	Self-directed learning (R^2)
M0	0.135 (0.137)	0.000 (0.220)	0.133 (0.073)	0.237 (0.065)
M1	-0.121 (0.113)	-0.005 (0.090)	-0.010 (0.055)	-0.100 (0.049)
M2	0.151 (0.140)	0.149 (0.153)	0.178 (0.080)	0.232 (0.047)
M3	0.122 (0.142)	0.120 (0.103)	0.175 (0.082)	0.253 (0.059)
M3+ (with disagreement)	0.133 (0.140)	0.080 (0.172)	0.177 (0.057)	0.249 (0.047)

The negative held-out R^2 values for M1 are reported as observed, in line with Section 3.7, and are not suppressed. A negative held-out R^2 means the model predicts worse than the training-set mean would. M1 is negative for logical reasoning, communication and self-directed learning, and approximately zero for creativity: interests measured by self-report alone carry almost no independent predictive signal once separated from academic context. The highest held-out R^2 observed anywhere in the study is 0.253 (M3, self-directed learning).

4.2. Early fusion versus late fusion

All three architectures were trained on the M3 data configuration under the same outer folds. Early fusion (F1) attains the lowest error for communication and self-directed learning, late fusion (F2) for logical reasoning, and the two are indistinguishable for creativity. Gated weighting (F3) is the weakest architecture for all four outcomes, and the difference against F2 excludes zero for creativity ($\Delta = +0.260$; 95% CI [+0.015, +0.506]) and communication ($\Delta = +0.195$; 95% CI [+0.006, +0.383]). The gate was not degenerate: fitted source weights varied substantially between students. Its additional parameters therefore cost more in estimation variance than they recovered in signal at this sample size.

Table 5. Fusion architectures compared on configuration M3 (mean MAE over five outer folds)

Outcome	MAE F1	MAE F2	MAE F3	Δ (F3 – F2) [95% CI]
logical reasoning	9.19	9.14	9.22	+0.076 [-0.136, +0.287]
creativity	8.87	8.90	9.16	+0.260 * [+0.015, +0.506]
communication	8.91	9.20	9.40	+0.195 * [+0.006, +0.383]
Self-directed learning	7.56	7.73	7.79	+0.055 [-0.321, +0.431]

4.3. Stability of SHAP rankings

Table 6. Mean Spearman rank correlation of feature importance across outer folds

Configuration	Logical reasoning	Creativity	Communication	Self-directed learning
M0	0.880	0.800	0.920	0.880
M1	0.696 *	0.693 *	0.561 *	0.496 *
M2	0.845	0.795	0.794	0.882
M3	0.554 *	0.549 *	0.678 *	0.576 *
M3 plus	0.490 *	0.530 *	0.599 *	0.641 *

Eight of the twenty configuration-by-outcome cells met the pre-specified stability threshold of 0.70. All four cells of M0 and all four of M2 passed; no cell of M1, M3 or M3+ did. Because the two richest configurations fail throughout, no SHAP ranking from M3 or M3+ is interpreted substantively anywhere in this paper. For the configurations that do pass, previous-term GPA is the largest single contributor to every outcome.

4.4. Source agreement, calibration and learning curves

Before modelling, the four reference sources were compared directly. Pairwise correlations between sources averaged 0.47 to 0.56 across the seven interest domains, and the mean absolute difference between any two sources was about 0.86 Likert points. Mean ratings differed systematically between informants (students 3.14, parents 3.08, behavioural records 3.01, teachers 2.96), which is the severity effect that Uto and Ueno [19] describe. The mean disagreement feature across domains was 0.64 scale points. These figures matter for interpreting the null result on H3: the sources do diverge, so a null is not explained by their being identical.

Calibration was assessed by regressing observed outcomes on out-of-fold predictions from M2. Slopes were below one for every outcome (logical reasoning 0.78, creativity 0.79, communication 0.81, self-directed learning 0.84), indicating that predictions are more dispersed than the outcomes they estimate the models overstate how far a student departs from the mean. Any student-facing use would need to shrink predictions accordingly.

marginally favours the baseline.

- H2 (M3 improves on M2): not supported for any outcome. All four intervals contain zero. The point

Learning curves for M2 were computed at 25, 50, 75 and 100 per cent of each training fold. Mean absolute error improves modestly and then flattens (logical reasoning 9.45 at 78 to 9.02 at 312; creativity 9.20 at 78 to 8.63 at 312; communication 9.30 at 78 to 8.94 at 312; self-directed learning 7.83 at 78 to 7.65 at 312). The curves give no indication that the null results for H2 and H3 would be overturned by a moderately larger sample; they do not exclude the possibility that a much larger one would.

5. DISCUSSION

5.1 Summary of principal findings

In this synthetic cohort the results are predominantly null. None of the three hypotheses is supported consistently across the four competency outcomes. All intervals below are paired differences over the five outer folds, computed with the $t(4)$ quantile of 2.776 rather than a normal quantile. As a proof of concept, what these results establish is that every decision rule in the protocol binds on data of this size and structure, and that the protocol returns a negative answer where the generator provides no signal rather than manufacturing one.

- H1 (M2 improves on M0): supported for one outcome only, creativity ($\Delta = -0.866$; 95% CI [-1.537, -0.194], $p = 0.023$). The remaining three intervals contain zero, and the creativity result does not survive Holm correction within the twelve-test family (adjusted $p = 0.278$). For self-directed learning the point estimate

estimates favour M3 for communication and self-directed learning and M2 for the other two, so the three

additional reference sources neither help nor harm reliably under early fusion.

- H3 (M3 plus improves on M3): no evidence for any outcome; all four intervals contain zero and no absolute point estimate exceeds 0.15 on the 0 to 100 outcome scale. Expressed in outcome standard deviations (Section 4.1), no standardised difference for H1, H2 or H3 exceeds 0.073 in absolute value, including the one interval that excludes zero; the maximum arises for H1 on creativity, where the difference of 0.866 points against a standard deviation of 11.8 gives 0.073.

In addition, configuration M1 (interests only) underperformed the academic baseline M0 on all four outcomes, with held-out R^2 at or below zero in every case. The gap reaches conventional significance for self-directed learning ($\Delta = +1.598$; 95% CI [+0.928, +2.269]). In this synthetic cohort, self-reported interests were therefore not a sufficiently strong independent predictive signal once isolated from academic context.

5.2 Interpreting null results

Null results are reported as observed and are not reinterpreted as supportive evidence. Specifically, a confidence interval containing zero does not establish that no effect exists; it establishes only that the available data cannot distinguish the effect from zero. Two substantively different explanations are equally consistent with these results:

of them. Three considerations temper these observations. First, the comparison is exploratory rather than pre-registered as primary. Second, the magnitudes are small: the largest architecture difference is 0.49 points on a 0 to 100 scale, against absolute errors of roughly 9 points. Third, no multiplicity correction has been applied within this exploratory family. A fourth qualification is more consequential: because M3 in Tables 2 to 4 is itself the early-fusion model, the ablation ladder is entirely an early-fusion sequence, and the conclusion recorded under H2 therefore holds only under early fusion. Substituting late fusion for early fusion did not change it F2 was indistinguishable from both M2 and M0 on every outcome.

5.4 Stability of model interpretation

This is the most methodologically concerning finding. Eight of twenty configuration-by-outcome cells met the pre-specified stability threshold of 0.70, and the pattern is not a simple function of model complexity. Both M0 (four features) and M2 (eleven features) passed on all four outcomes, whereas M1, M3 and M3+ failed on all four. Stability therefore tracks the ratio of signal to parameters rather than the raw number of features: M1 is small but carries almost no signal, while M3 and M3+ carry signal but spread it across thirty-two and thirty-nine correlated predictors. Because the four reference sources rate the same seven interest domains, the feature set of those configurations is correlated by construction, which is the behaviour Kumar et al. [29] anticipate.

The direct consequence is that for configurations falling below the threshold, SHAP importance rankings are NOT interpreted at all, cautiously or otherwise. A ranking that shifts between folds of the same dataset cannot support any claim about which

- Teacher, parent and behavioural ratings genuinely add nothing beyond what students already report about themselves.
- It does add information, but moving from M2 to M3 appends twenty-one further predictors seven interest domains for each of three additional sources to a sample of 390 students, and that inflates estimation variance more than the added signal reduces it. The number of features rises from 4 in M0 to 11 in M2, 32 in M3 and 39 in M3+, so the ratio of observations to features falls from about 98 to about 10. On this reading the finding concerns statistical power rather than the underlying phenomenon.

The present design offers no basis for distinguishing between these accounts. This is a limitation that should be stated plainly rather than passed over.

5.3 Comparing fusion strategies

No fusion architecture dominated. Early fusion (F1) produced the lowest error for communication and self-directed learning, late fusion (F2) for logical reasoning, and the two were indistinguishable for creativity; none of the four F2 against F1 intervals excludes zero. Gated weighting (F3) was the weakest for all four outcomes and significantly weaker than F2 for two

interest domain matters more than another. This is the behaviour Kumar et al. [29] anticipate for correlated predictors, and the four reference sources here rate the same seven interest domains, so the feature set is correlated by construction. Where rankings did pass the threshold, previous-term GPA was the largest single contributor to every outcome. That observation must be read correctly: it describes what the model relies upon when generating predictions, and says nothing about what produces competence in students still less in a cohort whose outcomes were generated by a script whose structure is known in advance.

5.5 Prediction versus causation

This is an observational study. Nothing reported above demonstrates that interests cause competencies, or the reverse. SHAP values characterise how the model uses its inputs, not the causal structure of the phenomenon [22]. At least three alternative explanations remain unexcluded: interests and competencies may share a common cause (family resources, school quality, access to extracurricular opportunity); the direction of association may be reversed, with more capable students finding it easier to access and sustain interests; or teacher ratings may reflect pre-existing expectations about particular students rather than independent observation of their interests.

5.6 Additional limitations arising from the results

- The synthetic cohort contains no missing values, so the analysis sample is the full 390 students and no complete-case exclusion was necessary. That is a property of the generator, not a strength of the study: the questions about non-random attrition

and about the observational limits of a single homeroom teacher that motivate multi-source designs cannot be examined here at all.

- Because no school or class identifiers were generated, cross-validation folds were formed over individual students. Any clustering of students within classes or schools is therefore

At the levels of accuracy and interpretive stability observed here, the models are not fit to support any decision about an individual student. The protocol already excludes use for admissions, grading, streaming, and discipline; these results reinforce that position. Even within a voluntary, human-reviewed counselling conversation, presenting model output to a student would require explicit communication of uncertainty and careful consideration of whether the benefit outweighs the risk of a labelling effect. The stakeholder-centred design advocated by XAI-ED [20], together with evidence that users respond differently to the same explanation [21], argues against presenting an identical technical SHAP plot to every student. The risk is not only that a prediction is wrong; Baker and Hawn [30] document how educational models can perform unevenly across student groups through the data and design choices made upstream of any explanation, which is a further reason to withhold this model from consequential decisions until subgroup performance has been examined directly.

5.8 Limitations

The discovered relationships are predictive, not causal; SHAP values reflect contributions to the model's predictions and do not establish that interests cause competencies. No reliability coefficient is reported for any of the four competency measures. Creativity is scored by a teacher panel and communication by an observational rubric, so inter-rater agreement is the relevant index in both cases, and it cannot be computed from the scores available here because only the aggregated value of each rubric was retained. Criterion validity of the targets is therefore assumed rather than demonstrated.

Non-academic interests risk acting as proxies for family socio-economic conditions, a mechanism through which educational models are known to reproduce existing inequities [30]; group-fairness results could not be computed because the dataset carries no demographic variables, so this limitation is stated rather than tested; the model should be used only to support counselling, never to classify or label students.

The sample of 390 is limited relative to the dimensionality of multi-source features: the ratio of observations to features falls

produces under a known generative model, not a recommended predictor. The protocol is now ready to be applied to an empirical cohort, where the same hypotheses can be tested against data that carry the school identifiers, rater-level rubric scores and demographic variables identified in Section 6.

REFERENCES

[1] W. Ahmed, M. A. Wani, P. Pławiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions,"

unmodelled, and the reported estimates should be read as an upper bound on what grouped validation would yield.

- The highest observed held-out R^2 was 0.253 (M3, self-directed learning), meaning that roughly three-quarters of the variance in competency outcomes remains unexplained. Any applied implication must begin from this figure.

5.7 Implications for practice

from about 98 under M0 to about 10 under M3+. The study applies regularisation and reports learning curves, but cannot separate a genuinely absent effect from one the design lacks power to detect.

The cohort is synthetic. It was generated to exercise the design end to end and to make every reported number reproducible from a published seed, not to describe any real population, which is both the purpose of a proof-of-concept study and the limit of what one can deliver. No claim in this paper is evidence about students, and the null results for H2 and H3 in particular are conditional on a generator in which the four informants are noisy readings of a common latent interest. A generator in which disagreement carried genuine signal would be required to show that the design can detect such an effect when it exists.

Three additions would strengthen any replication. Rubric scores should be retained at the level of the individual rater, so that inter-rater agreement can be reported for creativity and for communication. School and class identifiers should be recorded, so that validation folds can be formed over schools rather than over individual students and the leakage that clustering within classes would otherwise introduce can be ruled out. And demographic variables should be collected, so that the subgroup performance this study can only flag as a risk becomes something that can be measured.

6. CONCLUSION

This proof-of-concept study asked whether a multi-reference interest-to-competency pipeline can be specified, executed and interrogated end to end, and it can. On a synthetic cohort of 390 students the results are predominantly null: interest features improved prediction for one of four competencies and not after correction for multiplicity, additional reference sources and disagreement features improved nothing, per-student gated weighting was the weakest of the three fusion architectures, and SHAP rankings were stable enough to interpret in only eight of twenty cells and in none of the cells belonging to the two richest configurations. The contribution is therefore a validated and reproducible protocol, together with the negative evidence it

Scientific Reports, vol. 15, art. no. 26879, 2025, doi: 10.1038/s41598-025-12353-4.

[2] J. Lu et al., "Machine learning analysis of factors affecting college students' academic performance," *Frontiers in Psychology*, vol. 15, art. no. 1447825, 2024, doi: 10.3389/fpsyg.2024.1447825.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[4] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical

- features,” in *Advances in Neural Information Processing Systems* 31, 2018.
- [5] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
- [6] H. Gardner, *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books, 1983.
- [7] R. M. Ryan and E. L. Deci, “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being,” *American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000.
- [8] B. J. Zimmerman, “Becoming a self-regulated learner: An overview,” *Theory Into Practice*, vol. 41, no. 2, pp. 64–70, 2002.
- [9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems* 30, 2017, pp. 4765–4774.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [11] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [12] E. Alyahyan and D. Düşteğör, “Predicting academic success in higher education: Literature review and best practices,” *International Journal of Educational Technology in Higher Education*, vol. 17, art. no. 3, 2020, doi: 10.1186/s41239-020-0177-7.
- [13] W. Xing and D. Du, “Dropout prediction in MOOCs: Using deep learning for personalized intervention,” *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547–570, 2019, doi: 10.1177/0735633118757015.
- [14] M. Ji, J. Le, B. Chen, and Z. Li, “A predictive model for classifying college students’ academic performance based on visual-spatial skills,” *Frontiers in Psychology*, vol. 15, art. no. 1434015, 2024, doi: 10.3389/fpsyg.2024.1434015.
- [15] A. E. Poropat, “A meta-analysis of the five-factor model of personality and academic performance,” *Psychological Bulletin*, vol. 135, no. 2, pp. 322–338, 2009, doi: 10.1037/a0014996.
- [16] B. Ergen and S. Kanadlı, “The effect of self-regulated learning strategies on academic achievement: A meta-analysis study,” *Eurasian Journal of Educational Research*, no. 69, pp. 55–74, 2017, doi: 10.14689/ejer.2017.69.4.
- [17] X. Ochoa and M. Worsley, “Augmenting learning analytics with multimodal sensory data,” *Journal of Learning Analytics*, vol. 3, no. 2, pp. 213–219, 2016, doi: 10.18608/jla.2016.32.10.
- [18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.
- [19] M. Uto and M. Ueno, “Item response theory for peer assessment,” *IEEE Transactions on Learning Technologies*, vol. 9, no. 2, pp. 157–170, 2016, doi: 10.1109/TLT.2015.2476806.
- [20] H. Khosravi et al., “Explainable artificial intelligence in education,” *Computers and Education: Artificial Intelligence*, vol. 3, art. no. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.
- [21] C. Conati, O. Barral, V. Putnam, and L. Rieger, “Toward personalized XAI: A case study in intelligent tutoring systems,” *Artificial Intelligence*, vol. 298, art. no. 103503, 2021, doi: 10.1016/j.artint.2021.103503.
- [22] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.
- [23] C. D. Nye, R. Su, J. Rounds, and F. Drasgow, “Vocational interests and performance: A quantitative summary of over 60 years of research,” *Perspectives on Psychological Science*, vol. 7, no. 4, pp. 384–403, 2012, doi: 10.1177/1745691612449021.
- [24] A. De Los Reyes and A. E. Kazdin, “Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study,” *Psychological Bulletin*, vol. 131, no. 4, pp. 483–509, 2005.
- [25] E. Zell and Z. Krizan, “Do people have insight into their abilities? A metasynthesis,” *Perspectives on Psychological Science*, vol. 9, no. 2, pp. 111–125, 2014, doi: 10.1177/1745691613518075.
- [26] G. C. Cawley and N. L. C. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
- [27] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine learning algorithm validation with a limited sample size,” *PLOS ONE*, vol. 14, no. 11, art. no. e0224365, 2019, doi: 10.1371/journal.pone.0224365.
- [28] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement,” *BMC Medicine*, vol. 13, art. no. 1, 2015, doi: 10.1186/s12916-014-0241-z.
- [29] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, “Problems with Shapley-value-based explanations as feature importance measures,” in *Proc. 37th Int. Conf. Machine Learning (PMLR)*, vol. 119, 2020, pp. 5491–5500.
- [30] R. S. Baker and A. Hawn, “Algorithmic bias in education,” *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, 2022, doi: 10.1007/s40593-021-00285-9.