CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2765–2783
p-ISSN: 2086-0382; e-ISSN: 2477-3344

PCA-Based Labeling of Student Academic and Non-Academic Achievement: A Comparison of Data Transformation Methods

Salsabila Alya Putri Waluyo, Agung Teguh Wibowo Almais*, and Cahyo Crysdian

Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia

Abstract

Variance preservation does not necessarily establish meaningful achievement categories. This study compares six data transformations for principal component analysis (PCA)-based labeling using records from SMAN 1 Kertosono in 2024/2025. Academic data comprise 216 semester observations from 108 students in Grades X-XII; non-academic data comprise 144 observations from 72 students in Grades X-XI. Normalizer, StandardScaler, MinMaxScaler, MaxAbsScaler, RobustScaler, and QuantileTransformer were evaluated using student-grouped nested cross-validation. Transformations were selected by inner-validation macro F1 against teacher assessments. PCA fitting, sign orientation of the first principal component (PC1), and estimation of the one-third and two-thirds quantiles were confined to training data. These quantiles defined Low, Medium, and High categories for held-out students. MaxAbsScaler was selected in all five academic development sets, yielding 73.15% accuracy, macro F1 of 0.7233, and Cohen's kappa of 0.5982. Academic PC1 explained 81.07-82.16% of transformed training variance, but labeling did not outperform the raw-score tertile baseline. Non-academic selections varied across folds, yielding 27.78% accuracy, macro F1 of 0.2377, and kappa of -0.0904. Non-academic PC1 explained 17.48-27.86% of variance and primarily contrasted activity profiles, limiting its interpretation as an achievement scale. The contribution is an empirical evaluation separating variance retention, component interpretation, and teacher agreement. Results support a limited descriptive interpretation of academic PC1 but do not support the current non-academic categories as an overall achievement ranking. Findings are restricted to one school and require independent validation.

Keywords: Principal component analysis; data transformation; student achievement; PCA-based labeling; educational data analysis.

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

1. Introduction

Educational data mining and learning analytics provide methods for examining educational records and supporting the interpretation of student performance. Romero and Ventura describe a field encompassing different educational environments, data sources, and analytical objectives [1]. Within this field, a practical task is to summarize several recorded indicators into categories that teachers can inspect alongside the underlying records. Such categories are useful only to the extent that their meaning is clear: reducing a set of measurements to a short label does not establish that the label represents an educationally meaningful level of achievement.

*Corresponding author. E-mail: agung.twa@ti.uin-malang.ac.id

Received: August 21, 2026 Revised: October 14, 2026 Accepted: November 15, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.45741

Interpretability is therefore a substantive concern in the design of educational analyses. In their 2024 umbrella review, Gunasekara and Saarela examine how data, models, and evaluation metrics affect explainability in educational data mining and learning analytics [2]. They report that the reviewed literature discussed model performance and fairness metrics but did not evaluate explanation quality [2]. For achievement labeling, a related distinction is necessary between obtaining a compact numerical representation and justifying the educational interpretation assigned to it. A transparent calculation can still produce a category whose relation to the intended construct is weak. Examining this relation is especially relevant when academic grades and extracurricular records are processed using the same analytical procedure.

Principal component analysis (PCA) offers a linear representation of multivariate data through components that successively maximize variance under orthogonality constraints [3]. Its first principal component (PC1) can summarize a common direction when several indicators vary together. However, maximizing variance does not ensure that the resulting direction measures achievement. A component may instead express a contrast between indicators, and its numerical sign is arbitrary until an orientation convention is specified. In this study, PCA is consequently treated as dimensionality reduction. Assigning Low, Medium, and High categories using thresholds on PC1 is a separate labeling operation; the thresholds do not demonstrate that three naturally occurring clusters have been discovered.

A relevant methodological precedent is the study by Almais et al., which used PCA coordinates and value ranges to assign damage-level labels to post-disaster sector records [4]. This work motivates investigating component-based labeling in another application domain. Nevertheless, the meaning of a component and the suitability of its category boundaries depend on the measurements being analyzed. A rule developed for damage indicators cannot establish that school grades or activity codes form an equivalent ordered scale. The present study therefore examines the interpretation and validation of PCA-based labels in educational records, rather than transferring the earlier thresholds or assuming that variance preservation guarantees meaningful student categories.

Data transformation introduces an additional source of variation in this analysis. StandardScaler, MinMaxScaler, MaxAbsScaler, and RobustScaler modify feature scales using different reference quantities, whereas Normalizer rescales each observation and QuantileTransformer changes marginal distributions [5]. These operations can change the relative influence of indicators and the geometry presented to PCA. Consequently, explained variance describes the representation within a particular transformed space; a larger percentage across different spaces is not, by itself, a common-scale measure of better achievement labeling. A comparison of transformations must therefore examine category agreement and component interpretation as well as variance retention.

Evaluation design is equally important when several transformations are compared. Varma and Simon show that model selection and performance estimation within the same cross-validation process can yield optimistic error estimates, motivating a separate outer evaluation [6]. Kapoor and Narayanan further identify data leakage as a threat to reproducible claims in machine-learning-based science [7]. For school records containing repeated observations of the same student, this concern extends to separating students, rather than randomly separating semester records alone. Transformation fitting, component estimation, sign orientation, and threshold estimation must be confined to development data, with held-out students used to assess the resulting procedure.

Teacher assessments provide a practically relevant comparison, but agreement with teachers and objective measurement of achievement are different claims. Teachers may use observations that are not represented in the available numerical indicators, particularly for non-academic activities. Conversely, recorded activity codes may omit distinctions between unavailable information and an absence of documented participation. These issues motivate evaluating system-teacher agreement without assuming that either source is error-free. Accuracy is complemented here by

macro-averaged and per-class measures, while Cohen's kappa accounts for agreement expected from the marginal category frequencies [8]. This system-teacher comparison is also distinct from reliability among teachers, which requires overlapping ratings.

The empirical setting is SMAN 1 Kertosono during the 2024/2025 academic year. The academic analysis comprises 216 semester-level observations from 108 students in Grades X-XII, represented by seven subject aggregates. The primary non-academic analysis uses 144 observations from 72 students in Grades X-XI and 15 recorded activity indicators; Grade XII is excluded from this analysis because extracurricular activities were unavailable for that grade. Six transformations are compared using student-grouped nested validation, with transformation selection based on inner-validation macro F1. PC1 orientation and quantile thresholds are estimated from training records. Raw-score and simple category baselines, component diagnostics, and sensitivity analyses provide complementary checks on the interpretation of the resulting labels. Teacher labels inform transformation selection, so the complete evaluation pipeline is not described as wholly unsupervised.

The research gap addressed here is the need to establish, for these school records, whether variance-based compression supports defensible ordinal labels after accounting for transformation choice, repeated student observations, and disagreement between recorded indicators and teacher assessments. The objective is to compare six preprocessing methods for PCA-based academic and non-academic labeling and assess their agreement with teacher categories on held-out students. The novelty is empirical and analytical: a joint examination of variance retention, component direction, and category agreement within a reproducible validation design, rather than a new PCA algorithm. The contributions are an explicit labeling and evaluation procedure; evidence that the academic PC1 largely follows a common grade direction without outperforming the raw-score baseline; and a diagnosis of why non-academic PC1 categories fail to support an overall achievement ranking in this dataset. These findings identify conditions and limitations of component-based labeling within one school, rather than a universally superior transformation.

2. Methods

This section describes the school records and achievement indicators, then explains how the PCA labels were constructed and evaluated against teacher assessments.

2.1. Study design and data sources

This study examined whether principal component analysis (PCA) scores obtained after different data transformations could be used to assign relative Low, Medium, and High categories to student records. PCA provided a representation through linear combinations of the input features [3]. Category assignment was then performed using an explicit threshold rule. No clustering algorithm was applied. Academic and non-academic records were analyzed separately. The unit of analysis was a student-semester record, and the unit of separation between development and testing was the student.

The data originated from SMAN 1 Kertosono in the 2024/2025 academic year. Six school files represented Grades X, XI, and XII in semesters 1 and 2, with 36 records for each grade-semester combination. Records were joined across semesters using the school student identifier, and identifiers were replaced with study codes for analysis. All 216 supplied records from 108 distinct students were included in the academic analysis. The primary non-academic analysis used the 144 records from 72 students in Grades X and XI. Grade XII records were retained for the source audit and description, but were excluded from the activity-based PCA analysis: extracurricular activity fields were unavailable for those 72 records because Grade XII students did not participate in the recorded extracurricular program. Treating those structurally unavailable fields as zero activity would have introduced a different meaning of zero. The available school records constituted the study dataset; the files do not establish a random sample of a broader student population.

2.2. Academic indicators and activity coding

Seven academic indicators were formed from available numerical subject grades. For record i and academic indicator g, the aggregation rule was

$$x_{ig} = \text{round}_2\left(\frac{1}{|S_{ig}|}\sum_{j\in S_{ig}} v_{ij}\right), \tag{1}$$

where S_{ig} is the set of subjects with a numerical grade available for that record and indicator, and v_{ij} is the corresponding subject grade. Equation (1) gives equal weight to available subjects and rounds the mean to two decimal places. No subject weights were applied. An absent subject was omitted from the denominator rather than assigned a grade of zero. Table 1 lists the source subjects; their availability varied by grade and semester. In particular, advanced mathematics (MAT LJT) was recorded only in Grade XII semester 1. Every record had a numerical value for each of the seven resulting indicators, so no imputation of these indicators was needed.

Table 1: Academic indicators and constituent school subjects

Indicator	Source subjects	Availability	Observed range
Religion (AGAMA)	PAI, KRISTEN, KATOLIK	Student's recorded religion subject	86.00-97.00
Language (BAHASA)	BIN, BIG, JAWA	All grades and semesters	86.00-97.67
Mathematics (MATEMATIKA)	MAT, MAT LJT	MAT throughout; MAT LJT only XII semester 1	86.00-95.00
Science (SAINS)	IPA; BIO, KIM, FIS	IPA in X; BIO, KIM, FIS in XI-XII	83.00-95.33
Social studies (SOSIAL)	IPS, SEJARAH, PKn	IPS and PKn in X; SEJARAH and PKn in XI-XII	86.50-94.00
Skills (KETERAMPILAN)	PJOK, TIK, PKWU	PJOK/TIK in X; all three in XI; PJOK/PKWU in XII	85.00-97.00
Arts (SENI)	SENI	All grades and semesters	89.00-96.00

Source abbreviations denote Islamic, Christian, and Catholic religious education (PAI, KRISTEN, KATOLIK); Indonesian, English, and Javanese (BIN, BIG, JAWA); Mathematics and advanced Mathematics (MAT, MAT LJT); general Science, Biology, Chemistry, and Physics (IPA, BIO, KIM, FIS); Physical Education, Sports, and Health (PJOK); Information and Communication Technology (TIK); Crafts and Entrepreneurship (PKWU); and Arts (SENI). Ranges in Table 1 describe the supplied numerical grades after aggregation, not verified institutional grading limits. Different subject compositions limit direct comparisons across grades.

The non-academic dataset contained 15 recorded activity indicators, listed in Table 2. The source codes B and SB were represented as 1 and 2, respectively. In the analysis matrix for Grades X-XI, an activity without a recorded B or SB code was represented as 0. This zero means no such code was recorded in the available file. It does not establish non-participation, absence of achievement, or equal opportunity to participate. Source audits distinguished blank cells in existing columns from columns absent from a given file; both lead to a zero in the operational matrix when no code was recorded. Thus, this representation may conflate unrecorded participation with unavailable information. The spacing between 0, 1, and 2 was a modeling convention for the PCA input, not evidence of equal differences in underlying achievement. The four activities appearing as columns in every X-XI file were English Club, Journalism, PMR, and PPST PUSPENI. Twenty-six of the 144 X-XI observations had no recorded B or SB code in any activity. These observations were retained in the main analysis and separately examined in diagnostic summaries.

For every activity in Table 2, 1 denotes a recorded B, 2 a recorded SB, and 0 the absence of either code in the available file. English activity labels are used where appropriate, with the

Table 2: Non-academic input indicators and numerical representation

No.	Activity label (original dataset name)	Codes in analysis matrix
1	English Club (ENGLISH CLUB)	0, 1, 2
2	Music (MUSIK)	0, 1, 2
3	KIR	0, 1, 2
4	ICT	0, 1, 2
5	Journalism (JURNALISTIK)	0, 1, 2
6	CAGASKARA	0, 1, 2
7	ROHIS	0, 1, 2
8	PMR	0, 1, 2
9	PPST PUSPENI	0, 1, 2
10	SKK	0, 1, 2
11	TOD	0, 1, 2
12	BIVO	0, 1, 2
13	BRIGIP MASDA	0, 1, 2
14	Basketball (BASKET)	0, 1, 2
15	PPS	0, 1, 2

original dataset names retained in parentheses in Table 2. School-specific names and abbreviations are retained in their recorded form when a verified expansion or English equivalent is unavailable. PPS denotes *Putra Putri SMAKER*. Grade, semester, student, and record identifiers, teacher labels, and categories produced in the earlier version of the study were excluded from the PCA input matrix. Grade was used only when defining the validation splits and the grade-specific diagnostic baseline.

2.3. Teacher assessments and evaluation reference

Separate teacher-assigned academic and non-academic categories were available for the 216 student-semester records. Three homeroom teachers, each responsible for the students in one grade, provided the assessment reference. According to the reported assessment procedure, teachers assigned Low, Medium, or High categories without viewing the PCA-generated categories. Their academic assessments considered subject performance and consistency. Their non-academic assessments could also draw on participation in classroom, school, organizational, and other activities observed over an approximately one-year period, including information beyond the extracurricular fields available to the PCA model. Teacher labels were joined to the school records after their assignment; original assessments were retained and only category spelling and capitalization were normalized for analysis.

The reported procedure describes one annual assessment applied to two semester records. However, the supplied files contain different teacher categories between semesters for 36 of 108 academic student pairs and 27 of 108 pairs in the complete non-academic teacher file; 22 of these non-academic discrepancies occur among the 72 Grade X-XI students included in the primary activity analysis. The analysis therefore retained the actual label attached to each record rather than constructing a replacement annual label. A separate sensitivity analysis used only student pairs whose recorded categories matched across semesters. The files do not independently establish assessment dates, a historically documented three-level rubric, or ratings of the same student by multiple teachers. Consequently, inter-rater reliability cannot be estimated from these data. Agreement between the system and these teacher assessments was treated as agreement with an assessment reference, not proof that either set of categories was objectively correct. Although teachers did not see the PCA categories according to the reported procedure, some subject grades and activity records informing their judgments overlapped the model inputs; their assessments were therefore not informationally independent of those inputs.

2.4. Data transformations and PCA

The six transformations and their configurations are given in Table 3. L2 normalization acts on a record vector, whereas the other transformations act by feature; the collective term *data transformations* therefore distinguishes normalization, linear scaling, and distribution mapping [5]. The estimators were implemented in scikit-learn [9]. Parameters were estimated from the relevant training subset and applied unchanged to held-out records. No additional StandardScaler was applied after any candidate transformation. Constant features and zero interquartile ranges were recorded and handled by the configured implementation; they were not removed using statistics from the full dataset.

Table 3: Candidate transformations and settings used in each training fit

Code	Transformation	Configuration
NR	Normalizer	L2 norm per record; a zero vector remains zero
SS	StandardScaler	Training mean removal and standard deviation scaling; with_mean=True, with_std=True
MM	MinMaxScaler	Training range [0,1]; clip=False
MA	MaxAbsScaler	Division by the training feature maximum absolute value
RS	RobustScaler	Training median and interquartile range; quantile_range=(25, 75)
QT	QuantileTransformer	Normal target; n_quantiles=min(1000, n_train); subsample=None; random_state=42

PCA was fitted separately to each transformed academic or non-academic training matrix. The implementation used full singular value decomposition with two principal components (PC1 and PC2) and no whitening [10]. For a transformed training matrix $\mathbf{T}$ with n rows, centering and projection are expressed by Eq. (2):

$$\mathbf{Z} = \left(\mathbf{T} - \mathbf{1}_n \mu_T^{\mathrm{T}}\right) \mathbf{V}_2. \tag{2}$$

Here, μ_T is the column vector of training feature means, $\mathbf{1}_n$ is a vector of ones, and the two columns of $\mathbf{V_2}$ are the leading principal directions. The component eigenvalues and directions summarize variance in the transformed space. The same fitted transformation, training mean, and directions were used to project held-out records. Explained-variance ratios were recorded for both components to describe the transformed representation. Only the first component was used in the primary category rule; high explained variance was neither the criterion used to select a transformation in the primary analysis nor a substitute for agreement with teacher assessments. Values of explained variance obtained from different transformed feature spaces were interpreted within their respective spaces.

2.5. PC1 orientation and category assignment

The sign of a PCA eigenvector and its scores can be reversed without changing the PCA solution [3]. To orient the academic PC1, the Spearman correlation between training PC1 scores and the record-wise mean of the seven *untransformed* academic indicators was computed on the training subset. For non-academic PC1, the training reference was the record-wise sum of the 15 coded activities. If this correlation was negative, both PC1 scores and the corresponding eigenvector coefficients were multiplied by -1. The fitted sign was also applied to held-out records. No teacher category or held-out record was used to set the sign. A non-finite correlation or absolute correlation at most 10^{-12} made that fit unusable. An absolute correlation below 0.2 was flagged as indicating a weak descriptive association; it did not automatically invalidate the fit or establish a statistical significance threshold. A positive oriented association alone does not prove that PC1 measures overall achievement.

For an oriented training PC1 score vector z_{train}, the lower and upper boundaries were its empirical one-third and two-thirds quantiles, respectively, using NumPy's linear interpolation

convention (Hyndman-Fan type 7) [11, 12]. For sorted training scores $z_{(1)}, \ldots, z_{(n)}$ and $0 < q < 1$, this convention uses $h = 1 + (n-1)q$, $j = \lfloor h \rfloor$, and $\gamma = h - j$, giving $Q_q = (1-\gamma)z_{(j)} + \gamma z_{(j+1)}$. The boundaries in Eq. (3) were

$$a = Q_{1/3}\left(z_{\text{train}}\right), \qquad b = Q_{2/3}\left(z_{\text{train}}\right). \tag{3}$$

The classification rule for a training or held-out oriented score z was given by Eq. (4):

$$\hat{y}(z) = \begin{cases} \text{Low}, & z < a, \\ \text{Medium}, & a \le z < b, \\ \text{High}, & z \ge b. \end{cases} \tag{4}$$

Teacher labels were not used to calculate a or b. Tied PC1 scores were assigned by these inequalities without random jitter or forced equal-size groups. Boundaries were considered coincident when $b - a \le 10^{-12} \max(1, |a|, |b|)$; a fit was also ineligible if its training predictions did not contain all three categories. A valid held-out set was not required to contain all three predicted categories. These boundaries describe relative positions in a training-score distribution and are not fixed school achievement standards.

2.6. Student-grouped nested validation and transformation selection

Nested validation separates model selection from evaluation [6]. We used five outer folds for evaluation and three inner folds for transformation selection. Grouping records by student addressed the risk of leakage between dependent semester records [7]. Unique student identifiers were sorted, shuffled separately within each grade using NumPy's default_rng with seed 42, and allocated to outer folds using rotating round-robin assignment. Inner-fold assignments were generated within each outer development set with seed $42 + f$, where f was the outer-fold number. The non-academic analysis inherited the same student assignments after restricting the population to Grades X-XI. Teacher categories were not used to make these splits. Both semesters of a student always remained in the same development or held-out group, and each record was evaluated once in an outer test fold.

Within every outer development set, each of the six transformations was independently fitted in each inner training subset. The transformation, PCA, PC1 sign, and two category boundaries were re-estimated from that inner training subset before predictions were made for its inner validation subset. The primary selection criterion was the mean of the three inner-validation macro F1-scores computed against the teacher categories with a fixed class list comprising Low, Medium, and High. A candidate failing in any inner fit was ineligible. Ties within 10^{-12} were resolved using the mean unweighted Cohen's kappa, prioritizing candidates whose inner-fold kappa values were all defined, followed by the fixed order NR, SS, MM, MA, RS, QT. Among candidates with incomplete kappa values, the mean of available finite values was used, with candidates having no finite values ranked last.

The selected transformation was then refitted, together with PCA, orientation, and boundaries, on the full outer development set and applied once to the outer test students. If all candidates failed, or the selected candidate failed on outer refitting, the affected observations were retained with a missing-prediction status and coverage was reported. This protocol was recorded after the data audit and before the revised model fitting; the historical dataset had previously been explored, so the protocol is not a preregistration.

Teacher categories were used in inner validation to select the transformation and in outer testing to evaluate agreement. Thus, although PCA, orientation, and quantile boundaries do not directly use teacher categories, the complete selection-and-labeling workflow is not wholly unsupervised. The outer test categories played no role in choosing the transformation or its boundaries. This separation addresses the optimism caused by using evaluation data for model selection [6, 7].

2.7. Agreement measures and reporting

Predictions for held-out students were pooled once per observation. We reported accuracy, macro precision, macro recall, macro F1, unweighted Cohen's kappa, category-specific metrics and support, and a confusion matrix with teacher categories as rows and system categories as columns. Category-specific precision, recall, and F1 were defined as in Eq. (5) [13]:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F_{1,c} = \frac{2TP_c}{2TP_c + FP_c + FN_c}. \tag{5}$$

Here, TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives relative to teacher category c. With three fixed categories and N valid predictions, Equation (6) defines the macro average for $M \in \{P, R, F_1\}$ and overall accuracy:

$$M_{\text{macro}} = \frac{1}{3}\sum_{c=1}^{3} M_c, \qquad \text{Accuracy} = \frac{\sum_{c=1}^{3} TP_c}{N}. \tag{6}$$

The unweighted kappa in Eq. (7) adjusts observed agreement for the agreement expected from the marginal label distributions [8]:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_e = \frac{1}{N^2}\sum_{c=1}^{3} n_{c\cdot} n_{\cdot c}, \qquad p_o = \text{Accuracy}. \tag{7}$$

In Eq. (7), $n_{c\cdot}$ and $n_{\cdot c}$ are the teacher and prediction totals for category c. Undefined kappa was recorded as missing. For reproducible macro scoring, an undefined category metric contributed zero; undefined values in the category-specific report were separately marked, consistent with the implementation conventions [14]. Support-weighted recall equals accuracy in this single-label multiclass setting and was not treated as additional independent evidence. These metrics quantify system-teacher agreement; they do not estimate reliability between teachers who did not independently rate the same records.

Coverage was the number of valid predictions divided by the number of eligible observations. When predictions were unavailable, metrics were explicitly conditional on the successfully predicted observations. Pooled scores were distinguished from the unweighted outer-fold mean and sample standard deviation (SD; ddof=1); this standard deviation was not interpreted as a confidence interval. Paired comparisons used observations with valid predictions from both procedures and retained their coverage denominators. Differences in point estimates were not treated as evidence of statistical significance.

2.8. Baseline and sensitivity analyses

Three baselines fitted on training data supplied context: the globally most frequent teacher category, the most frequent teacher category within each grade (a diagnostic baseline, not a PCA feature), and the same training-quantile rule applied directly to the mean of seven academic indicators or the sum of activity codes.

We examined the sensitivity of the main labeling procedure to training quantiles (0.30, 0.70), (0.35, 0.65), and (0.25, 0.75) while holding the primary selected transformation fixed within each outer fold. Alternative transformation criteria were inner-validation kappa, mean training explained-variance ratio of PC1, and mean training explained-variance ratio of PC1 plus PC2. Additional analyses refitted the procedure with 14 activities excluding PPS or only the four activities present in all X-XI files; restricted the population to one semester at a time or to students whose teacher labels matched across semesters; and summarized predictions by grade and absence of recorded activity. All restricted analyses retained the inherited student folds and reported both eligible and successfully predicted counts. Alternatives were evaluated as sensitivity analyses rather than selected retrospectively as a new primary rule.

To examine whether additional component information could support agreement with teacher categories, an L2-regularized multinomial logistic regression was included as a supervised comparator [15]. It used either one PC, two PCs, or the smallest number of training PCs required to reach at least 80% cumulative explained variance. Within each outer fold it retained the transformation selected for the primary procedure; all transformations and PCA models were refitted within inner training subsets. Its regularization parameter $C \in \{0.01, 0.1, 1, 10, 100\}$ was selected by inner macro F1, then kappa, then the smaller C in a tie. The implementation used the lbfgs solver, a maximum of 5,000 iterations, and no class weighting or added scaling of PCA scores. Three teacher categories were required in a comparator training subset; failure to converge or failure of all candidate C values was reported as an unavailable model. Because this model learns decision boundaries using teacher categories, its performance is not the performance of quantile-based PCA labeling, and differences cannot be attributed to the number of PCs alone.

2.9. Computational reproducibility

All analyses were performed in Python 3.12.14 with NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, and scikit-learn 1.8.0. Source-file checksums, student-fold assignments, per-fit transformation and PCA parameters, selected candidates, individual held-out predictions, and calculation outputs were retained to permit audit of the revised experiments.

3. Results and discussion

The student-grouped nested evaluation showed higher agreement with teacher assessments for academic labeling than for non-academic labeling. All eligible records received a prediction in the primary analysis. The results below separate held-out agreement from training explained variance, report category-level errors, and examine the sensitivity of these findings to transformation selection, category boundaries, and the available inputs.

3.1. Results

3.1.1. Evaluation coverage and overall agreement

The academic evaluation comprised 216 student-semester observations from 108 students in Grades X-XII. The primary non-academic evaluation comprised 144 observations from 72 students in Grades X-XI; the 72 Grade XII observations were outside the activity-based evaluation population. Each student contributed two semester records that remained in the same outer fold. Academic outer test sets contained 44, 44, 44, 42, and 42 observations while non-academic test sets contained 30, 30, 28, 28, and 28 observations. Prediction coverage was 100% in all five outer folds for both domains.

Table 4 reports pooled predictions for held-out students. Academic labels agreed with teachers for 158 of 216 observations, yielding a macro F1-score of 0.7233 and Cohen's kappa of 0.5982. Non-academic labels agreed for 40 of 144 observations, with macro F1 of 0.2377 and kappa of -0.0904. The negative non-academic kappa indicates less agreement than expected from the observed marginal category distributions. These are estimates of internal agreement within the study school, using teacher assessments as the evaluation reference.

Across the five outer folds, the unweighted mean accuracy was 73.23% (SD 6.98 percentage points) for academic labels and 27.62% (SD 10.55 percentage points) for non-academic labels. Mean macro F1 was 0.7212 (SD 0.0719) and 0.2369 (SD 0.0685), respectively. These fold summaries describe variation across partitions; their standard deviations are not confidence intervals and are distinct from the pooled metrics in Table 4.

3.1.2. Category distributions and error patterns

Table 5 and Fig. 1 derive from the same pooled predictions as Table 4. Academic teacher counts were 109 Low, 47 Medium, and 60 High, whereas predicted counts were 71, 74, and 71.

Table 4: Pooled agreement with teacher categories in the primary evaluation

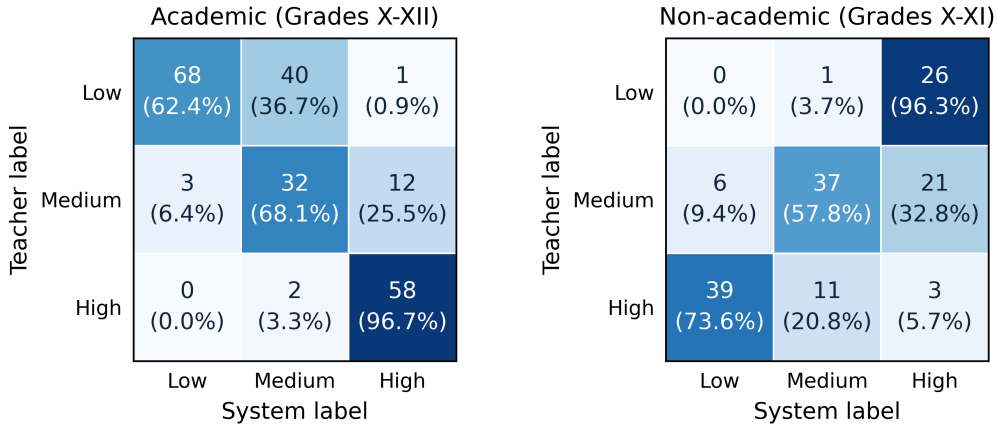
Measure	Academic	Non-academic
Students	108	72
Observations/Valid Predictions	216/216	144/144
Prediction Coverage (%)	100.00	100.00
Accuracy (%)	73.15	27.78
Macro Precision	0.7357	0.2717
Macro Recall	0.7571	0.2116
Macro F1	0.7233	0.2377
Cohen's Kappa	0.5982	-0.0904

Non-academic teacher counts were 27 Low, 64 Medium, and 53 High, compared with predicted counts of 45, 49, and 50. Training tertiles therefore did not impose equal category sizes on the held-out records.

Table 5: Category counts and category-specific agreement

Domain	Category	Teacher n	Predicted n	Precision	Recall	F1
Academic	Low	109	71	0.9577	0.6239	0.7556
Academic	Medium	47	74	0.4324	0.6809	0.5289
Academic	High	60	71	0.8169	0.9667	0.8855
Non-academic	Low	27	45	0.0000	0.0000	0.0000
Non-academic	Medium	64	49	0.7551	0.5781	0.6549
Non-academic	High	53	50	0.0600	0.0566	0.0583

Academic disagreement was concentrated near the boundary between Low and Medium: 40 teacher-labeled Low observations were assigned Medium. The Medium category had precision of 0.4324 because its 74 predictions included these 40 Low observations and two High observations. In contrast, 58 of 60 teacher-labeled High observations were recovered as High, giving recall of 0.9667. Only one academic observation crossed directly between Low and High.



Cells show counts and percentages within each teacher category.

Fig. 1: Pooled confusion matrices for academic and non-academic analyses

Non-academic errors were more severe. None of the 27 teacher-labeled Low observations was predicted as Low; 26 were assigned High. Of 53 teacher-labeled High observations, 39 were assigned Low and only three were recovered as High. Thus, 65 of 144 observations crossed between the two extreme categories. The Medium category accounted for 37 of the 40 agreements. The weak overall result cannot be explained only by an imbalance in category frequencies, because

the extreme category reversal is visible in the individual confusion-matrix cells.

3.1.3. Transformation selection and fitted category boundaries

Table 6 summarizes the six transformation candidates during inner validation. Every candidate produced valid predictions in all 15 primary inner fits per domain. MaxAbsScaler was selected in each of the five academic outer development sets. Non-academic selections varied across folds: MinMaxScaler in fold 1, QuantileTransformer in folds 2, 3, and 5, and Normalizer in fold 4. The reported inner means describe the selection stage; they are not additional independent test estimates or a post hoc criterion for choosing a single final winner.

Table 6: Inner-validation macro F1 and outer-fold selection frequency

Transformation	Academic mean macro F1	Academic selected	Non-academic mean macro F1	Non-academic selected
Normalizer	0.5270	0/5	0.3026	1/5
StandardScaler	0.6836	0/5	0.2873	0/5
MinMaxScaler	0.7039	0/5	0.3899	1/5
MaxAbsScaler	0.7269	5/5	0.3899	0/5
RobustScaler	0.6085	0/5	0.3069	0/5
QuantileTransformer	0.6100	0/5	0.4373	3/5

Table 7 gives the fitted PC1 boundaries and representation diagnostics for each selected outer model. Academic lower boundaries ranged from -0.040745 to -0.038361 and upper boundaries from 0.035968 to 0.039766. Non-academic boundaries varied more widely across transformations; in fold 4, both boundaries were positive. Categories therefore followed the training quantiles specified in Subsection 2.5, rather than a rule equating all negative PC1 scores with Low and all positive scores with High.

Table 7: Training parameters of the selected outer-fold models

Domain	Fold	Transformation	Lower	Upper	PC1 EVR (%)	PC1 + PC2 EVR (%)	Anchor rho
Academic	1	MA	-0.038361	0.039695	81.71	91.67	0.9958
Academic	2	MA	-0.040745	0.035968	82.16	91.37	0.9955
Academic	3	MA	-0.040459	0.036108	81.07	91.17	0.9954
Academic	4	MA	-0.040001	0.037018	81.89	91.68	0.9950
Academic	5	MA	-0.040270	0.039766	82.09	91.60	0.9959
Non-academic	1	MM	-0.117275	0.085273	19.77	33.80	0.1163
Non-academic	2	QT	-0.627321	0.191594	17.48	33.99	0.0929
Non-academic	3	QT	-1.330992	0.937031	21.57	36.14	0.0933
Non-academic	4	NR	0.078179	0.136806	27.86	44.80	0.1915
Non-academic	5	QT	-1.146322	0.649282	18.98	32.82	0.0243

MA = MaxAbsScaler; MM = MinMaxScaler; QT = QuantileTransformer; NR = Normalizer. EVR = explained-variance ratio. Boundaries are rounded for display; calculations used full precision. Anchor rho is the training Spearman correlation after sign orientation. PC1 scores and boundaries from different fitted spaces are not interchangeable.

3.1.4. Variance representation and component interpretation

Academic PC1 explained 81.07-82.16% of the transformed training variance, and PC1 plus PC2 explained 91.17-91.68%. Its correlation with the raw mean academic anchor was 0.9950-0.9959 after orientation (Fig. 2). All seven academic PC1 coefficients were positive in all five selected models. The skills aggregate had the largest coefficient in each fold (0.5820-0.5969), whereas

social studies had the smallest (0.1694-0.1806). The academic component therefore followed a common direction across the recorded aggregates, although their contributions were unequal.

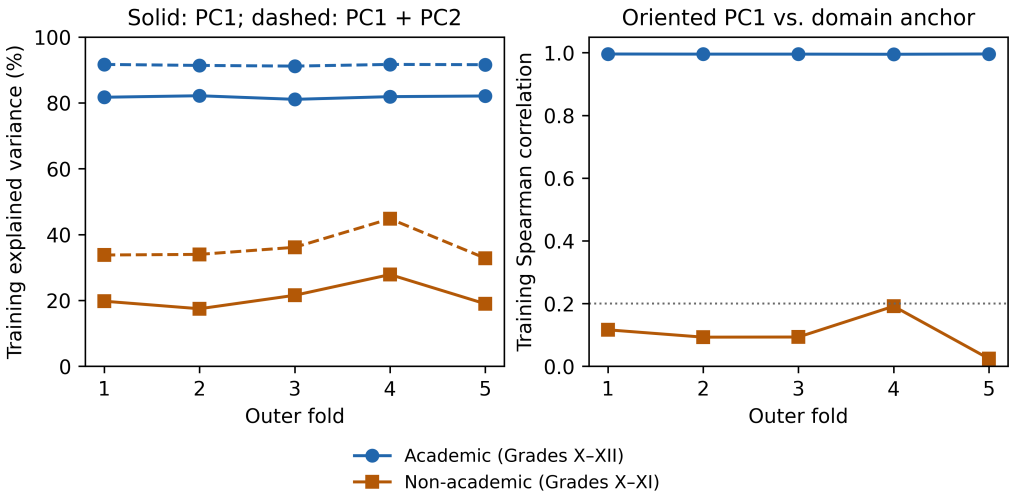


Fig. 2: Training explained variance and oriented PC1 correlation with the domain anchor

Non-academic PC1 explained only 17.48-27.86% of transformed training variance, leaving 72.14-82.52% outside the component used for primary labeling. The first two components together explained 32.82-44.80%, so even this two-dimensional representation left more than half of the variance outside the displayed components. Oriented correlations with the raw activity-code sum ranged from 0.0243 to 0.1915; all five models met the protocol's weak-association flag. Sign reversal was required in four non-academic outer models, whereas the academic signs were already positively aligned.

Fig. 3 shows the non-academic PC1 coefficients. Journalism and English Club had the two largest absolute coefficients in every selected model, but with opposite signs: Journalism was negative and English Club positive after orientation. The leading direction therefore primarily contrasted these recorded activity profiles. It did not show a uniformly positive contribution from all activities. Table 8 also shows that PC2 did not have the same leading pair in every fold: Music and English Club ranked first and second in fold 5. These findings qualify the interpretation of PC1 as an overall non-academic achievement scale.

Table 8: Non-academic features ranked first and second by absolute component coefficient

Fold	PC1 leading features	PC2 leading features
1	Journalism, English Club	English Club, Journalism
2	English Club, Journalism	Journalism, English Club
3	Journalism, English Club	Journalism, English Club
4	Journalism, English Club	English Club, Journalism
5	Journalism, English Club	Music, English Club

3.1.5. Baselines and supervised comparisons

Table 9 compares the primary procedure with the prespecified baselines and supervised logistic models on the same held-out observations. All procedures in this table achieved full coverage. For academic labels, applying training tertiles directly to the raw mean score gave macro F1 of 0.7288, compared with 0.7233 for PCA-based labeling. Only 10 of 216 predictions differed between these two procedures. Thus, the present point estimates did not show an incremental agreement benefit from PCA over the raw-score baseline.

The academic grade-specific majority baseline had higher accuracy (76.39%) than the primary procedure but lower macro F1 (0.5725). The global majority baseline reached 50.46% accuracy

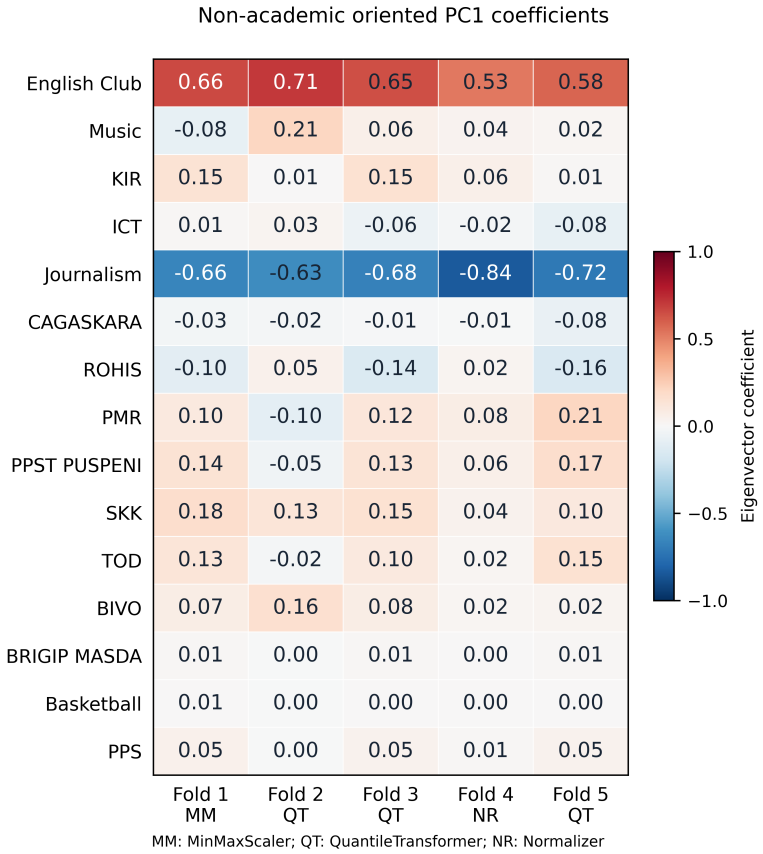


Fig. 3: Non-academic PC1 eigenvector coefficients across selected outer models

and macro F1 of 0.2236. In the non-academic domain, the primary procedure had lower accuracy than both majority baselines and the raw-score tertile baseline. Its macro F1 exceeded the global majority value of 0.2051, but remained below the grade-specific majority value of 0.3680 and the raw-score value of 0.2740. These comparisons show why accuracy alone is insufficient to characterize the observed category agreement.

Academic logistic regression using PC1 achieved 87.96% accuracy and macro F1 of 0.8432; using the first two PCs gave 85.19% and 0.7958. Retaining at least 80% of training variance required only one academic PC in every outer fold and reproduced the PC1 logistic results. Non-academic logistic regression achieved macro F1 of 0.8182 with PC1, 0.8101 with two PCs, and 0.8279 with the 80%-variance representation, which required seven PCs in every outer fold. Adding PC2 alone did not increase agreement in either domain. The larger differences from the primary method also involve learning category boundaries from teachers, so they cannot be attributed to additional components alone. No significance test was used to establish superiority.

3.1.6. Sensitivity to selection criteria and category boundaries

Changing the transformation-selection criterion from inner macro F1 to kappa did not change any primary outer-test prediction in either domain. Academic selection based on training PC1 variance or cumulative PC1 + PC2 variance also produced identical predictions. For non-academic labels, the two variance-based criteria each yielded 27.78% accuracy, macro F1 of 0.2510, and kappa of -0.0670 (Table 10). Despite their identical aggregate metrics, these alternatives changed 22 and 20 predictions, respectively, relative to the primary procedure. Selection by explained variance therefore did not resolve the low non-academic agreement.

Changing only the training quantile cut points affected individual labels and aggregate scores (Table 11 and Fig. 4). Across the three alternatives, academic macro F1 ranged from 0.6958 to

Table 9: Held-out agreement for baselines and supervised comparisons

Domain	Procedure	n	Accuracy (%)	Macro F1	Kappa
Academic	Primary PCA tertiles	216	73.15	0.7233	0.5982
Academic	Global majority	216	50.46	0.2236	0.0000
Academic	Grade-specific majority	216	76.39	0.5725	0.5865
Academic	Raw-score tertiles	216	73.61	0.7288	0.6051
Academic	Logistic regression: PC1	216	87.96	0.8432	0.8019
Academic	Logistic regression: PC1 + PC2	216	85.19	0.7958	0.7559
Academic	Logistic regression: ≥80% cumulative EVR	216	87.96	0.8432	0.8019
Non-academic	Primary PCA tertiles	144	27.78	0.2377	-0.0904
Non-academic	Global majority	144	44.44	0.2051	0.0000
Non-academic	Grade-specific majority	144	50.00	0.3680	0.1579
Non-academic	Raw-score tertiles	144	30.56	0.2740	-0.0584
Non-academic	Logistic regression: PC1	144	81.25	0.8182	0.6988
Non-academic	Logistic regression: PC1 + PC2	144	80.56	0.8101	0.6885
Non-academic	Logistic regression: ≥80% cumulative EVR	144	82.64	0.8279	0.7221

Table 10: Alternative transformation-selection criteria

Domain	Selection criterion	Accuracy (%)	Macro F1	Kappa	Changed / paired n
Academic	Inner kappa	73.15	0.7233	0.5982	0/216
Academic	Training PC1 EVR	73.15	0.7233	0.5982	0/216
Academic	Training PC1 + PC2 EVR	73.15	0.7233	0.5982	0/216
Non-academic	Inner kappa	27.78	0.2377	-0.0904	0/144
Non-academic	Training PC1 EVR	27.78	0.2510	-0.0670	22/144
Non-academic	Training PC1 + PC2 EVR	27.78	0.2510	-0.0670	20/144

0.7445, with 11-31 changed predictions out of 216. Non-academic macro F1 ranged from 0.2433 to 0.2929, with 8-20 changed predictions out of 144. The widest middle interval, using the 0.25 and 0.75 quantiles, reduced academic accuracy to 67.59% while increasing non-academic accuracy to 39.58%. Non-academic agreement remained low across all tested boundaries. These sensitivity results were not used to replace the primary one-third and two-thirds rule after examining the outer-test outcomes.

Table 11: Sensitivity to training quantile boundaries

Domain	Quantiles	Accuracy (%)	Macro F1	Kappa	Changed / paired n
Academic	1/3, 2/3	73.15	0.7233	0.5982	0/216
Academic	0.30, 0.70	73.61	0.7445	0.6111	14/216
Academic	0.35, 0.65	75.00	0.7302	0.6206	11/216
Academic	0.25, 0.75	67.59	0.6958	0.5328	31/216
Non-academic	1/3, 2/3	27.78	0.2377	-0.0904	0/144
Non-academic	0.30, 0.70	33.33	0.2686	-0.0168	8/144
Non-academic	0.35, 0.65	26.39	0.2433	-0.0890	9/144
Non-academic	0.25, 0.75	39.58	0.2929	0.0604	20/144

3.1.7. Sensitivity to semester coverage and available inputs

Table 12 reports separately refitted analyses restricted by semester, teacher-label consistency, and activity availability. Academic agreement was lower in the semester-1 analysis (macro F1

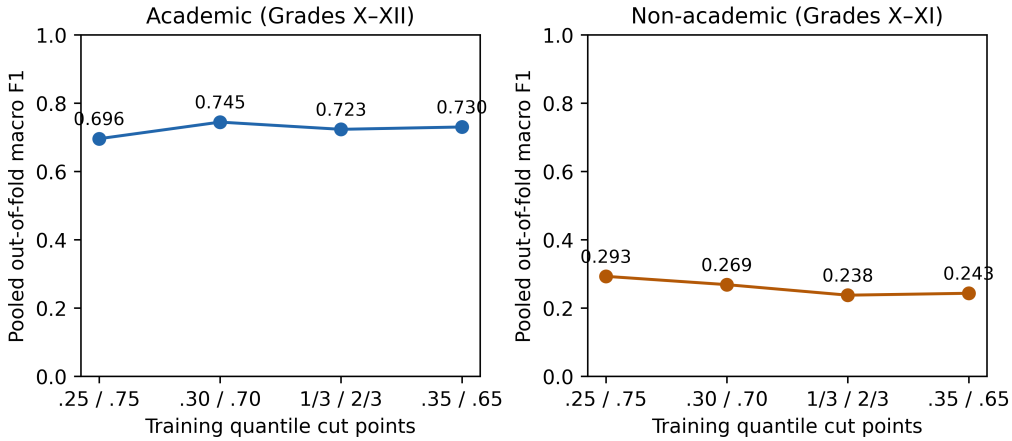


Fig. 4: Pooled macro F1 across training quantile cut points

0.5578) than in semester 2 (0.8579). Restricting the analysis to the 72 students with matching academic teacher labels across semesters retained 144 observations and yielded macro F1 of 0.6553. The corresponding non-academic consistency restriction retained 50 students and 100 observations, with macro F1 of 0.2958 and kappa of -0.0134. These restrictions changed both the development data and the evaluation populations; the differences should not be interpreted as effects of semester or label consistency alone.

Table 12: Refitted analyses of restricted populations and feature sets

Domain	Restriction	Predicted / eligible n	Coverage (%)	Accuracy (%)	Macro F1	Kappa
Academic	Semester 1 only	108/108	100.00	60.19	0.5578	0.4007
Academic	Semester 2 only	108/108	100.00	86.11	0.8579	0.7916
Academic	Matching teacher labels	144/144	100.00	70.83	0.6553	0.5636
Non-academic	Semester 1 only	72/72	100.00	34.72	0.3023	0.0160
Non-academic	Semester 2 only	72/72	100.00	43.06	0.4281	0.1854
Non-academic	Matching teacher labels	100/100	100.00	32.00	0.2958	-0.0134
Non-academic	Exclude PPS	144/144	100.00	29.86	0.2646	-0.0516
Non-academic	Four common activities	28/144	19.44	0.00	0.0000	-0.2874

For a matched-population comparison, the primary academic predictions on the 144 consistency-restricted observations had macro F1 of 0.7119, compared with 0.6553 after refitting on that restricted population. On the 100 corresponding non-academic observations, the values were 0.2760 and 0.2958. These paired summaries distinguish changes in the fitted procedure from comparisons that would otherwise mix different evaluation records.

Excluding PPS retained all 144 non-academic observations and changed 14 predictions relative to the primary procedure. Accuracy was 29.86%, macro F1 was 0.2646, and kappa remained negative at -0.0516. Restricting the input to the four activities present in every Grade X-XI source file produced predictions for only 28 of 144 observations (19.44% coverage). Outer refitting yielded coincident boundaries in folds 1 and 3, while folds 2 and 5 had no valid candidate selected by inner validation. Only fold 4 produced predictions, with no agreement among its 28 records. The resulting 0% accuracy and kappa of -0.2874 describe that successfully predicted subset, not the complete 144-observation population.

3.1.8. Agreement by grade and recorded activity

Table 13 disaggregates the primary held-out predictions without refitting the models. Academic accuracy differed across grades, from 44.44% in Grade XI to 88.89% in Grade X. The apparently high Grade X accuracy coexisted with macro F1 of 0.3137 and kappa of -0.0549. Teacher

categories were highly concentrated by grade: 69 of 72 Grade X observations were Low, while 57 of 72 Grade XII observations were High. Thus, pooled performance and within-grade agreement describe different aspects of the same predictions.

Table 13: Primary prediction agreement within grade and activity-record strata

Domain	Stratum	n	Accuracy (%)	Macro F1	Kappa
Academic	Grade X	72	88.89	0.3137	-0.0549
Academic	Grade XI	72	44.44	0.3986	0.0857
Academic	Grade XII	72	86.11	0.4965	0.4786
Non-academic	Grade X	72	23.61	0.2250	-0.1266
Non-academic	Grade XI	72	31.94	0.2396	-0.0398
Non-academic	At least one B/SB code	118	14.41	0.1656	-0.2869
Non-academic	No B/SB code recorded	26	88.46	0.3129	0.0000

Among the 26 non-academic observations with no recorded B or SB activity code, 23 teacher labels were Medium, two High, and one Low. All 26 primary predictions were Medium, giving 88.46% accuracy but kappa of 0 and macro F1 of 0.3129. Among the remaining 118 observations, accuracy was 14.41%, macro F1 was 0.1656, and kappa was -0.2869. The high accuracy in the no-code stratum therefore reflected its predominantly Medium reference labels and constant predictions, rather than successful separation of three achievement levels. Taken together, the category, component, and sensitivity results constrain the interpretation of the non-academic PC1 labels despite complete coverage in the primary evaluation.

3.2. Discussion

The academic and non-academic contrast shows that variance retention, component interpretation, and agreement with teachers require separate evaluation. PCA maximizes variance rather than agreement with achievement categories [3]. Academic PC1 followed a common grade direction, with positive coefficients and close alignment with the raw academic mean (Fig. 2). However, raw-score tertiles slightly exceeded PCA labeling on macro F1 (Table 9). The additional representation therefore described shared variation without demonstrating an incremental labeling benefit. Its categories remain relative to the training distribution, not established educational standards.

Transformation effects depend on what each procedure changes. StandardScaler, MinMaxScaler, MaxAbsScaler, and RobustScaler alter feature weights through different scaling quantities; Normalizer removes record magnitude, while QuantileTransformer changes marginal distributions [5]. Similar academic feature maxima made the scaling applied by MaxAbsScaler comparatively uniform, plausibly explaining its alignment with the raw mean, although this mechanism was not causally tested. Normalization can remove magnitude relevant to achievement, and quantile mapping cannot recover distinctions absent from tied activity codes. Explained variance across these different spaces consequently provides no universal ranking of labeling quality.

For non-academic data, low PC1 variance retention, weak association with the activity-code sum, and opposing Journalism and English Club coefficients indicate an activity-profile contrast (Table 7; Fig. 3). Sign orientation makes the numerical direction reproducible but does not validate an achievement ordering. Frequent reversals between Low and High teacher categories further undermine that interpretation (Fig. 1). Teachers reportedly considered classroom, school, and organizational participation beyond the recorded extracurricular inputs. This difference in scope may contribute to disagreement, but its effect was not isolated. Zero also combines absent codes with unavailable information; it cannot establish non-participation. High accuracy among all-zero records coexisted with zero kappa (Table 13), illustrating agreement driven by category concentration rather than useful differentiation.

The baseline and sensitivity analyses clarify these limits. The grade-majority baseline ranked differently under accuracy and macro F1, emphasizing the importance of category-sensitive evaluation [13]. Supervised logistic comparators achieved stronger agreement even using non-

academic PC1 alone (Table 9), indicating predictive information that the ordered tertile rule did not exploit. Adding PC2 did not improve agreement; the seven-component alternative produced only a descriptive improvement. These comparators learned teacher-based boundaries and inherited the primary transformation selections, so they do not establish that additional variance alone resolves the interpretation problem. Threshold and selection alternatives changed individual predictions without resolving weak non-academic agreement (Tables 10 and 11). Removing PPS also failed to resolve it, whereas restricting activities sharply reduced prediction coverage (Table 12). These alternatives should not be selected retrospectively from test outcomes [6].

Student-grouped nested validation addressed repeated observations and separated selection from evaluation, but remained internal validation of an already explored dataset from one school and year. Academic category concentration by grade and varying subject aggregates limit pooled interpretation. Teacher assessments were non-overlapping across raters, preventing inter-rater reliability estimation. Recorded semester disagreements also conflicted with the reported annual labeling procedure, and historical assessment dates and a formal rubric were not established. Kappa measures system-teacher agreement rather than objective correctness [8]. Numerical activity-code spacing remains a modeling assumption. No confidence intervals were computed, and no significance tests were performed.

The contribution is an empirical demonstration that reproducible PCA labels still require domain-specific validation of their educational meaning. Current non-academic categories are not supported as an overall student ranking, and academic PCA offers no demonstrated advantage over the raw-score baseline. Future work should distinguish missing records from non-participation, align inputs with explicit assessment rubrics, obtain overlapping independent teacher ratings, and validate new cohorts and schools. Adaptive thresholds and actual clustering methods warrant separate evaluation within development data; their outputs would still require educational interpretation.

4. Conclusion

This study compared six data transformations for PCA-based labeling of academic and non-academic achievement at SMAN 1 Kertosono. Student-grouped nested validation evaluated 216 academic observations from 108 students and 144 non-academic observations from 72 students. Transformation selection, PC1 orientation, and quantile boundary estimation were restricted to development data. The findings show that the suitability of PC1 for ordinal labeling depended on the domain and could not be established from explained variance alone.

For academic labeling, MaxAbsScaler was selected in all five outer development sets. Held-out predictions achieved 73.15% accuracy, a macro F1-score of 0.7233, and Cohen's kappa of 0.5982. Academic PC1 explained 81.07-82.16% of transformed training variance and closely followed the raw mean academic score. However, the raw-score tertile baseline performed slightly better on macro F1. Thus, PCA described a common direction in the recorded grades but did not demonstrate an incremental labeling advantage over the simpler baseline.

Non-academic labeling achieved 27.78% accuracy, a macro F1-score of 0.2377, and a kappa of -0.0904. Transformation choices varied across folds, and PC1 explained only 17.48-27.86% of transformed training variance. Its weak association with the activity-code sum and opposing Journalism and English Club coefficients supported an interpretation as an activity-profile contrast rather than an overall achievement scale. The tested threshold and selection alternatives did not resolve this weakness. Higher agreement from supervised comparators indicated that predictive information about teacher categories remained available, but did not validate the ordinal interpretation imposed by the PC1 tertile rule.

The contribution is an empirical assessment of the conditions under which component-based categories agree with teacher judgments. Explained variance, reproducible sign orientation, and teacher agreement represent distinct properties that must be assessed separately. The results

support a limited descriptive use of academic PC1 and do not support the current non-academic labels as an overall student ranking. No transformation was established as universally superior.

These conclusions are limited to one school and one academic year. Internal validation does not establish performance in new schools or cohorts, and teacher assessments remain a comparison reference rather than an objective standard. Ambiguous activity coding, differences between recorded inputs and teacher assessment scope, inconsistent semester labels, and the absence of overlapping teacher ratings further constrain interpretation. Future research should improve record-keeping and assessment rubrics, obtain independent overlapping ratings, and evaluate new cohorts and schools. Adaptive thresholds and actual clustering algorithms could be investigated as separate alternatives, with their selection confined to development data and their educational meaning independently assessed.

CRediT authorship contribution statement

Salsabila Alya Putri Waluyo: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing–Original Draft, Visualization.
Agung Teguh Wibowo Almais: Conceptualization, Methodology, Validation, Writing–Review and Editing, Supervision, Project Administration.
Cahyo Crysdian: Methodology, Validation, Writing–Review and Editing, Supervision.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5) as an aid for refining the language, organizing the manuscript structure, and enhancing the clarity of the academic writing. All AI-assisted output was reviewed and edited by the authors. The authors bear full responsibility for the accuracy, integrity, and content of the final manuscript. The research methodology, data analysis, interpretation of results, and research conclusions were determined and verified by the authors.

Declaration of competing interest

The authors declare that there are no conflicts of interest that could influence the conduct of the research or the preparation of this manuscript.

Ethics approval and consent to participate

This study did not require approval from an ethics committee. Permission to use the data for research purposes was obtained from SMAN 1 Kertosono. The data were used solely for academic and research purposes. Students' personal identifying information was not disclosed in the manuscript or the published research findings.

Funding and acknowledgments

The authors thank SMAN 1 Kertosono for granting permission to use the data, providing the data, and supporting this study. The authors also thank the teachers who participated in the assessment and validation of the students' academic and non-academic achievement labels, enabling the system-generated labeling results to be evaluated against the teachers' assessments.

Data and code availability

Data supporting the findings of this study are not publicly available because they contain students' educational information. Anonymized data may be obtained from the corresponding author upon reasonable request and subject to approval from the relevant institutions.

The program code used for data preprocessing, the application of six data transformations, principal component analysis (PCA), PC1-based labeling, visualization, and result validation is available from the corresponding author upon reasonable request.

References

[1] C. Romero and S. Ventura. "Educational data mining and learning analytics: An updated survey". In: *WIREs Data Mining and Knowledge Discovery* 10.3, e1355 (2020). DOI: 10.1002/widm.1355.

[2] S. Gunasekara and M. Saarela. "Explainability in Educational Data Mining and Learning Analytics: An Umbrella Review". In: *Proceedings of the 17th International Conference on Educational Data Mining*. 2024, pp. 887–892. DOI: 10.5281/zenodo.12729987.

[3] I. T. Jolliffe and J. Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, 20150202 (2016). DOI: 10.1098/rsta.2015.0202.

[4] A. T. W. Almais et al. "Principal Component Analysis-Based Data Clustering for Labeling of Level Damage Sector in Post-Natural Disasters". In: *IEEE Access* 11 (2023), pp. 74590–74601. DOI: 10.1109/access.2023.3275852.

[5] scikit-learn developers. *Preprocessing data*. Version 1.8.0. https://scikit-learn.org/1.8/modules/preprocessing.html (visited on 10/07/2026).

[6] S. Varma and R. Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7, 91 (2006). DOI: 10.1186/1471-2105-7-91.

[7] S. Kapoor and A. Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science". In: *Patterns* 4.9, 100804 (2023). DOI: 10.1016/j.patter.2023.100804.

[8] J. Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[9] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.

[10] scikit-learn developers. *PCA*. Version 1.8.0. https://scikit-learn.org/1.8/modules/generated/sklearn.decomposition.PCA.html (visited on 10/07/2026).

[11] R. J. Hyndman and Y. Fan. "Sample Quantiles in Statistical Packages". In: *The American Statistician* 50.4 (1996), pp. 361–365. DOI: 10.1080/00031305.1996.10473566.

[12] NumPy developers. *numpy.quantile*. Version 2.3. NumPy. https://numpy.org/doc/2.3/reference/generated/numpy.quantile.html (visited on 10/07/2026).

[13] M. Sokolova and G. Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4 (2009), pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002.

[14] scikit-learn developers. *precision_recall_fscore_support*. Version 1.8.0. https://scikit-learn.org/1.8/modules/generated/sklearn.metrics.precision_recall_fscore_support.html (visited on 10/07/2026).

[15] scikit-learn developers. *LogisticRegression*. Version 1.8.0. https://scikit-learn.org/1.8/modules/generated/sklearn.linear_model.LogisticRegression.html (visited on 10/07/2026).