applied system innovation

MDPI

Article

A Survey-Aware Evaluation Protocol for Feature Selection in Educational Data Mining: Variable-Length Particle Swarm Optimisation Benchmarked on PISA 2018

Odai Amer Hamid [1,2,*], Dalbir Singh [1], Faizan Qamar [3,*] and Azuraliza Abu Bakar [4]

[1] Center for Software Technology and Management, Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; dalbir@ukm.edu.my
[2] College of Technical Management Mosul, Northern Technical University, Mosul 41002, Iraq
[3] Center of Cyber Security, Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia
[4] Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; azuraliza@ukm.edu.my
* Correspondence: oudayamer@ntu.edu.iq (O.A.H.); faizanqamar@ukm.edu.my (F.Q.)

Abstract

Educational data mining (EDM) often reports classifiers and feature selectors fitted to large-scale assessments without respecting how these surveys are designed. This study specifies an evaluation protocol for such data—a published predictor allowlist enforced in code, school-grouped nested cross-validation, outcome categories derived from each of ten plausible values and combined by Rubin's rules, and design-based standard errors from replicate weights for the fitted prediction rules—and uses it to benchmark variable-length particle swarm optimisation (VLPSO) against fifteen alternatives on the Spanish sample of PISA 2018 (29,786 students, 1087 schools). Mathematics proficiency bands were discriminated with an area under the curve of 0.8850 (Low vs. High), 0.7520 (Low vs. Medium) and 0.6963 (Medium vs. High); a label-permutation null averaged 0.4989, and measurement error contributed about half of the variance. Across 12,000 matched selector fits, no method improved on the full predictor set. VLPSO retained four to six of 31 items at a cost of 0.0257–0.0368 in AUC, did not differ from binary swarm optimisation, and was less stable than most filter and embedded selectors. SHAP rankings were concordant across 50 models per task, and models transferred to Portugal without loss. VLPSO compresses predictors at a measurable cost; it does not improve prediction.

Keywords: educational data mining; feature selection; particle swarm optimisation; data leakage; nested cross-validation; plausible values; survey design; explainable machine learning; PISA 2018

Check for updates

Academic Editors: Nikos Antonopoulos, Stamatios Papadakis and Teen-Hang Meen

Received: 16 July 2026
Revised: 8 October 2026
Accepted: 23 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Educational data mining (EDM) applies statistical learning to data generated by educational systems in order to describe, predict, and explain student outcomes [1–3]. Artificial intelligence is increasingly being applied to educational measurement [4], which raises questions of data privacy, algorithmic bias, and fairness [5]. Large-scale assessments such as the Programme for International Student Assessment (PISA) are an attractive source for this work: they combine an achievement measure with several hundred questionnaire variables on home background, attitudes, and school context for nationally representative samples of 15-year-olds [6]. Two methodological threads run through the resulting literature. The

first is feature selection, motivated by the dimensionality of the questionnaire and by the expectation that a compact predictor set is easier to interpret [7,8]. The second is post hoc explanation, principally SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which are used to describe which variables a fitted model relies on [9–11].

Both threads presuppose that the predictive estimates they qualify are sound; on assessment data, however, three properties of the source make this presupposition fragile. First, the achievement scores from which the outcome is derived are stored in the same file as the questionnaire responses, so a table of predictors (a design matrix) built by keeping every numeric column admits the outcome among the predictors; the resulting estimate is then a lookup, not a prediction [12,13]. Second, students are sampled within schools, so cross-validation that splits at the level of the student places a test student's classmates in the training data [14]. Third, proficiency is not observed: PISA reports it as ten plausible values drawn from a posterior distribution, with sampling uncertainty carried by 80 replicate weights; an analysis that averages the plausible values or ignores the weights will report intervals that are too narrow [15–17]. A feature selector evaluated without regard to these properties may be credited with gains that are actually artefacts of the evaluation rather than properties of the selector.

This article is organised around the above dependency. Its primary contribution is an evaluation protocol for supervised learning on large-scale assessment data, specified in enough detail to be re-executed and checked. This includes a published allowlist of candidate predictors enforced by exception, nested cross-validation with schools as the grouping unit, outcome categories derived separately from each plausible value and combined by Rubin's rules, design-based standard errors from the Fay balanced repeated replication (BRR) weights for the fitted prediction rules, two permutation nulls with distinct purposes, and an explanation protocol with attributions replicated across plausible values and folds. Its secondary contribution is an evaluation under the above protocol of variable-length particle swarm optimisation (VLPSO) [18] for selecting questionnaire predictors. The evaluation is performed against fifteen comparators that include filter, wrapper, and embedded selectors, the fixed-length binary particle swarm optimiser (BPSO) [19], four VLPSO ablations, and the unselected allowlist. The VLPSO variant evaluated here is specified at the level of implementation (Section 4.4); the variable-length idea itself is from Tran, Xue, and Zhang [18].

This study addresses four research questions on the Spanish cohort of PISA 2018:

RQ1. *How well can mathematics proficiency bands be discriminated from allowlisted questionnaire items when the evaluation respects the clustering, measurement, and sampling design of the survey?*

RQ2. *Does VLPSO select subsets that discriminate as well as the full allowlist and as well as established selectors, and are its selections stable across folds, plausible values, and random seeds?*

RQ3. *Are the global and local attributions of the fitted models reproducible across plausible values, folds, and explainer seeds?*

RQ4. *Do the fitted models transfer to a second national sample (Portugal) without re-estimation?*

This study makes the following three contributions:

1. **An evaluation protocol for large-scale assessment data.** The protocol of Section 4 combines leakage control by allowlist, school-grouped nested resampling, per-plausible-value outcome derivation with Rubin's rules, and Fay BRR standard errors. Its components are individually established; the contribution is their combination into one

executable pipeline in which each safeguard is enforced by code that raises an exception. A demonstration (Section 5.1) shows the size of the error that each safeguard prevents on these data.

2. **A benchmark of VLPSO under the above protocol.** VLPSO is compared with fifteen alternatives on three binary tasks, ten plausible values, and twenty-five outer folds per plausible value, with identical swarm settings for VLPSO and BPSO, wrapper fitness evaluated on school-grouped folds, and the subset size of ranking methods tuned on inner folds (Section 5.4). The stability of VLPSO is examined across seeds and in a one-at-a-time sensitivity analysis of its control parameters (Section 5.5).
3. **Replicated explanations and external validation.** Global SHAP attributions are computed on 50 fitted models per task (ten plausible values by five folds), and their agreement is quantified. LIME stability is reported with statistics that are not distorted by unnamed features, and the Spanish models are applied unchanged to the Portuguese sample of the same cycle (Sections 5.6 and 5.7).

The rest of the paper is organised as follows: Section 2 reviews related work; Section 3 describes the data, the survey design, and an exploratory analysis; Section 4 specifies the protocol, VLPSO, and comparison design; Section 5 reports the results and Section 6 discusses them; Section 7 states the limitations; and Section 8 concludes the paper. Abbreviations are listed at the end of the article.

2. Related Work

2.1. Prediction of Student Performance

EDM has moved from descriptive analysis towards prediction of attainment, progression, and risk [1,2]. Comparative studies of student performance prediction typically fit several learners, most often logistic regression, decision trees, random forests, gradient boosting, support vector machines, and neural networks, then report performance results [20,21]. Alternatively, some studies first cluster students, then fit a classifier within the selected clusters [22]. Large-scale assessments have been used in the same way, with PISA in particular serving as a source of questionnaire predictors for proficiency outcomes [11,23]. Two features of this literature bear on the present study: first, model choice is frequently made on the same data used to report performance, which yields optimistic estimates [24,25]; second, the survey design of the assessment, namely, the plausible values, sampling weights, and clustering of students in schools, is seldom carried into the evaluation.

2.2. Feature Selection and Particle Swarm Optimisation

Feature selection methods are conventionally grouped into filters, which rank variables by a criterion computed without a learner; wrappers, which score candidate subsets by the performance of a learner; and embedded methods, in which selection is a by-product of fitting [7]. Common filters include symmetric uncertainty [26], mutual information [27], the chi-squared statistic, and ReliefF [28]. Common wrappers include recursive feature elimination (RFE) [29] and sequential forward selection (SFS). Embedded methods include the L_1-penalised (lasso) path [30], impurity-based importance from random forests [31], and Boruta, which compares each feature's importance with that of permuted "shadow" copies [32]. Selection has also been combined with feature extraction, for example by clustering features after a principal component analysis [33].

Evolutionary and swarm methods treat subset selection as a search problem [8]. Particle swarm optimisation (PSO) [34] moves a population of candidate solutions under the attraction of each particle's best position and the swarm's best position; its binary variant (BPSO) maps velocities to bit-flip probabilities through a sigmoid transfer function [19]. In both, every particle has the dimensionality of the full candidate set. Comprehensive

learning PSO (CLPSO) [35] lets each dimension of a particle learn from a different exemplar, which preserves diversity on multimodal problems. Variable-length PSO (VLPSO) [18] combines a ranking of features with particles of different lengths, so that short particles search only the highest-ranked features, then adapts particle lengths during the search; originally, it was proposed for high-dimensional classification problems with thousands of features. To the best of our knowledge, its behaviour on survey questionnaires with tens of candidate items, strong redundancy, and clustered observations has not yet been examined.

The stability of a selector, that is, the extent to which it returns the same subset on perturbed data, is a property distinct from the accuracy of the resulting model. Nogueira, Sechidis, and Brown [36] showed that most stability indices violate desirable properties, including Kuncheva's index [37], which is undefined when the compared subsets differ in size. In [36], the authors proposed an estimator defined for subsets of unequal size and that has a known asymptotic sampling distribution; the same estimator is used here.

2.3. Explainable Machine Learning in Educational Data Mining

SHAP attributes a prediction to features through Shapley values of a cooperative game defined by the model [9]; for tree ensembles, TreeSHAP computes them exactly in polynomial time [38]. LIME fits a sparse linear surrogate to the model's predictions on perturbed samples around one instance [10]. Both have been applied to student performance in programming courses [39], massive open online courses [40], and PISA mathematics literacy [11] as well as to the discovery of student profiles in PISA 2022 [23]. Two properties of these methods are rarely examined in previous work: LIME is stochastic, and its explanations can change between runs on the same instance [41]; meanwhile, a global SHAP ranking computed from a single fitted model inherits the sampling variability of that model. Agreement of rankings across repeated fits can be summarised by Kendall's coefficient of concordance W [42].

2.4. Leakage and Validation Failure

Kaufman et al. [12] provided a formal definition and taxonomy of leakage in data mining. Two of their categories must be kept apart here, as the diagnostics reported in Section 5 test them separately and neither diagnostic tests the other.

Outcome-in-predictors leakage is the presence among features of either the target itself or of a deterministic or near-deterministic function of it. It can easily arise when the table of predictors given to a model (the design matrix) is built by keeping every numeric column of a data file, as a selection rule based on how a value is stored cannot distinguish between a questionnaire response and the achievement score that defines the label, so the score is kept as a predictor. Kapoor and Narayanan [13] have documented that this class of error has produced irreproducible findings across many fields and that it is usually detected by outside replication rather than by the original authors.

Protocol leakage is different in kind. The predictors carry no outcome information, but information flows from the outer test fold into model construction through a scaler or imputer fitted before splitting, features ranked on all rows, hyperparameters tuned on the folds used to report performance, or observations split at the level of the student when they are clustered within schools [14]. Varma and Simon [24] and Cawley and Talbot [25] have shown that selecting models or hyperparameters on the folds used to report performance yields optimistically biased estimates, and that nested cross-validation is the remedy.

The two failure modes require different tests. Permuting labels destroys the association between the outcome and every predictor at once, including a leaked copy of the outcome, so a pipeline that still contains such a copy returns a chance-level metric just as a clean pipeline does. Therefore, a label permutation null is silent on outcome-in-predictors leakage by construction; it tests whether the fitting, tuning, and selection machinery can manufacture above-chance discrimination when no signal remains.

2.5. Analysis of Large-Scale Assessment Data

Using an item response model conditioned on background variables, PISA estimates each student's proficiency from responses to a subset of the item pool and reports ten random draws from each student's posterior distribution; these draws are the plausible values [16,43]. An analysis is run once per plausible value, and the results are combined by Rubin's rules [17]. The point estimate is the mean of the ten estimates, and its variance is the mean within-plausible-value sampling variance plus $(1 + 1/10)$ times the between-plausible-value variance. The fraction of missing information (FMI) is the share of the total variance attributable to the between-plausible-value component, that is, to the measurement error in proficiency. Sampling variance is estimated by balanced repeated replication with Fay's modification: the statistic is recomputed under each of 80 replicate weights, and the variance is $\frac{1}{G(1-k)^2}\sum_{g=1}^{G}(\hat{\theta}_g - \hat{\theta})^2$ with $G = 80$ and Fay factor $k = 0.5$ [15,44]. Unequal weights reduce precision by the Kish design effect $1 + \mathrm{CV}^2$, where CV is the coefficient of variation of the weights [45]. Averaging the plausible values before analysis or analysing only one of them will understate the uncertainty of any estimate that depends on proficiency [15,16].

2.6. Research Gaps

Three gaps follow from the above review. First, feature selectors, VLPSO among them, have been applied to educational data without an evaluation that respects the clustering, measurement, and sampling design of large-scale assessments; thus, it is unknown how much of a reported gain survives such an evaluation. Second, selector stability is rarely reported, and when it is, the indices are undefined for the subsets of unequal size that variable-length methods produce. Third, explanations of models fitted to assessment data are typically computed once from a single fitted model, so their reproducibility is unknown. The present study addresses each of these gaps within a single protocol.

3. Data

3.1. Source and Analytic Sample

Our analysis uses the student questionnaire file of PISA 2018 [46], downloaded from the OECD and read without modification. The school questionnaire file is not used. The primary sample is Spain; the Portugal sample from the same file and cycle is held out for external validation (Section 4.8), and plays no part in any choice made on the Spanish data.

Table 1 gives the exclusion flow. Records were retained where at most 400 of the 1119 columns of the student file were missing. The threshold is consequential, as the sample halves at 200 and grows by a sixth at 600; both alternatives are reported in Table 1. The analytic sample comprises 29,786 students in 1087 schools. Following the same rule, the Portuguese sample comprises 4668 students in 272 schools.

Table 1. Analytic sample flow and survey design of the Spanish sample. Counts of students and schools after each step; the design statistics refer to the primary analytic sample. —, not reported.

Step	Students	Schools
Student file, Spain and Portugal read	41,875	—
Spain	35,943	1089
≤400 of 1119 columns missing (primary)	29,786	1087
≤200 missing (alternative)	14,728	1056
≤600 missing (alternative)	34,953	1089
Portugal, same rule (external validation)	4668	272
Survey design of the primary sample		
Students per school, mean (range)		27.4 (1–42)
Explicit strata		74
Sum of final weights		344,601
Final weight, range		1.00–145.62
Final weight, coefficient of variation		1.065
Kish design effect ($1 + CV^2$)		2.134
Kish effective sample size		13,958
Replicate weights (Fay, $k = 0.5$)		80

3.2. Survey Design

PISA draws schools with probability proportional to size within explicit strata, then samples students within schools [43]. Each student carries a final weight (`W_FSTUWT`), the inverse of the selection probability adjusted for non-response, and 80 replicate weights (`W_FSTURWT1`–`W_FSTURWT80`) constructed for Fay balanced repeated replication with $k = 0.5$ (Section 2.5). The lower panel of Table 1 summarises the design of the analytic sample. The Kish design effect of 2.134 is a weighting effect—it equals $1 + CV^2$ of the final weight—and quantifies the precision lost to unequal selection probabilities. It is not a measure of clustering. Clustering is handled structurally; schools define the boundaries of every cross-validation fold and are the resampling unit of every bootstrap reported here (Sections 4.1 and 4.2).

3.3. Outcome and Plausible Values

The outcome is a band of mathematics proficiency. Each student answers a fraction of the item pool, so proficiency is reported as ten plausible values (`PV1MATH`–`PV10MATH`). For each plausible value separately, a student is assigned to Low if the value is below 482.38 (the lower bound of proficiency Level 3), to High if it is at least 606.99 (the lower bound of Level 5), and to Medium otherwise [6]. Three binary tasks are formed: Low versus High, Low versus Medium, and Medium versus High; in each, the first-named category is coded 0 and the second 1, with students in the third category excluded. Every analysis is run once per plausible value and the ten results are combined by Rubin's rules. Averaging the ten values and thresholding the average is a different procedure; the mean of posterior draws has deflated variance, so the resulting category carries no defensible standard error [16].

3.4. Candidate Predictors and Leakage Guard

Candidate predictors are fixed by an allowlist enforced by an exception: a column enters a design matrix only if it appears in the allowlist, which carries one row per variable with its PISA item code, official label, response scale, and a stated justification. It is reproduced in full as Supplementary Table S1. The allowlist contains 31 student questionnaire items in six blocks (Table 2): possessions related to study and culture (ST011, 13 items), household possessions (ST012, 8), number of books at home (ST013), a situational judgment task on a phishing e-mail (ST166, 5), gender (ST004), and country of birth of the student and

their parents (ST019, 3). A denylist of anchored patterns runs as well, so an allowlist error is still caught, and the guard is called at every point at which a design matrix is built (inside every selector and model); thus, a violation halts the run rather than degrading the estimate silently.

Exclusions and the reasons for them are as follows: the mathematically plausible values, which define the outcome; the reading and science plausible values, which are generated by the same conditioning model and hence are contaminated; the final and replicate weights, which are used for weighting and variance estimation and never as features; identifiers; and the stratum variable, which is a sampling design variable.

Ordinal items keep their original codes and binary items keep their 1/2 codes, which for a two-level item is equivalent to a single indicator column; no one-hot encoding is applied. Missing responses are imputed by the median and accompanied by a missingness indicator per item with any missing value, giving 61 design columns from the 31 items. All columns are standardised. Imputation, indicators, and standardisation are fitted on the training partition of each fold only (Section 4.1).

Table 2. The 31 allowlisted items by questionnaire block: number of items, response scale, range of item missingness in the analytic sample, and range of survey-weighted item means. Item-level statistics for all 31 items can be found in Supplementary Table S2.

Block	Content	Items	Scale	Missing (%)	Weighted Mean
ST011	Possessions for study and culture	13	1 yes, 2 no	0.1–1.4	1.02–1.54
ST012	Household possessions (counts)	8	1 none–4 three or more	0.3–1.9	1.50–3.91
ST013	Books at home	1	1 (0–10)–6 (>500)	0.3–0.3	3.34–3.34
ST166	Phishing e-mail: appropriateness of a response	5	1 not at all–6 very	1.6–2.5	2.50–4.30
ST004	Gender	1	1 female, 2 male	0.0–0.0	1.50–1.50
ST019	Country of birth (student, mother, father)	3	1 Spain, 2 other	0.7–1.9	1.09–1.16

3.5. Item Identification

Several items in these blocks are easy to misidentify from their variable names alone; therefore, every interpreted variable is resolved against the official instrument (PISA 2018 Student Questionnaire, computer-based main survey) and the value labels distributed with the data file. A guard is raised if a variable would be interpreted without a documented source. Four cases matter for the interpretation of the results: ST012Q01TA counts televisions, not books; the number of books is ST013Q01TA. The ST166 items are not measures of device access but ratings of the appropriateness of responses to a phishing e-mail, two of which (ST166Q01HA and ST166Q03HA) are reverse-scored in the sense that endorsing them is the unsafe response. ST004D01T is gender. Finally, the machine-readable codebook covering all 31 items is part of the repository (see Data Availability Statement).

3.6. Exploratory Analysis

For each block, Table 2 reports the response scale, range of item-level missingness, and survey-weighted mean. Table 3 reports the size of each task and its prevalence. The survey-weighted distribution of proficiency, with Fay BRR standard errors, is Low 0.461 (SE 0.0064), Medium 0.459 (SE 0.0054), and High 0.080 (SE 0.0028). The High category holds about one student in twelve, which makes the Medium versus High task the most imbalanced (prevalence 0.162 at plausible value 1), while Low versus High, which excludes the largest band, is the smallest task. The weighted mean of the ten mathematics plausible values ranges from 486.2 to 489.4 across plausible values.

Table 3. The three binary tasks: n and schools at plausible value 1; range of n over the ten plausible values; prevalence is the share of the second-named category, unweighted and survey-weighted.

Task	n (PV1)	n, PV Range	Schools	Prevalence	Weighted
Low vs. High	15,237	15,047–15,419	1084	0.184	0.148
Low vs. Medium	26,977	26,751–27,133	1087	0.539	0.499
Medium vs. High	17,358	17,293–17,660	1074	0.162	0.148

The assignment of a student to a proficiency band is not determined by the data for most students. Across the ten plausible values, 69.69% of students (survey-weighted) are assigned to more than one band; the mean share of the modal band is 0.819, and only 30.31% are assigned to the same band by all ten plausible values. This is the source of the substantial fraction of missing information reported in Section 5.2, and is why task membership (and hence task size) varies across plausible values (Table 3).

4. Methods

Figure 1 summarises the protocol. Its defining property is the order of operations: the data are first split into school-disjoint outer folds. Every step that learns from the data (imputation, scaling, feature ranking, feature selection, hyperparameter tuning, and choice of model family) is fitted on the outer training partition only. The outer test fold is scored once, after all decisions have been taken.

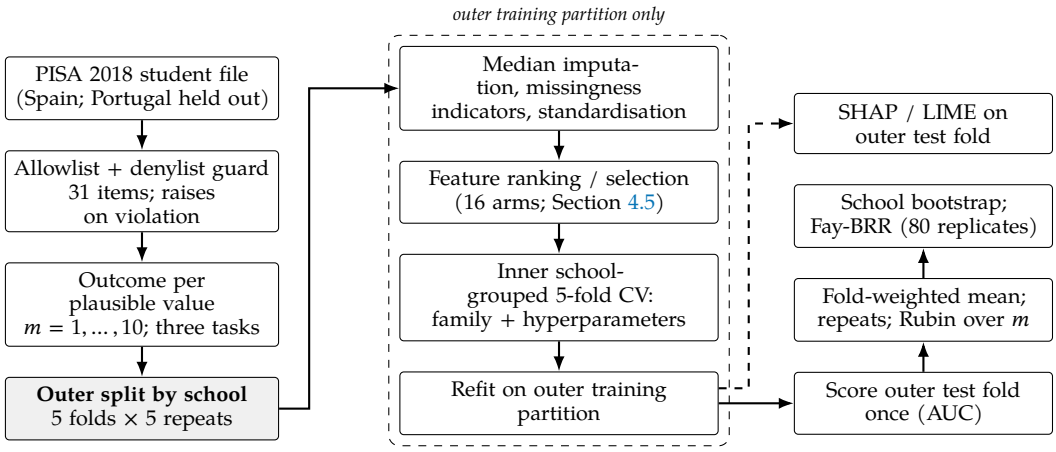


Figure 1. The evaluation protocol. The outer split by school precedes every step that learns from the data; steps inside the dashed box are refitted within each outer training partition, and never see the outer test fold. The procedure is repeated for each of the ten plausible values, five repeats, and five outer folds; the results are combined as described in Section 4.2.

4.1. Nested School-Grouped Cross-Validation

Let $\mathcal{D}_t^{(m)} = \{(\mathbf{x}_i, y_i^{(m)}, g_i)\}$ denote the data for task t under plausible value $m \in \{1, \dots, 10\}$, where $\mathbf{x}_i$ is the vector of 31 allowlisted items for student i, $y_i^{(m)} \in \{0, 1\}$ the binary outcome, and g_i the school identifier (`CNTSCHID`). The outer level partitions the schools into $K = 5$ folds by stratified group K-fold, with the random seed set to $42 + r$ for repeat $r \in \{0, \dots, 4\}$; the same seeds, and hence the same folds, are used by every analysis in this article. The inner level partitions the outer training schools into five folds in the same way. A check is raised if any school appears on both sides of any split.

Within each outer training partition, the inner folds are used to choose among three model families: gradient boosting [47] (100 or 200 trees, learning rate 0.01 or 0.1, depth 2 or 3), random forest [31] (100 or 300 trees, maximum depth 5, 10 or unlimited, minimum leaf size 1 or 10), and L_2-penalised logistic regression ($C \in \{0.01, 0.1, 1, 10\}$), together with

their hyperparameters, by mean inner area under the receiver operating characteristic curve (AUC). The selected configuration is refitted on the whole outer training partition and scored on the outer test fold. Because the family is chosen inside the inner loop, the reported figure refers to this selection procedure and not to any single family; no per-family comparison on the outer folds is reported here. The implementation uses scikit-learn [48].

The performance estimand is the size-weighted mean of the outer fold AUCs. For repeat r and plausible value m,

$$\hat{\theta}_r^{(m)} = \frac{\sum_{k=1}^{K} n_{rk}\,\mathrm{AUC}\left(\mathbf{y}_{rk}^{(m)},\, \hat{f}_{rk}^{(m)}(\mathbf{X}_{rk})\right)}{\sum_{k=1}^{K} n_{rk}}, \qquad \hat{\theta}^{(m)} = \frac{1}{R}\sum_{r=0}^{R-1} \hat{\theta}_r^{(m)}, \tag{1}$$

where n_{rk} is the size of outer test fold k in repeat r, $\mathbf{X}_{rk}$ and $\mathbf{y}_{rk}^{(m)}$ are its predictors and labels, $\hat{f}_{rk}^{(m)}$ is the model selected and fitted on the complementary training partition, and $R = 5$. Out-of-fold predictions are not pooled into a single ranking before the AUC is computed, as each outer fold selects its own model, so a pooled ranking would compare scores produced by different models. On these data, pooling would change the Medium versus High estimate by up to 0.073 (Section 5.2).

4.2. Uncertainty: Plausible Values, Schools, and Survey Weights

Two sources of uncertainty are distinguished. The first is the uncertainty in evaluating fitted prediction rules: given the models selected and fitted on the outer training partitions, the estimate depends on which schools and students form the outer test folds. The second is the uncertainty of the model development process as a whole: repeating preprocessing, feature selection, the choice of model family, and hyperparameter tuning on another sample from the same population would produce different fitted rules. The two interval estimates defined below quantify the first source; neither refits the models within its resamples or replicates. The second source is reflected only in part: in the between-plausible-value component of Rubin's rules, as every model is refitted under the labels of each plausible value, as well as in the spread over the five repeats of the outer partition, which is reported separately (Section 5.2). In the selector comparison, the Nadeau–Bengio correction (Section 4.5) approximates the variance that arises because the training sets, and hence the selected subsets and fitted models, differ between folds.

Resampling interval. For each plausible value, the sampling variance of $\hat{\theta}^{(m)}$ is estimated by a bootstrap that resamples schools with replacement (2000 resamples per repeat), with each resampled school remaining in its outer fold, then recomputes the fold-weighted statistic [49]. The within-plausible-value variance $U^{(m)}$ is the mean of the five per-repeat bootstrap variances. The ten estimates and their variances are combined by Rubin's rules [17]:

$$\begin{aligned} \bar{\theta} &= \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}^{(m)}, & \bar{U} &= \frac{1}{M}\sum_{m=1}^{M} U^{(m)}, \\ B &= \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{\theta}^{(m)} - \bar{\theta}\right)^2, & T &= \bar{U} + \left(1 + \frac{1}{M}\right)B \end{aligned} \tag{2}$$

with $M = 10$, and the interval uses a t reference distribution with Rubin's degrees of freedom $(M-1)\left[1 + \bar{U}/\{(1+1/M)B\}\right]^2$. The fraction of missing information $\mathrm{FMI} = (1+1/M)B/T$ is the share of the total variance contributed by measurement error in proficiency. Because the outer-fold models and their predictions are held fixed within the bootstrap, this interval describes the uncertainty in evaluating the fitted prediction rules on the sampled schools, together with the between-plausible-value variation.

Design-based standard error. The population quantity of interest for a fixed prediction rule is the survey-weighted AUC, in which each positive–negative pair of students is weighted

by the product of their final weights. It is estimated from the outer-test predictions of repeat 0 as the fold-weighted mean of the weighted fold AUCs, and its sampling variance is estimated by Fay BRR. The same statistic is recomputed with each of the $G = 80$ replicate weights in place of the final weight, and

$$\widehat{\mathrm{Var}}_{\mathrm{BRR}}(\hat{\theta}_w) = \frac{1}{G(1-k)^2} \sum_{g=1}^{G} \left(\hat{\theta}_{w,g} - \hat{\theta}_w\right)^2, \qquad k = 0.5. \tag{3}$$

The fitted models are held fixed across replicates, so this variance reflects the sampling of schools and students under the PISA design but not the variability of preprocessing, feature selection, model family choice, or hyperparameter tuning; the ten plausible values are again combined by Rubin's rules. The resulting standard error refers to the discrimination, in the population represented by the survey weights, of the prediction rules fitted in repeat 0, not to the procedure that produced them. The unweighted estimand of Equation (1) is the primary one, as in comparable machine learning studies; the weighted estimand and its BRR standard error are reported alongside it.

4.3. Leakage Diagnostics and Permutation Nulls

Three diagnostics accompany the estimates: (i) to show what the allowlist prevents, every column of the student file is classified against the PISA documentation, and a depth-one decision tree is fitted to each task with the averaged mathematics score as its only input; (ii) for every allowlisted item and task, a single-variable screen computes the AUC of the item on its own, with any item above 0.95 referred for review; and (iii) two label permutation nulls are computed on Low versus High at plausible value 1 under the specification of the main estimate (five outer and five inner school-grouped folds, model family chosen in the inner loop, repeat 0). The unrestricted null permutes labels across the whole sample (100 draws) and tests whether the fitting machinery manufactures discrimination when no signal remains. The within-school null permutes labels only among students of the same school (30 draws), thereby preserving each school's composition; its excess over 0.5 measures the discrimination recoverable from school membership alone. The two nulls use disjoint seed streams. The p-value is $(1 + \#\{b : \hat{\theta}_b \geq \hat{\theta}_{\mathrm{obs}}\})/(1 + B)$.

4.4. Variable-Length Particle Swarm Optimisation

This subsection specifies the VLPSO variant evaluated here exactly as it is implemented; Algorithm 1 gives the complete procedure and Table 4 its settings. Throughout, VLPSO is applied to the standardised design matrix of the outer training partition, for which the $p = 61$ columns are indexed by j.

Ranking and representation. The columns are first ordered by symmetric uncertainty with the label [26],

$$\mathrm{SU}(X_j, Y) = \frac{2\, I(X_j; Y)}{H(X_j) + H(Y)} \in [0, 1], \tag{4}$$

computed on the outer training partition after discretising each column with more than ten distinct values into ten bins of equal frequency, giving an ordering $\pi(1), \ldots, \pi(p)$. Particle i carries an integer length $L_i \in [L_{\min}, p]$, position $\mathbf{x}_i \in [0,1]^{L_i}$, and velocity $\mathbf{v}_i \in \mathbb{R}^{L_i}$. The dimension d of every particle refers to the same column $\pi(d)$, so a particle of length L_i searches the L_i highest-ranked columns, and two particles of different lengths address the same columns over their shared prefix. The subset selected by particle i is

$$S_i = \left\{\pi(d) : \; d \leq L_i, \; x_{i,d} > \tau\right\}, \qquad \tau = 0.5. \tag{5}$$

Objective. The fitness of a subset S is *maximised*:

$$F(S) = \gamma\,\mathrm{BAcc}_{\mathrm{CV}}(S) - \lambda\,\frac{|S|}{p} - \mu\,\mathrm{Red}(S), \qquad \mathrm{Red}(S) = \binom{|S|}{2}^{-1} \sum_{\substack{j<l \\ j,l\in S}} \mathrm{SU}(X_j, X_l), \tag{6}$$

with $\gamma = 0.8$, $\lambda = 0.15$, and $\mu = 0.05$. $\mathrm{BAcc}_{\mathrm{CV}}(S)$ is the balanced accuracy of a five-nearest-neighbour classifier on the columns in S, averaged over three school-grouped folds of a stratified subsample of 3000 students drawn from the outer training partition; the folds are fixed once per search. The second term is a cardinality penalty: of two subsets with equal accuracy and redundancy, the smaller is preferred, and λ is the accuracy the search is willing to give up in order to remove the whole candidate set. The third term penalises redundancy among the selected columns, measured by their mean pairwise symmetric uncertainty. No term of Equation (6) is evaluated on the outer test fold or by resubstitution.

Velocity and position update. At iteration t, for each particle i and each dimension $d \le L_i$,

$$v_{i,d} \leftarrow \omega_t\, v_{i,d} + c_1 r_1 (p_{i,d} - x_{i,d}) + c_2 r_2 (s_{i,d} - x_{i,d}), \qquad v_{i,d} \in [-v_{\max}, v_{\max}], \tag{7}$$

$$x_{i,d} \leftarrow \min\{1, \max\{0,\ x_{i,d} + v_{i,d}\}\}, \tag{8}$$

where $r_1, r_2 \sim U(0,1)$ are drawn independently per dimension, ω_t decreases linearly from 0.9 to 0.4 over the iteration budget, $c_1 = c_2 = 1.49445$, $v_{\max} = 0.5$, and $\mathbf{p}_i$ is the particle's personal best position. The cognitive term is omitted in a dimension that the personal best does not span. Positions are continuous propensities and are mapped to a subset only through the threshold of Equation (5); no sigmoid transfer is applied.

Interaction between particles of unequal length. The social target $s_{i,d}$ is the global best position $\mathbf{g}$ (of length L_g) wherever the two overlap. In a dimension that the global best does not span, it is taken from the personal best of the highest-fitness particle for which the personal best does span d; if no particle spans d, the particle's own personal best is used, and if that does not span d either, the social term is omitted:

$$s_{i,d} = \begin{cases} g_d, & d \le L_g, \\ p_{q,d}, \ \ q = \arg\max\{F_q^{\mathrm{pb}} \ :\ L_q^{\mathrm{pb}} \ge d\}, & d > L_g \text{ and such } q \text{ exists}, \\ p_{i,d}, & \text{otherwise, if } L_i^{\mathrm{pb}} \ge d. \end{cases} \tag{9}$$

In this way, guidance is defined in every dimension that any particle has explored. Unlike CLPSO [35], exemplars are not assigned per dimension by tournament; the global best is the exemplar wherever it is defined.

Length adaptation. Each particle keeps a counter of iterations without improvement of its personal best. When the counter reaches $\beta = 9$, the particle's length is changed by one and the counter is reset; with probability 0.5, the particle grows, otherwise it shrinks within $[L_{\min}, p]$ with $L_{\min} = 2$. On growth, the new dimension $d = L_i + 1$ (column $\pi(L_i + 1)$) receives the current position of the highest-fitness particle that spans d, or a uniform draw if none does, and receives a velocity of zero, since a velocity accumulated in other dimensions carries no information about the new one. On shrinkage, the last dimension is removed from the position and velocity. Lengths can both grow and shrink, so a column dropped early can re-enter the search. The ablation in which lengths can only shrink isolates this mechanism (Section 4.5).

Initialisation and termination. The $P = 60$ particles are spread over $D = 12$ length strata between $L_{\min}$ and p. Under the default ranked initialisation, positions decrease linearly from 0.85 in the first dimension to 0.35 in the last, plus Gaussian noise with standard deviation 0.15, so that higher-ranked columns start with higher selection propensity; the random initialisation ablation draws positions from $U(0,1)$. Velocities are drawn from $N(0, 0.1^2)$.

The search stops after $T = 100$ iterations or when the global-best fitness has not improved by more than 10^{-6} for 15 consecutive iterations. The returned subset is that of the global best. If it contains fewer than $L_{\min}$ columns, the $L_{\min}$ highest-ranked columns are returned instead and the event is logged.

Algorithm 1 VLPSO as implemented (one search on a single outer training partition)

Require: training matrix $\mathbf{X}$ $(n \times p)$, labels $\mathbf{y}$, schools $\mathbf{g}$; settings of Table 4
$\pi \leftarrow$ columns sorted by decreasing $\mathrm{SU}(X_j, Y)$ ▷ Equation (4), training partition only
draw a stratified subsample of 3000 students; fix 3 school-grouped folds on it
for $i = 1, \ldots, P$ **do**
    $L_i \leftarrow$ length from stratum i mod D; initialise $\mathbf{x}_i, \mathbf{v}_i$; $\mathbf{p}_i \leftarrow \mathbf{x}_i$; $F_i^{\mathrm{pb}} \leftarrow F(S_i)$
end for
$\mathbf{g} \leftarrow \mathbf{p}_{i^*}$, $i^* = \arg\max_i F_i^{\mathrm{pb}}$ ▷ maximisation
for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, P$ **do**
        update $\mathbf{v}_i, \mathbf{x}_i$ by Equations (7) and (8) with $s_{i,d}$ from Equation (9)
        $f \leftarrow F(S_i)$ ▷ Equation (6), grouped internal CV
        if $f > F_i^{\mathrm{pb}} + 10^{-6}$ **then** $\mathbf{p}_i \leftarrow \mathbf{x}_i$; $F_i^{\mathrm{pb}} \leftarrow f$; $c_i \leftarrow 0$
        else $c_i \leftarrow c_i + 1$
        end if
        if $c_i \geq \beta$ **then** grow (prob. 0.5) or shrink L_i by one; $c_i \leftarrow 0$
        end if
    end for
    update $\mathbf{g}$ if some F_i^{pb} exceeds $F(\mathbf{g})$ by more than 10^{-6}
    if $\mathbf{g}$ unchanged for 15 iterations **then break**
    end if
end for
return $S_{\mathbf{g}}$ (at least $L_{\min}$ columns)

Table 4. Settings of the swarm searches in the selector comparison. Settings above the rule are shared by VLPSO and BPSO, and are asserted equal in the code base; those below apply to VLPSO only. —, no symbol.

Setting	Symbol	Value
Population size	P	60
Maximum iterations	T	100
Inertia weight (linear decrease)	ω_t	$0.9 \rightarrow 0.4$
Acceleration coefficients	c_1, c_2	1.49445, 1.49445
Objective weights (accuracy, cardinality, redundancy)	γ, λ, μ	0.8, 0.15, 0.05
Wrapper learner	—	5-nearest neighbours, balanced accuracy
Internal folds (school-grouped)	—	3, on a stratified subsample of 3000
Minimum subset size	$L_{\min}$	2
Maximum subset size	—	none (all p columns)
Early stopping	—	no gain $> 10^{-6}$ in 15 iterations
Length strata at initialisation	D	12
Selection threshold	τ	0.5
Velocity bound	$v_{\max}$	0.5
Stagnation threshold for length change	β	9
Growth probability; length step	—	0.5; 1
Feature ranking	—	symmetric uncertainty, 10 bins

4.5. Design of the Selector Comparison

Sixteen arms are compared (Table 5): the unselected allowlist; four filters (symmetric uncertainty, mutual information, chi-squared, ReliefF); two wrappers (RFE with logistic regression, SFS); three embedded methods (L_1 path, random forest impurity importance, Boruta); BPSO; VLPSO; and four VLPSO ablations. Each changes one element: no cardi-

nality penalty ($\lambda = 0$), shrink-only length adaptation, random initialisation, and mutual information ranking instead of symmetric uncertainty.

Every arm is run on the same outer folds as the headline estimate for all three tasks, all ten plausible values, five repeats, and five folds, giving 12,000 selector fits. Within each outer training partition, the preprocessing of Section 3.4 is fitted, the selector is fitted on the transformed training data, and a single downstream learner (L_2-penalised logistic regression with balanced class weights) is fitted on the selected columns, with $C \in \{0.01, 0.1, 1, 10\}$ chosen on five school-grouped inner folds. The AUC of that model on the outer test fold is the arm's score. Using one downstream learner for every arm isolates the selector from the choice of model; the consequence, that the absolute level of this comparison differs from the headline estimate (Section 5.2), which selects among three families, is stated where the two are compared.

Three requirements make the comparison fair. First, every wrapper evaluates candidate subsets on school-grouped folds of the training partition (stratified group K-fold on `CNTSCHID`), and the code raises if groups are not supplied. Second, BPSO and VLPSO share every setting they have in common—population, iterations, inertia schedule, learning coefficients, objective and its weights, wrapper learner, internal folds, subsample, minimum subset size and stopping rule—and neither is capped in cardinality; a test in the code base asserts this equality. BPSO uses the sigmoid transfer of Kennedy and Eberhart [19] over all p columns. Third, the ranking methods (the four filters, RFE, L_1 path, and tree importance) return a ranking, and the number of columns kept is chosen from $k \in \{5, 10, 15, 20, 30\}$ by the inner-fold AUC of the downstream learner, refitting the ranking within each inner training partition; SFS stops by its own criterion (at most 30 columns, five steps without improvement of the grouped inner AUC) and Boruta by its own statistical test (30 iterations, Bonferroni correction at $\alpha = 0.05$). The upper end of the grid, 30 of the 61 design columns, and the 30-column limit of SFS are caps on subset size that the two swarm searches do not have; in these data, a subset of 30 columns comprises between 13 and 30 items, since an item and its missingness indicator count as one item.

Table 5. The sixteen arms of the selector comparison. "Tuned k": the number of columns is chosen from $\{5, 10, 15, 20, 30\}$ on school-grouped inner folds, so that at most 30 of the 61 columns are kept.

Family	Arm	Subset Size and Notes
Reference	None	all 61 columns
Filter	Symmetric uncertainty [26]	tuned k; 10-bin discretisation
Filter	Mutual information [27]	tuned k; nearest-neighbour estimator
Filter	Chi-squared	tuned k; columns shifted to be non-negative
Filter	ReliefF [28]	tuned k; 10 neighbours, 500 sampled instances
Wrapper	RFE [29]	tuned k; L_2 logistic regression, one column removed per step
Wrapper	SFS	own stopping: $\leq$30 columns, patience 5, grouped inner AUC
Embedded	L_1 path [30]	tuned k; order of entry along the lasso path
Embedded	Tree importance [31]	tuned k; random forest, 200 trees, impurity decrease
Embedded	Boruta [32]	own test: 30 iterations, 100 trees of depth 7, Bonferroni $\alpha = 0.05$
Swarm	BPSO [19]	own search; Table 4
Swarm	VLPSO (Section 4.4)	own search; Table 4
Ablation	VLPSO, $\lambda = 0$	cardinality penalty removed
Ablation	VLPSO, shrink-only	lengths can only decrease
Ablation	VLPSO, random initialisation	positions drawn from $U(0, 1)$
Ablation	VLPSO, MI ranking	mutual information replaces symmetric uncertainty

Statistical comparison. For each task, plausible value, and arm, the 25 outer-fold AUCs are averaged and their variance is estimated with the Nadeau–Bengio correction for overlapping training sets [50], $\left(\frac{1}{J} + \frac{n_{\text{test}}}{n_{\text{train}}}\right)s^2$ with $J = 25$. This is because the naive variance of resampled estimates is anti-conservative; the correction is intended to approximately account for the variability of the training sets between folds, and hence that of the selected subsets and fitted models. Estimates are combined over plausible values by Rubin's rules. Paired contrasts are computed on matched folds in the same way, against two references—VLPSO and the unselected allowlist—and the p-values within each task and reference are adjusted by Holm's procedure [51]. The standardised effect d is the mean paired difference divided by the within-plausible-value standard deviation of the differences, averaged over plausible values, and is described as negligible ($|d| < 0.2$), small (0.2–0.5), medium (0.5–0.8), or large ($\geq$0.8). A (task, plausible value) block enters the comparison only if every arm has every fold, so all contrasts are on identical folds.

Stability. For each arm, the 25 subsets selected within a plausible value are summarised by the stability estimator of Nogueira et al. [36], which equals 1 for identical subsets, has expectation 0 under random selection of subsets of the same sizes, and is defined for subsets of unequal size. It is computed over items (a column and its missingness indicator count as one item) and averaged over plausible values. The mean pairwise Jaccard index is reported for comparison with earlier work.

4.6. Stability and Sensitivity of the Swarm Searches

To examine the variability of VLPSO separately from the variability of the data, VLPSO and BPSO are each run with ten random seeds on each of the five outer folds of Low versus High at plausible value 1 (repeat 0), with all other settings as in Table 4. Variability across seeds within a fold is attributable to the search; variability across folds for a fixed seed includes the data. A one-at-a-time sensitivity analysis then varies a single setting of VLPSO from its comparison value on the same five folds: population 20 or 100; iteration budget 30 or 200; $\lambda \in \{0, 0.05, 0.30\}$; threshold $\tau \in \{0.40, 0.60\}$; stagnation threshold $\beta \in \{5, 15\}$; and fitness subsample 1500 or 6000. The global-best fitness at every iteration is recorded for every swarm run.

4.7. Explanation Protocol

Explanations are computed for the models of the headline procedure on repeat 0: for each task, plausible value, and outer fold, the inner loop of Section 4.1 is rerun on the outer training partition, which reproduces the selected model and its outer-fold AUC, and the fitted model is explained on its outer test fold. This gives 50 fitted models per task.

Global SHAP. Attributions are computed on the log-odds margin, by TreeSHAP [38] for tree ensembles and by the linear explainer for logistic regression, with interventional feature perturbation against a background of 200 k-means centroids of the transformed outer training partition (the tree explainer of SHAP 0.52 uses a random subsample of 100 of them, with which the base value is also computed). They are computed for 2000 outer-test students drawn by stratified sampling on the true label. The global importance of a column is its mean absolute attribution, with a 1000-resample bootstrap interval over students. Across the 50 models of a task, a column's importance is summarised by its mean, its between-plausible-value standard deviation and the median and interquartile range of its rank, and the agreement of the 50 rankings by Kendall's W [42], which is 1 for identical rankings and 0 for no agreement. The signed mean attribution is not used to infer direction; direction is read from the sign of the single-variable association (Section 4.3).

Local SHAP and LIME. For each model at plausible value 1, six students are chosen at fixed quantiles (0.05, 0.25, 0.50, 0.75, 0.95, 0.99) of the predicted probability, rather than

at its extremes. Each is explained by local SHAP and by LIME [10], the latter 30 times with different seeds (5000 perturbation samples, 15 features per explanation, probability scale). For each feature and student, the selection frequency is the share of seeds for which explanation names the feature; its weight is summarised by the mean, standard deviation, and coefficient of variation over the seeds that name it, and its sign consistency is the share of those seeds that agree on the sign. A feature that a seed does not name has no weight in that seed, and is excluded rather than being counted as zero, since counting it as zero would confound how often a feature is named with how variable its weight is. Rank agreement between seeds is the mean pairwise Spearman correlation over the features named by both seeds.

Every interpreted variable is resolved against the official codebook (Section 3.5). Attributions are associational statements about the fitted models, and are not interpreted causally (Section 6).

4.8. External Validation Design

For each task and plausible value, the inner procedure (model family and hyperparameters chosen on five school-grouped folds) is applied to the complete Spanish task sample, and the selected model without any refitting or recalibration is applied to the Portuguese students of the same task and plausible value. The Spanish and Portuguese school sets are disjoint by construction. Unweighted and survey-weighted AUCs are computed, their variances are estimated by a 2000-resample bootstrap over Portuguese schools, and the ten plausible values are combined by Rubin's rules.

4.9. Software, Determinism, and Reproducibility

All analyses were run in Python 3.12 with scikit-learn 1.6.1, NumPy, pandas, SHAP 0.52, and LIME 0.2. Every unit of work (one selector on one fold, one explained model, one permutation draw) is a "cell" with an identifier from which its random seed is derived, and is written atomically together with the configuration fingerprint and software versions; a reconciliation step counts completed cells against the configured design, and every table is generated from completed cells only. The headline estimate (Section 5.2) was produced in a separate run of the same code on the same folds; the explanation cells refit the same models, and the agreement of their outer-fold AUCs with the stored headline values is reported in Section 5.6 as an end-to-end reproduction check. The allowlist, denylist, codebook, seeds, grids, and all configuration files are in the repository (see the Data Availability Statement) together with the per-cell outputs of the selector, stability, permutation, explanation, external validation and replicate weight analyses, and the commands that reassemble them into the corresponding tables. Because cells were distributed over four machines in two data centres, 1157 selector and stability cells happened to be computed twice on different machines; all 1157 pairs returned identical AUCs and identical selected subsets.

5. Results

5.1. Leakage Diagnostics

The analysis frame has 1122 columns, consisting of the student file and the columns derived from it at ingestion. Classifying them against the PISA documentation gives 936 admissible questionnaire variables, 83 design or weight variables, 80 outcome-derived columns (the reading and science plausible values), 12 outcome columns (the mathematics plausible values and the scores derived from them), and 11 identifiers. All 92 outcome and outcome-derived columns are numeric and pass a filter keyed on storage type. A depth-one decision tree using the averaged mathematics score as its only input reproduces every task label exactly (accuracy 1.0000 on all three tasks), as each label is a threshold on that input.

A classifier given such a design matrix performs a lookup; its accuracy is at the ceiling, and does not change when the model is changed.

Under the allowlist, no such column can enter. The single-variable screen (Table 6) found no allowlisted item above the 0.95 ceiling on any task; the largest single-item AUC was 0.7472, for the number of books at home (ST013Q01TA). This magnitude is typical of a questionnaire item rather than a leaked outcome.

Table 6. Single-variable discrimination screen: the five allowlisted items with the largest AUC on Low vs. High, with their values on the other tasks (plausible value 1). The review ceiling was 0.95; no item in any task exceeded it. For reverse-scored items, the unsigned AUC is shown.

Item	Content	Low vs. High	Low vs. Med.	Med. vs. High
ST013Q01TA	books at home	0.7472	0.6540	0.6108
ST166Q03HA	phishing: click the link (rev.)	0.7273	0.6211	0.6099
ST166Q05HA	phishing: check operator site	0.6730	0.5981	0.5773
ST166Q02HA	phishing: check sender	0.6714	0.6081	0.5669
ST012Q06NA	computers	0.6419	0.5841	0.5606
Items above 0.95		0	0	0

5.2. Predictive Performance

Table 7 answers RQ1. The AUC was 0.8850 (95% CI 0.874–0.896) for Low versus High, 0.7520 (0.743–0.761) for Low versus Medium, and 0.6963 (0.679–0.714) for Medium versus High. Discrimination falls as the two categories become adjacent, which is the ordering a genuine signal would produce; a leaking design matrix returns values at the ceiling for all three tasks alike. The inner loop chose gradient boosting in 728 of the 750 outer folds, random forest in 22, and logistic regression in 0.

The fraction of missing information lies between 47.0% and 60.0%. Roughly half of the variance of each estimate is due to measurement error in proficiency rather than to the sampling of schools. An analysis that used a single plausible value or their average would report intervals that are too narrow, and would attribute none of their width to this source. The repeat-to-repeat spread (0.00059–0.00123) is an order of magnitude smaller than the interval widths, so re-partitioning the same schools moves the estimate little.

Had the out-of-fold predictions of a repeat been pooled into a single ranking, the Medium versus High estimate would have differed from the fold-weighted value by up to 0.0730 in a single (plausible value, repeat) cell against at most 0.0010 and 0.0003 for the other tasks, since the models selected in different folds produce scores on different scales.

Table 7. Discrimination under the protocol (750 outer folds: 3 tasks × 10 plausible values × 5 repeats × 5 folds). AUC: fold-weighted mean (Equation (1)), averaged over repeats and combined over plausible values by Rubin's rules, with a 95% interval from the school bootstrap. FMI: fraction of missing information. Repeat spread: mean within-plausible-value standard deviation over the five repeats. Weighted AUC: survey-weighted AUC of the repeat-0 models with its Fay-BRR standard error (80 replicates), combined over plausible values. Both the 95% interval and the Fay-BRR standard error hold the fitted outer-fold models fixed (Section 4.2). *n* and schools at plausible value 1.

Task	*n*	Schools	AUC	95% CI	FMI	Repeat Spread	Weighted AUC (SE)
Low vs. High	15,237	1084	0.8850	0.8740–0.8960	0.568	0.00087	0.8823 (0.0069)
Low vs. Medium	26,977	1087	0.7520	0.7432–0.7608	0.470	0.00059	0.7487 (0.0059)
Medium vs. High	17,358	1074	0.6963	0.6792–0.7135	0.600	0.00123	0.6960 (0.0115)

Two survey-weighted analyses address the population represented by the weights. Applying the final student weight to the metric and the fold weighting gives 0.8824

(0.8685–0.8962), 0.7490 (0.7380–0.7599), and 0.6960 (0.6719–0.7201); every weighted estimate lies inside the corresponding unweighted interval. The design-based analysis of Equation (3) gives weighted AUCs of 0.8823, 0.7487 and 0.6960 with Fay BRR standard errors of 0.0069, 0.0059, and 0.0115 (95% intervals 0.869–0.896, 0.737–0.760, and 0.673–0.719, combined over plausible values). Both kinds of interval hold the fitted models fixed and describe the evaluation of the fitted prediction rules (Section 4.2), but differ in their reference populations. The bootstrap interval of Table 7 refers to the schools of the cross-validated sample, whereas the BRR standard error refers to the population of 15-year-olds in Spanish schools represented by the survey weights.

5.3. Permutation Nulls

Table 8 reports both nulls at the configured budgets. The unrestricted null averaged 0.4989 (SD 0.0068, 100 draws), and no draw exceeded 0.5125. Its mean minus 0.5 is −0.0011, with a standard error of 0.0007, so the optimism that the fitting, tuning, and model selection machinery could contribute is bounded above by about 0.0003 in AUC (mean minus 0.5 plus two standard errors). The observed value under the same specification was 0.8845, exceeding every draw of both nulls, so $p = 0.0099$ and $p = 0.0323$, which are the smallest values attainable with 100 and 30 draws.

The within-school null averaged 0.6188 (SD 0.0046), so keeping each school's composition while shuffling students within schools leaves a substantial part of the discrimination intact. Comparing the two nulls with the observed value decomposes the 0.386 by which the observed AUC exceeds the unrestricted null ($0.885 - 0.499$) into 0.120 ($0.619 - 0.499$; 31.1%) that is recoverable from knowing which school a student attends and 0.266 ($0.885 - 0.619$; 68.9%), which requires students within a school to be distinguished. The two nulls were drawn from disjoint seed streams, and are independent.

Table 8. Permutation nulls for Low vs. High at plausible value 1, computed under the specification of the headline estimate (five outer and five inner school-grouped folds, model family chosen in the inner loop, repeat 0). The two nulls use disjoint seed streams. $p = (1 + \#\{\text{draws} \geq \text{observed}\})/(1 + \text{draws})$. —, not applicable (single fit on the observed labels).

Permutation Scheme	Draws	Mean AUC	SD	2.5–97.5%	Max	*p*
Unrestricted (fitting machinery test)	100	0.4989	0.0068	0.4850–0.5104	0.5125	0.0099
Within-school (school composition kept)	30	0.6188	0.0046	0.6127–0.6271	0.6278	0.0323
Observed, same specification	—	0.8845				

5.4. Comparison of Feature Selectors

All 12,000 selector fits were completed (3 tasks × 10 plausible values × 25 outer folds × 16 arms), so every contrast below is on identical folds. Table 9 reports the AUC of the downstream logistic regression, the number of items selected, and the stability of the selection. Table 10 reports the paired contrasts, and Figure 2 plots discrimination against subset size. The absolute level is below that of Table 7 because one regularised linear learner is used for every arm, whereas the headline procedure chose gradient boosting in almost every fold.

No selector improved on the full allowlist. No arm exceeded the unselected allowlist by a Holm-significant margin on any task. Three methods—RFE, SFS, and the L_1 path—matched it to within 0.0006 in AUC (all Holm-adjusted $p \geq 0.54$) while keeping between 17.2 and 21.6 of the 31 items on average, a number set largely by the maximum subset size (see below). The filters and the remaining embedded methods lost at most 0.0122, a loss that was Holm-significant for several of them on Low versus Medium, the largest task.

Table 9. Selector comparison under a common protocol: AUC of the downstream L_2 logistic regression on the outer test folds (25 folds per plausible value, combined over plausible values by Rubin's rules), mean number of items selected (of 31), and stability $\hat{\Phi}$ of Nogueira et al. [36] over items (mean over plausible values; 1 = identical subsets, 0 = chance). —, not defined, because no selection is made.

	Low vs. High			Low vs. Medium			Medium vs. High		
Arm	**AUC**	**Items**	$\hat{\Phi}$	**AUC**	**Items**	$\hat{\Phi}$	**AUC**	**Items**	$\hat{\Phi}$
None (full allowlist)	0.8747	31.0	—	0.7410	31.0	—	0.6905	31.0	—
Symmetric uncertainty	0.8732	23.4	0.93	0.7355	22.6	0.92	0.6921	22.5	0.81
Mutual information	0.8691	23.5	0.48	0.7346	21.9	0.35	0.6845	19.1	0.16
Chi-squared	0.8625	20.5	0.91	0.7337	19.1	0.94	0.6865	18.7	0.84
ReliefF	0.8728	27.2	0.72	0.7361	22.5	0.17	0.6914	27.3	0.73
RFE	0.8743	20.8	0.79	0.7405	21.6	0.80	0.6906	17.2	0.70
SFS	0.8741	21.3	0.65	0.7403	21.1	0.72	0.6902	18.9	0.53
L_1 path	0.8745	21.6	0.76	0.7403	21.5	0.76	0.6911	17.8	0.75
Tree importance	0.8731	27.9	0.95	0.7377	27.9	0.96	0.6914	27.4	0.86
Boruta	0.8703	22.0	0.98	0.7321	20.3	0.96	0.6892	13.7	0.95
BPSO	0.8541	9.7	0.31	0.7207	9.1	0.23	0.6521	8.6	0.09
VLPSO	0.8414	6.1	0.45	0.7153	5.5	0.49	0.6537	4.4	0.37
VLPSO, $\lambda = 0$	0.8644	16.0	0.34	0.7289	13.3	0.33	0.6760	12.9	0.22
VLPSO, shrink-only	0.8420	6.0	0.46	0.7149	5.5	0.48	0.6523	4.2	0.37
VLPSO, random init.	0.8385	5.6	0.45	0.7123	5.2	0.48	0.6466	4.0	0.30
VLPSO, MI ranking	0.8430	5.6	0.50	0.7144	5.3	0.50	0.6506	3.9	0.34

Table 10. Paired contrasts on matched outer folds (method minus reference, in AUC) against the full allowlist and against VLPSO. Nadeau–Bengio corrected variance within plausible values, Rubin's rules across them; * Holm-adjusted $p < 0.05$ within task and reference. Intervals, p-values, and effect sizes are reported in Supplementary Table S3. —, not applicable (VLPSO is the reference).

	vs. Full Allowlist			vs. VLPSO		
Arm	**L–H**	**L–M**	**M–H**	**L–H**	**L–M**	**M–H**
Symmetric uncertainty	−0.0015	−0.0054 *	+0.0017	+0.0319	+0.0202 *	+0.0385 *
Mutual information	−0.0056	−0.0064 *	−0.0059	+0.0277	+0.0193 *	+0.0309
Chi-squared	−0.0122 *	−0.0072 *	−0.0040	+0.0211	+0.0184 *	+0.0328
ReliefF	−0.0019	−0.0048 *	+0.0010	+0.0315	+0.0208 *	+0.0378 *
RFE	−0.0004	−0.0005	+0.0001	+0.0330	+0.0252 *	+0.0369 *
SFS	−0.0006	−0.0006	−0.0002	+0.0327	+0.0250 *	+0.0366 *
L_1 path	−0.0002	−0.0006	+0.0006	+0.0332	+0.0250 *	+0.0374 *
Tree importance	−0.0016	−0.0033 *	+0.0009	+0.0318	+0.0224 *	+0.0377 *
Boruta	−0.0044	−0.0089 *	−0.0013	+0.0289	+0.0168 *	+0.0355
BPSO	−0.0206 *	−0.0203 *	−0.0383	+0.0127	+0.0054	−0.0016
VLPSO	−0.0334	−0.0257 *	−0.0368	—	—	—
VLPSO, $\lambda = 0$	−0.0104	−0.0121 *	−0.0144	+0.0230	+0.0136	+0.0224
VLPSO, shrink-only	−0.0327 *	−0.0260 *	−0.0382	+0.0007	−0.0004	−0.0014
VLPSO, random init.	−0.0362 *	−0.0287 *	−0.0438 *	−0.0029	−0.0030	−0.0070
VLPSO, MI ranking	−0.0317 *	−0.0266 *	−0.0399 *	+0.0017	−0.0009	−0.0031

VLPSO selected the smallest subsets at the largest cost. VLPSO kept between 4.4 and 6.1 items on average and lost −0.0334, −0.0257, and −0.0368 in AUC relative to the full allowlist (95% intervals −0.0566 to −0.0101, −0.0366 to −0.0147, and −0.0617 to −0.0118; Holm-adjusted $p = 0.0500$, 0.0001, and 0.0502). It ranked 15th, 13th, and 12th of the 16 arms. BPSO, run with identical settings, kept more items (9.7, 9.1, and 8.6) and differed from VLPSO by +0.0127, +0.0054, and −0.0016 (Holm-adjusted $p = 1$ in each task): the variable-length representation neither helped nor hurt discrimination relative to fixed-length search

under matched conditions. On Low versus Medium, every non-swarm arm, including the full allowlist, exceeded VLPSO after Holm correction (ten arms); on Medium versus High, seven did; on Low versus High, none did after correction, although the point differences were of similar size (0.0211 to 0.0334) and their intervals were wider.

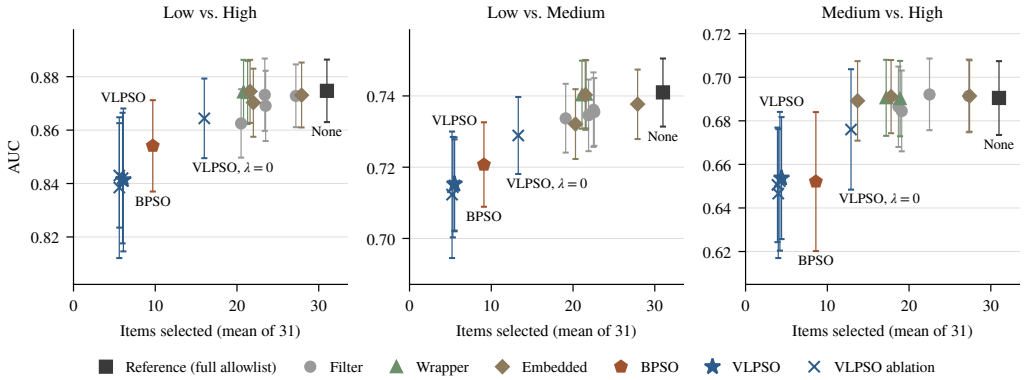


Figure 2. Discrimination against subset size for the 16 arms. AUC of the downstream logistic regression (Rubin's rules over ten plausible values, 25 outer folds each) with 95% intervals, against the mean number of items selected.

The ablations locate the cost in the cardinality penalty. Removing the penalty ($\lambda = 0$) recovered most of the loss; the unpenalised variant kept 16.0, 13.3, and 12.9 items and fell 0.0104–0.0144 below the full allowlist. The other three ablations (shrink-only length adaptation, random initialisation, and a mutual-information ranking) were indistinguishable from VLPSO (largest absolute difference 0.0070, all Holm-adjusted $p = 1.00$). Neither bidirectional length adaptation nor the symmetric uncertainty ordering made a measurable contribution on these data.

Stability. The filters and embedded methods returned largely the same items across the 25 folds of a plausible value ($\hat{\Phi}$ from 0.70 to 0.98, with the exceptions of mutual information, for which the nearest-neighbour estimator is itself stochastic, ReliefF on Low versus Medium, and SFS). The swarm methods were among the least stable ($\hat{\Phi}$ from 0.09 to 0.50); among them, VLPSO was more stable than BPSO on every task ($\hat{\Phi}$ = 0.45, 0.49, and 0.37 against 0.31, 0.23, and 0.09). The items selected by either swarm are not interpreted educationally; item-level selection frequencies for every arm and task are given in Supplementary Table S4.

Maximum subset size. For the seven ranking methods, the inner loop chose the largest k in the grid (30 of 61 columns) in 92% of the 5250 tuned folds (98% on Low versus High, 99.8% on Low versus Medium, and 78% on Medium versus High), and SFS stopped at its 30-column limit in 82% of folds. For RFE and the L_1 path, the cap was reached in 94–100% of folds on the first two tasks and in 40–46% on Medium versus High. Therefore, the cap was binding for most fits of these arms, which has three consequences for the comparison. First, the subset sizes reported for the ranking methods and SFS in Table 9, including the 17.2–21.6 items retained by RFE, SFS, and the L_1 path, are determined largely by the cap and are not estimates of an optimal subset size. Second, the finding that no selector improved on the full allowlist holds for the configurations evaluated: subsets of 31 to 60 columns were not evaluated for these arms, and whether one of them would exceed the full allowlist is not established by these data. Third, the contrasts between these arms and VLPSO compare capped procedures that retained 17.2 to 27.9 items on average with an uncapped search that retained 4.4 to 6.1; they compare the selection procedures as configured, not selectors at a common subset size.

Computational cost. A swarm search took 73–137 s per fold on average, against at most 53 s for any other selector.

5.5. Stability, Sensitivity, and Convergence of the Swarm Searches

Across ten seeds on a fixed fold (Low versus High, plausible value 1), VLPSO returned subsets for which the stability across seeds averaged $\hat{\Phi} = 0.50$ (range over folds 0.44–0.58), against 0.32 (0.27–0.34) for BPSO under identical settings. VLPSO selected fewer items (mean 6.9, range 2–12, against 10.5, 7–16) and its downstream AUC varied more across seeds (mean within-fold standard deviation 0.0102 against 0.0066). The ranked-prefix encoding makes the identity of the selected items more reproducible than in BPSO, since every particle searches the same highest-ranked columns; however, the small subsets it converges to make the resulting discrimination more sensitive to those few items that are kept.

Table 11 shows that one setting dominates. The cardinality weight λ moves the mean subset size from 17.2 items ($\lambda = 0$) through 6.8 ($\lambda = 0.15$) to 4.4 ($\lambda = 0.30$), and the mean AUC from 0.8687 through 0.8521 to 0.8419. No other one-at-a-time change moved the mean AUC outside 0.8403–0.8563 or the mean subset size outside 6.0–8.8 items. Doubling the iteration budget did not change the result (0.8538, 8.0 items) and the searches ended by the stopping rule after a mean of 35 iterations at the comparison setting, so the swarm converges well within the budget; enlarging the population to 100 (0.8475) or doubling the fitness subsample (0.8563) did not materially improve the downstream AUC. Therefore, the compactness of VLPSO and the price paid for it are set by λ rather than by an insufficient search.

Table 11. One-at-a-time sensitivity of VLPSO on Low vs. High, plausible value 1, the five outer folds of repeat 0 (seed 0). AUC of the downstream logistic regression: mean and SD over folds; items: mean (range); $\hat{\Phi}$: stability across the five folds; stop: mean iteration at which the search ended; s: mean selector time in seconds.

Setting	AUC	SD	Items	$\hat{\Phi}$	Stop	s
Comparison setting (Table 4)	0.8521	0.0175	6.8 (3–9)	0.55	35	147
$P = 20$	0.8543	0.0184	8.8 (6–11)	0.44	35	50
$P = 100$	0.8475	0.0222	7.0 (4–11)	0.39	33	214
$T = 30$	0.8524	0.0228	7.2 (3–11)	0.49	23	85
$T = 200$	0.8538	0.0165	8.0 (6–11)	0.49	31	119
$\lambda = 0$	0.8687	0.0133	17.2 (11–20)	0.35	57	176
$\lambda = 0.05$	0.8583	0.0147	10.0 (9–11)	0.54	28	109
$\lambda = 0.30$	0.8419	0.0146	4.4 (3–5)	0.58	20	68
$\tau = 0.40$	0.8508	0.0196	7.0 (3–11)	0.50	37	175
$\tau = 0.60$	0.8403	0.0386	6.0 (3–9)	0.38	35	124
$\beta = 5$	0.8500	0.0184	6.4 (3–9)	0.55	32	134
$\beta = 15$	0.8554	0.0164	8.6 (7–11)	0.53	25	78
fitness subsample 1500	0.8536	0.0231	7.6 (5–13)	0.41	38	55
fitness subsample 6000	0.8563	0.0136	7.0 (6–8)	0.65	30	334

Convergence. In the comparison runs, half of the VLPSO searches had stopped by iteration 23, 23, and 20 on the three tasks and half of the BPSO searches by iteration 36, 35, and 35; no search of either kind reached the 100-iteration limit (Figure 3). The global-best fitness plateaus early in both, so the lower discrimination of the swarm-selected subsets is not a consequence of an exhausted budget. VLPSO also reached a higher final fitness than BPSO on every task (median 0.551, 0.495, and 0.427 against 0.519, 0.465, and 0.394). Though it is the better optimiser of Equation (6), its subsets discriminate no better on the outer folds. What limits discrimination is the objective, consisting of a nearest-neighbour wrapper on a subsample traded against a cardinality penalty, rather than the search.

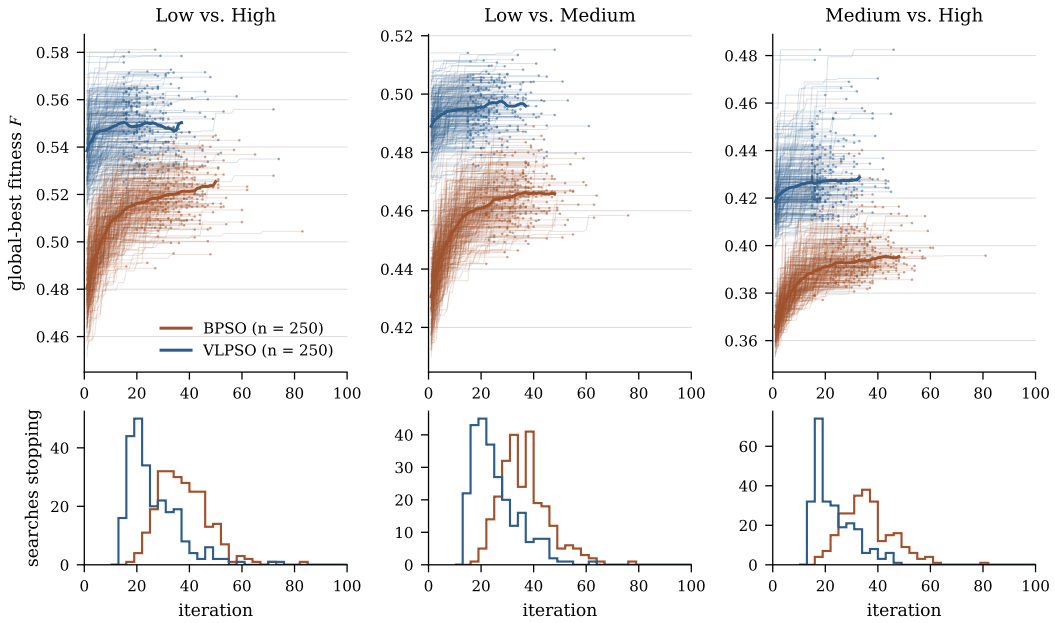


Figure 3. Convergence of the swarm searches in the selector comparison (250 searches per method and task, with 10 plausible values × 25 outer folds). Top: each thin line is the global-best fitness F (Equation (6)) of one search, drawn up to the iteration at which that search stopped (dot); the thick line is the median over the searches still running, shown while at least 10% of them run, so its late fluctuations reflect which searches remain rather than further improvement. Bottom: the iteration at which the searches stopped (stopping rule: no improvement for 15 iterations; limit 100).

5.6. Replicated Explanations

Reproduction. The 150 explained models refit the headline inner loop on repeat 0. All 150 selected the same model family as the headline run (gradient boosting in 147, random forest in three), 131 reproduced its outer-fold AUC exactly, and the largest difference was 4.1×10^{-4}, consistent with the change of NumPy version between the two runs. Therefore, the models being explained are the models behind Table 7.

Global SHAP. Table 12 and Figure 4 summarise the attributions over the 50 models per task. The ranking is highly reproducible across plausible values and folds: Kendall's W over the 61 design columns is 0.779, 0.809, and 0.604 for the three tasks, and 0.910, 0.906, and 0.811 over the 31 items alone (missingness indicators rank no higher than 18th in any task). The same two items rank first and second on every task: ST166Q03HA (phishing: click the link), ranked first in every task with an interquartile range of ranks of at most 1–2, and the number of books at home (ST013Q01TA). Gender ranks 3, 4, and 3. The televisions item (ST012Q01TA), which the variable name invites reading as books, ranks 6, 7, and 5. No direction is read from the signed mean attribution. Direction is taken from the single-variable screen: for ST166Q03HA, the signed AUC on Low versus High is below 0.5, consistent with its reverse scoring (endorsing the unsafe response is associated with the lower band).

In 141 of the 150 models, the attributions plus the base value reproduced the model's log-odds output to within 10^{-4} for every explained student. In six gradient-boosting models, the largest discrepancy over 2000 students reached 1.513, while in three random forest models it reached 0.0025; the number of students affected was not recorded. Excluding these models leaves the ten highest-ranked columns of every task identical and the full rankings correlated at $\rho \geq 0.997$, so the global summaries do not depend on them.

Table 12. Global SHAP for Low vs. High over 50 fitted models (10 plausible values × 5 outer folds, repeat 0): mean absolute attribution on the log-odds margin (2000 stratified test students per model), its standard deviation between plausible values, the median rank with interquartile range across models, and the percentage of models in which the column ranks in the top five. Kendall's *W* over the 50 rankings is 0.779.

Rank	Item	Content	Mean |SHAP|	SD (PV)	Rank, Median (IQR)	Top 5 (%)
1	ST166Q03HA	phishing: click the link	0.564	0.032	1 (1–2)	100
2	ST013Q01TA	books at home	0.548	0.033	2 (1–2)	100
3	ST004D01T	gender	0.393	0.041	3 (3–3)	100
4	ST166Q02HA	phishing: check sender	0.357	0.031	4 (4–4)	100
5	ST166Q05HA	phishing: check operator site	0.289	0.030	5 (5–5)	94
6	ST012Q01TA	televisions	0.200	0.048	6 (6–7)	6
7	ST012Q08NA	e-book readers	0.167	0.034	8 (7–8)	0
8	ST012Q06NA	computers	0.157	0.038	8 (7–9)	0
9	ST012Q03TA	rooms with bath/shower	0.144	0.027	8 (7–10)	0
10	ST166Q04HA	phishing: delete e-mail	0.129	0.020	9 (9–11)	0

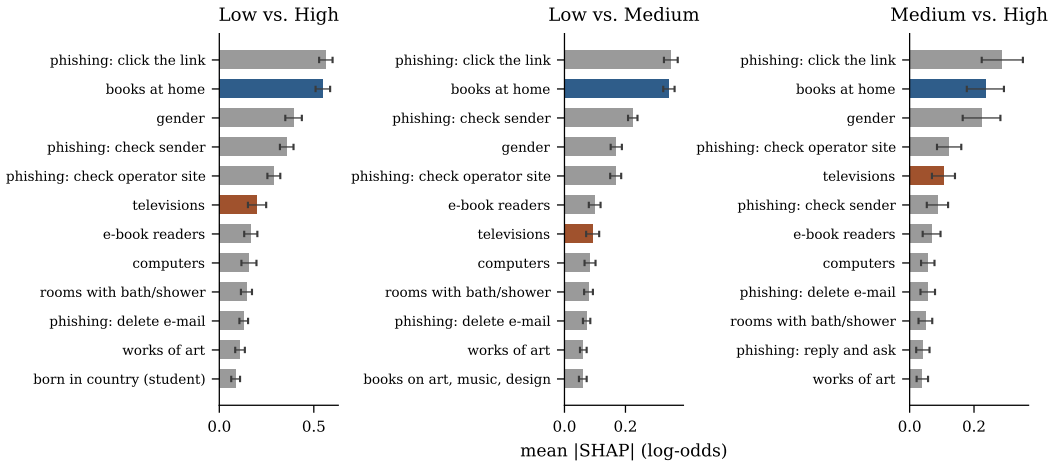


Figure 4. Global SHAP: the twelve columns with the largest mean absolute attribution per task, averaged over the 50 models of each task (10 plausible values × 5 outer folds); bars, mean; whiskers, ±1 standard deviation across models. Log-odds scale. Books at home (ST013Q01TA) in blue; the televisions item (ST012Q01TA), which the code invites reading as books, is in brown.

Local SHAP and LIME. LIME was run 30 times with different seeds on each of 90 students (six quantiles of predicted probability, five folds, three tasks, plausible value 1). Each run names 15 features, but the 30 runs on one student named between 25 and 39 distinct features between them. On average 6.7 features per student were named by all 30 seeds. For these, the weight was stable when named (median coefficient of variation 0.065) and its sign never changed (mean sign consistency 1.000). For features named by fewer than half of the seeds, the median coefficient of variation over the seeds that named them (at least two) was 0.263 and the mean sign consistency 0.871. Over the features named by both runs of a pair, the mean Spearman correlation between runs was 0.840 (range over students 0.619–0.964); counting unnamed features as zero lowers it to 0.771. Statistics that count the seeds which do not name a feature as zero weights are not used, as they make the standard deviation exceed the mean for any feature named by fewer than half of the seeds regardless of LIME's behaviour. A LIME explanation of these models is reproducible in its leading terms; beyond them, which features appear varies from seed to seed, but their weights (when they appear) are moderately stable (Figure 5).

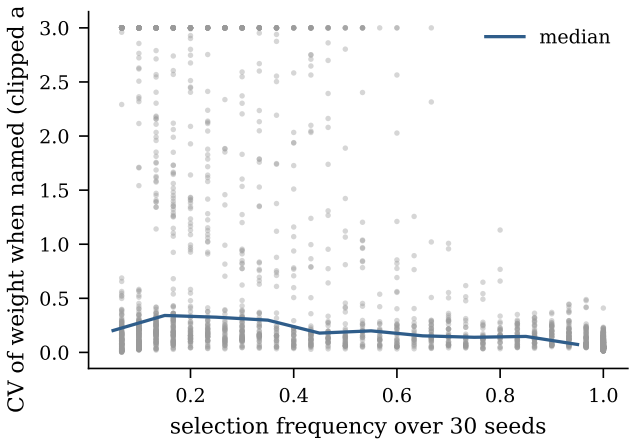


Figure 5. LIME stability over 30 seeds and 90 students: the coefficient of variation of a feature's weight over the seeds that name it, against the share of seeds that name it. Points, feature–student pairs named by at least two seeds; line, median by decile of selection frequency.

5.7. External Validation

Applied unchanged to the Portuguese sample (Table 13), the Spanish models discriminated at least as well as in the Spanish cross-validation: 0.9009 (95% CI 0.883–0.919) for Low versus High, 0.7678 (0.748–0.787) for Low versus Medium, and 0.7139 (0.679–0.749) for Medium versus High, for differences of +0.0159, +0.0158, and +0.0176 from the Spanish estimates. The two figures are not strictly comparable; the external models were fitted on the complete Spanish sample rather than on four-fifths of it, and the Portuguese tasks have their own composition and prevalence. What the comparison does show is that the discrimination obtained in Spain is not an artefact of the Spanish folds and carries over to a neighbouring education system assessed with the same instrument, with the ordering of the three tasks being preserved. Survey weighting changed the Portuguese estimates by at most 0.0024.

Table 13. External validation: Models selected and fitted on the complete Spanish task sample, applied unchanged to the Portuguese sample of PISA 2018. AUC with 95% interval from a bootstrap over Portuguese schools, combined over plausible values; weighted: survey-weighted AUC. Spain: the cross-validated estimate of Table 7. *n* and schools: median over plausible values.

Task	*n*	Schools	AUC	95% CI	Weighted	Spain (CV)	Difference
Low vs. High	2493	270	0.9009	0.8831–0.9188	0.9004	0.8850	+0.0159
Low vs. Medium	4079	272	0.7678	0.7484–0.7872	0.7696	0.7520	+0.0158
Medium vs. High	2769	247	0.7139	0.6789–0.7489	0.7163	0.6963	+0.0176

6. Discussion

6.1. What the Protocol Changes

The evaluation protocol is the part of this study that transfers independently of the selector it was built to assess. Three of its findings bear on any classification study of large-scale assessment data. First, a design matrix assembled by storage type contains the outcome; on these data, a one-split tree reproduces every label exactly (Section 5.1), so a published candidate predictor list is what allows a reader rather than only the authors to rule this out. Second, roughly half the variance of every estimate (FMI 47.0–60.0%) comes from measurement error in proficiency, and 69.69% of students are assigned to more than one band across plausible values; dichotomising an averaged score produces labels that are less certain than they appear and intervals that are too narrow. Third, the choice of estimand is consequential: pooling out-of-fold predictions from folds that each selected

their own model changed one estimate by up to 0.0730, an error that is invisible unless the fold-weighted quantity is computed alongside it.

The two permutation nulls answer different questions, and both are needed. The unrestricted null bounds the optimism contributed by fitting, tuning, and model selection at about 0.0003 in AUC. The within-school null shows that about a third of the discrimination above chance (31.1%) is recoverable from school membership alone, which is the expected consequence of between-school differences in achievement in a stratified clustered sample, and provides the reason that ungrouped folds would overstate what an individual's responses predict.

The two kinds of interval are also complementary. The school bootstrap of Table 7 quantifies the uncertainty in evaluating the fitted rules for the sampled schools; the Fay BRR standard error quantifies the uncertainty of the fitted rules' discrimination in the population represented by the survey weights. On these data, the weighted estimates differ little from the unweighted ones and the BRR standard errors (0.0069, 0.0059 and 0.0115) are somewhat larger than the bootstrap standard errors of the unweighted estimates, as the design effect of 2.134 leads one to expect; neither changes a conclusion.

6.2. Why VLPSO Underperforms

With the full design and matched settings, VLPSO did not improve on the full allowlist, BPSO, or on any filter, wrapper, or embedded selector; it was among the least accurate arms on every task while selecting the fewest items. The analyses of Sections 5.4 and 5.5 locate the cause in the objective rather than in the search. The searches converge well within the iteration budget, and doubling it changes nothing; VLPSO attains a higher value of its own objective than BPSO does; of the four ablations, only removing the cardinality penalty moves discrimination materially, recovering most of the loss. The objective trades a nearest-neighbour balanced accuracy, estimated on a subsample of the training data against subset size. On a candidate set of 31 items where the signal is spread across many weak correlated predictors (Table 6), this trade favours subsets of four to six items, which a regularised linear model on the full set outperforms. Methods that rank items by their contribution to the downstream learner itself (RFE, the L_1 path, SFS) do not face this mismatch, and match the full set at subset sizes set mostly by the 30-column cap (Section 5.4).

The variable-length mechanism had no measurable effect on discrimination: shrink-only adaptation and random initialisation were indistinguishable from the full method, and VLPSO and BPSO did not differ significantly on any task. What the ranked-prefix encoding changes is stability: VLPSO selected more reproducible subsets than BPSO across folds and across seeds, as all particles search the same highest-ranked columns. This stability remains below that of most filters and embedded methods and is bought with a smaller less discriminating subset. The value of VLPSO on data of this kind is limited to compression, (a subset of four to six of 31 items at a loss of 0.0257–0.0368 in AUC), and the size of that loss is set by the cardinality weight, which the sensitivity analysis shows to be the single setting that matters.

6.3. Explanations

Replicating the explanations over 50 models per task avoids resting the interpretation on a single fitted model. The global rankings were highly concordant across plausible values and folds (Kendall's W over items 0.811–0.910) and the same two items led on every task: the reverse-scored phishing item, and the number of books at home, with gender third or fourth. Measurement error in proficiency affects which students are labelled low or high far more than it affects which items the fitted models rely on.

A SHAP or LIME attribution describes how a fitted model uses a variable within a particular dataset. It is an associational statement about the model, not evidence that the variable causes achievement or evidence about what would follow from intervening on it. A SHAP plot orders features by attribution magnitude and has no time axis, so it cannot be read as a trajectory, while an attribution does not license a mechanism. The association between household resources and measured achievement is long-established [52–54]. That the number of books at home is among the strongest predictors here is unsurprising, and on its own is uninformative about why. The prominence of the phishing items is less familiar. They record judgments about how to respond to a suspicious e-mail, and their association with mathematics proficiency may reflect general reading and reasoning ability; however, the present design cannot distinguish this from other explanations.

Repeated-seed LIME shows that a single LIME explanation of these models is reliable in its leading terms but not beyond them; reporting the selection frequency and the conditional dispersion of each weight, as here, separates the two.

6.4. External Validity

The Spanish models transferred to Portugal without loss of discrimination and with the ordering of the three tasks preserved. Portugal shares the PISA instrument and cycle, so this is evidence of transfer between comparable systems rather than of generalisability to education systems in general, and the higher Portuguese AUCs may partly reflect that the external models were fitted on the complete Spanish sample.

6.5. Implications for Reporting Practice

Four elements of the protocol are inexpensive and could be adopted generally: publishing the candidate predictor list; grouping cross-validation folds by school; deriving outcome categories per plausible value and combining by Rubin's rules; and screening every admitted predictor on its own against a stated ceiling, which takes seconds and detects an outcome among the predictors at the point where it is introduced. For feature selection specifically, the results suggest that a selector should be compared with the unselected candidate set and with an embedded method under the same downstream learner before any benefit is claimed, and that its stability should be reported with an index that is defined for subsets of unequal size.

7. Limitations

Two countries, one cycle. The analytic sample is Spain in PISA 2018 and the external sample Portugal in the same cycle. Neither the discrimination estimates nor the selector comparison is shown to hold for other education systems, cycles, or assessments. Pooling student-level data across countries is constrained by the data licence; privacy-preserving federated designs, in which only model updates leave each data holder [55], are one route to such a comparison.

A small curated candidate set. The allowlist admits 31 items by design, so that every predictor is documented and none is outcome-derived. VLPSO was proposed for problems with thousands of features [18]; its behaviour on a larger candidate set from the same survey, where compression may matter more, was not examined.

One downstream learner in the selector comparison. Every arm was scored with L_2 logistic regression, so differences reflect the selector rather than the model. The ranking of selectors could differ under a tree ensemble.

Bounded tuning grids. The tuned subset size of the ranking methods was limited to 30 of 61 columns, and that of SFS to 30 columns by its stopping rule; these limits were reached in 92% and 82% of folds, respectively. Subsets of 31 to 60 columns were not evaluated for

these arms, and their comparison with the uncapped swarm searches is not a comparison at equal subset size. The VLPSO settings were fixed in advance; the sensitivity analysis varied one setting at a time on one task and plausible value, and interactions between settings were not explored.

Explanations. Explanations were computed for repeat 0 only (50 models per task) and LIME at plausible value 1 only. In 9 of 150 models, the SHAP attributions did not sum exactly to the model output for some students; the global summaries were unchanged when these models were excluded.

Uncertainty of the fitted rules only. The school bootstrap intervals and Fay BRR standard errors hold the fitted models fixed; refitting the full nested procedure, including feature selection, within each bootstrap resample or under each of the 80 replicate weights was not attempted. Both quantify the uncertainty of evaluating the fitted prediction rules and understate the uncertainty of the model development process as a whole, which they reflect only through the between-plausible-value component; in the selector comparison, it is approximated by the Nadeau–Bengio correction.

Cross-sectional design. Nothing reported here supports a claim about what would follow from changing any student's circumstances.

8. Conclusions

This study has specified an evaluation protocol for classification on large-scale assessment data and used it to evaluate variable-length particle swarm optimisation for selecting questionnaire predictors of mathematics proficiency in the Spanish sample of PISA 2018.

Under school-grouped nested cross-validation, with outcome categories derived per plausible value and combined by Rubin's rules, the AUC was 0.8850 (95% CI 0.874–0.896) for Low versus High, 0.7520 (0.743–0.761) for Low versus Medium, and 0.6963 (0.679–0.714) for Medium versus High (RQ1). An unrestricted permutation null averaged 0.4989 over 100 draws, and measurement error in proficiency contributed 47.0–60.0% of the variance of these estimates.

In a comparison of 16 arms over three tasks, ten plausible values, and 25 outer folds each, no feature selector improved on the full allowlist. RFE, SFS, and the L_1 path matched it with a maximum subset size of 30 of 61 design columns, which was reached in most folds. VLPSO selected four to six items and lost 0.0257–0.0368 in AUC; it did not differ significantly from fixed-length binary PSO under matched settings, its length-adaptation mechanism and ranking had no measurable effect, and its cost was set by the cardinality penalty. VLPSO was more stable than binary PSO but less stable than most filter and embedded methods (RQ2). On these data, VLPSO is a means of compressing the predictor set at a quantified cost, not of improving prediction.

Global SHAP rankings were highly concordant across 50 models per task, and LIME explanations were reproducible in their leading terms only (RQ3). The Spanish models transferred to Portugal without loss of discrimination (RQ4).

The protocol, the code and the aggregate results are public. Applying the protocol to other countries, cycles, and larger candidate sets and comparing selectors under a tree-ensemble learner are the natural next steps.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/asi9110228/s1, Supplementary Tables S1–S6 (one workbook): Table S1: Predictor allowlist; Table S2: Item-level descriptives; Table S3: Paired contrasts of the selector comparison; Table S4: Item selection frequencies by selector and task; Table S5: Selector comparison summary; Table S6: External validation by plausible value.

Author Contributions: All authors contributed to the conceptualisation, methodology, validation, and review of the manuscript. O.A.H. conducted the formal analysis, software implementation, data curation, visualisation, and original draft preparation. D.S., F.Q. and A.A.B. supervised and reviewed the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Grant Scheme (FRGS) of the Ministry of Higher Education, Malaysia, grant number FRGS/1/2022/ICT01/UKM/02/2, and by Universiti Kebangsaan Malaysia Internal Funding.

Data Availability Statement: The PISA 2018 data analysed in this study are publicly available from the OECD PISA 2018 database (https://www.oecd.org/en/data/datasets/pisa-2018-database.html, accessed on 1 October 2026); the OECD licence permits download but not redistribution, so the microdata are not redistributed. The analysis code, configuration files (including the predictor allowlist, denylist and codebook, random seeds and hyperparameter grids), the cell runner and the analysis outputs described below are available at https://github.com/Qussai2026/article-1 (accessed on 1 October 2026; code on branch `main`, commit `23f1734`; per-cell outputs of the selector comparison, stability and sensitivity, permutation, explanation, external-validation and replicate-weight analyses on branches `r2-results-a` to `r2-results-d`, from which `scripts/cells.py merge` and `aggregate` rebuild the corresponding tables; the headline nested cross-validation and the leakage audit are produced by `scripts/run_all.py --config default`; reproduction instructions in the README) and are archived at Zenodo (DOI: https://doi.org/10.5281/zenodo.23181755). Student-level derived data (individual predictions) are not published, in line with the OECD licence; they are regenerated by the code from the public data.

Acknowledgments: The authors acknowledge the Ministry of Higher Education, Malaysia, and Universiti Kebangsaan Malaysia for support through the Fundamental Research Grant Scheme and internal funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the receiver operating characteristic curve
BPSO	Binary particle swarm optimisation
BRR	Balanced repeated replication
CLPSO	Comprehensive learning particle swarm optimisation
CV	Coefficient of variation (of the survey weights); cross-validation
EDA	Exploratory data analysis
EDM	Educational data mining
FMI	Fraction of missing information
LIME	Local Interpretable Model-agnostic Explanations
MI	Mutual information
PISA	Programme for International Student Assessment
PSO	Particle swarm optimisation
PV	Plausible value
RFE	Recursive feature elimination
SFS	Sequential forward selection
SHAP	SHapley Additive exPlanations
SU	Symmetric uncertainty
VLPSO	Variable-length particle swarm optimisation
XAI	Explainable artificial intelligence

Glossary

Allowlist	The list of the 31 questionnaire items admitted as candidate predictors; any other column in a design matrix raises an error.
Design matrix	The table of predictor values (rows: students; columns: transformed items) given to a selector or model.
Fay BRR	Variance estimation that recomputes a statistic under each of 80 replicate weights, each perturbing the final weights of half of the sampled schools by a factor $1 \pm k$ with $k = 0.5$.
Kish design effect	$1 + \mathrm{CV}^2$ of the final weights: the variance inflation due to unequal weighting alone.
Plausible values	Ten random draws from each student's posterior distribution of proficiency, reported instead of a single score.
Rubin's rules	The combination of an analysis repeated over plausible values: mean of the estimates; variance equal to the mean within-PV variance plus $(1 + 1/M)$ times the between-PV variance.
School-grouped fold	A cross-validation fold in which all students of a school fall on the same side of the split.

References

1. Romero, C.; Ventura, S. Educational data mining: A review of the state of the art. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2010**, *40*, 601–618. [CrossRef]
2. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
3. Fischer, C.; Pardos, Z.A.; Baker, R.S.; Williams, J.J.; Smyth, P.; Yu, R.; Slater, S.; Baker, R.; Warschauer, M. Mining big data in education: Affordances and challenges. *Rev. Res. Educ.* **2020**, *44*, 130–160. [CrossRef]
4. Othman, N.; Mohd Matore, M.E.E.; Yunus, F.W. Artificial intelligence in educational measurement: A bibliometric review (1997 to 2024). *J. Appl. Sci. Eng. Technol. Educ.* **2025**, *7*, 314–325. [CrossRef]
5. Ying, L.; Bahar, N.; Singh, D.; Asri, F.H.M.; Al-mzary, M.M.; Falaki, N.; Alzoubi, M.R.; Alshurideh, M.T.; Al-Hawary, S.I.S. Investigating the ethical impact of AI in vocational education. In *Intelligence-Driven Circular Economy: Regeneration Towards Sustainability and Social Responsibility, Volume 2*; Hannoon, A., Mahmood, A., Eds.; Studies in Computational Intelligence; Springer: Cham, Switzerland, 2025; Volume 1174, pp. 97–114. [CrossRef]
6. OECD. *PISA 2018 Results (Volume I): What Students Know and Can Do*; OECD Publishing: Paris, France, 2019. [CrossRef]
7. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182. [CrossRef]
8. Xue, B.; Zhang, M.; Browne, W.N.; Yao, X. A survey on evolutionary computation approaches to feature selection. *IEEE Trans. Evol. Comput.* **2016**, *20*, 606–626. [CrossRef]
9. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
10. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144. [CrossRef]
11. Huang, Y.; Zhou, Y.; Chen, J.; Wu, D. Applying machine learning and SHAP method to identify key influences on middle-school students' mathematics literacy performance. *J. Intell.* **2024**, *12*, 93. [CrossRef]
12. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15. [CrossRef]
13. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef]
14. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
15. OECD. *PISA Data Analysis Manual: SPSS, Second Edition*; OECD Publishing: Paris, France, 2009. [CrossRef]
16. von Davier, M.; Gonzalez, E.; Mislevy, R.J. What are plausible values and why are they useful? *IERI Monogr. Ser.* **2009**, *2*, 9–36.
17. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, NY, USA, 1987. [CrossRef]
18. Tran, B.; Xue, B.; Zhang, M. Variable-length particle swarm optimization for feature selection on high-dimensional classification. *IEEE Trans. Evol. Comput.* **2019**, *23*, 473–487. [CrossRef]
19. Kennedy, J.; Eberhart, R.C. A discrete binary version of the particle swarm algorithm. In Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics: Computational Cybernetics and Simulation, Orlando, FL, USA, 12–15 October 1997; Volume 5, pp. 4104–4108. [CrossRef]

20. Badal, Y.T.; Sungkur, R.K. Predictive modelling and analytics of students' grades using machine learning algorithms. *Educ. Inf. Technol.* **2023**, *28*, 3027–3057. [CrossRef]
21. Chen, Y.; Zhai, L. A comparative study on student performance prediction using machine learning. *Educ. Inf. Technol.* **2023**, *28*, 12039–12057. [CrossRef]
22. Zainudin, S.; Ibrahim, R.; Sarim, H.M. Combining cluster quality index and supervised learning to predict students' academic performance. *Asia-Pac. J. Inf. Technol. Multimed.* **2024**, *13*, 72–92. [CrossRef]
23. Alvarez-Garcia, M.; Arenas-Parra, M.; Ibar-Alonso, R. Uncovering student profiles. An explainable cluster analysis approach to PISA 2022. *Comput. Educ.* **2024**, *223*, 105166. [CrossRef]
24. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [CrossRef]
25. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
26. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; pp. 856–863.
27. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
28. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
29. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
30. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
33. Ananda, R.; Rachmawaty, D.; Pratikno, B.; Hamid, O.A.; Mohd Amin, M.B. Improved clustering-based feature selection using feature extraction based on principal component analysis. In *Proceedings of the Innovations of Intelligent Informatics, Networking, and Cybersecurity: Second International Conference, 3INC 2024*; Communications in Computer and Information Science; Springer: Cham, Switzerland, 2025; Volume 2329, pp. 15–38. [CrossRef]
34. Kennedy, J.; Eberhart, R. Particle swarm optimization. In *Proceedings of the ICNN'95—International Conference on Neural Networks*; IEEE: Piscataway, NJ, USA, 1995; Volume 4, pp. 1942–1948. [CrossRef]
35. Liang, J.J.; Qin, A.K.; Suganthan, P.N.; Baskar, S. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. *IEEE Trans. Evol. Comput.* **2006**, *10*, 281–295. [CrossRef]
36. Nogueira, S.; Sechidis, K.; Brown, G. On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **2018**, *18*, 1–54. [CrossRef]
37. Kuncheva, L.I. A stability index for feature selection. In Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, Innsbruck, Austria, 12–14 February 2007; pp. 390–395.
38. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
39. Hoq, M.; Brusilovsky, P.; Akram, B. Analysis of an explainable student performance prediction model in an introductory programming course. In Proceedings of the 16th International Conference on Educational Data Mining, Bengaluru, India, 11–14 July 2023; pp. 79–90. [CrossRef]
40. Swamy, V.; Radmehr, B.; Krco, N.; Marras, M.; Käser, T. Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In Proceedings of the 15th International Conference on Educational Data Mining, Durham, UK, 24–27 July 2022; pp. 98–109. [CrossRef]
41. Alvarez-Melis, D.; Jaakkola, T.S. On the robustness of interpretability methods. *arXiv* **2018**, arXiv:1806.08049.
42. Kendall, M.G.; Babington Smith, B. The problem of m rankings. *Ann. Math. Stat.* **1939**, *10*, 275–287. [CrossRef]
43. OECD. *PISA 2018 Technical Report*; OECD Publishing: Paris, France, 2020.
44. Judkins, D.R. Fay's method for variance estimation. *J. Off. Stat.* **1990**, *6*, 223–239.
45. Kish, L. *Survey Sampling*; John Wiley & Sons: New York, NY, USA, 1965.
46. OECD. PISA 2018 Database. 2019. Available online: https://www.oecd.org/en/data/datasets/pisa-2018-database.html (accessed on 1 October 2026).
47. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
48. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
49. Field, C.A.; Welsh, A.H. Bootstrapping clustered data. *J. R. Stat. Soc. Ser. B* **2007**, *69*, 369–390. [CrossRef]
50. Nadeau, C.; Bengio, Y. Inference for the generalization error. *Mach. Learn.* **2003**, *52*, 239–281. [CrossRef]
51. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

52. Sirin, S.R. Socioeconomic status and academic achievement: A meta-analytic review of research. *Rev. Educ. Res.* **2005**, *75*, 417–453. [CrossRef]
53. OECD. *Equity in Education: Breaking Down Barriers to Social Mobility*; OECD Publishing: Paris, France, 2018. [CrossRef]
54. Park, H. Home literacy environments and children's reading performance: A comparative study of 25 countries. *Educ. Res. Eval.* **2008**, *14*, 489–505. [CrossRef]
55. Noori Alshami, A.S.; Ali, S.M.; Sadiq, A.T.; Muhsen, D.K.; Fadhil, S.A.; Hamid, O.A. Privacy-preserving federated clustering for financial data analytics in cloud environments. In Proceedings of the 2025 International Conference on Electrical Engineering and Informatics (ICEEI), Kuching, Malaysia, 13–15 November 2025; pp. 1–6. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.