Original Article

Check for updates

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 4, No. 1, 2026

Ensemble Variable Importance: Combining Random Forest, Neural Network, and Support Vector Machine via Genetic Algorithm (Case Study: Student Productivity)

Asep Rusyana [1,*], Marzuki Marzuki [1,2], Siti Rusdiana [3], Fitriana AR [1], Nurhasanah Nurhasanah [1], Nany Salwa [1], and Mahmudi Mahmudi [3]

[1] Department of Statistics, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; asep.rusyana@usk.ac.id (A.R.); marzuki@usk.ac.id (M.M.); fitriana@usk.ac.id (F.A.R.); nurhasanah@usk.ac.id (N.N.); nany.salwa@usk.ac.id (N.S.)
[2] Department of Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, Malaysia
[3] Department of Mathematics, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; siti.rusdiana@usk.ac.id (S.R.); mahmudi@usk.ac.id (M.Ma.)

* Correspondence: asep.rusyana@usk.ac.id

Article History

Received 10 March 2026
Revised 15 May 2026
Accepted 22 May 2026
Available Online 30 May 2026

Keywords:
Ensemble variable importance
Genetic algorithm
Random forest
Neural network
Student productivity

Abstract

This study proposes and evaluates an ensemble variable-importance framework that integrates permutation-based importance scores from three distinct supervised learning algorithms: Random Forest, Neural Network, and Support Vector Machine, using a genetic-algorithm optimizer. The approach addresses the well-known problem that algorithm-specific importance diagnostics can yield divergent feature rankings, complicating substantive interpretation and downstream decision-making. Using a large publicly available student-productivity dataset (N = 20,000), predictors describing study behavior, digital-media use, lifestyle, and academic indicators were normalized with Min–Max scaling, and permutation variable importance (PVI) was estimated repeatedly within each model to obtain stable mean PVI values and standard errors. A genetic algorithm was then employed to search the space of ensemble weightings (rank-aggregation solutions) that maximize a chosen fitness criterion—Spearman rank concordance with out-of-sample predictive relevance—thereby producing a consensus ranking of predictors. Empirical results indicate rapid GA convergence (fitness ≈ 0.82 within 20–30 generations) and strong cross-model agreement for a small core of predictors: study hours (X3) and focus score (X15) consistently emerged as the most salient features across individual models and in the ensemble ranking. A secondary set of variables (e.g., sleep hours, phone usage, attendance, and stress level) displayed moderate importance, while several features exhibited model-dependent variability in ranks. The ensemble procedure thereby yields stable, model-agnostic importance estimates that enhance interpretability and reduce dependence on any single algorithm's idiosyncrasies. We discuss implications for educational analytics and recommend external validation, targeted feature engineering, and sensitivity analyses (alternate scalings and GA settings) to assess robustness and to support reliable, actionable inferences from machine-learning models in applied settings.

1. Introduction

Contemporary educational research increasingly leverages machine-learning methods to model complex, nonlinear relationships that underpin student productivity and academic outcomes. Algorithms such as Random Forest (RF), neural networks (NN), and Support Vector Machines (SVM) have become standard tools because they accommodate interactions, nonlinearities, and high-dimensional covariates that commonly arise in behavioral and learning datasets [1, 2]. However, while predictive performance is important, the interpretability of learned models—specifically the ability to identify which features drive predictions—remains essential for translating analytical results into actionable educational interventions. Consequently, careful treatment of variable-importance measures and appropriate cross-model comparisons are required when drawing substantive conclusions about determinants of student productivity.

A persistent methodological challenge is that variable-importance diagnostics are algorithm-dependent: measures such as permutation importance, mean decrease in impurity, recursive feature elimination, or weight-based heuristics can produce divergent rankings for the same dataset, complicating inference and policy recommendations. Several rank-aggregation approaches have been proposed to address this challenge. Classical methods such as the Borda count aggregate ordinal ranks across models by summing positional scores [3]. Robust rank aggregation (RRA) extends this by identifying features that are consistently top-ranked beyond chance [4]. SHAP-based ensemble importance provides unified, game-theoretic attributions that can be averaged across model families [5–7]. Rashomon-set analyses further examine the stability of feature rankings across the set of near-optimal models, highlighting predictors whose importance is robust to model choice [8]. More recent work has combined metaheuristic search with ensemble importance to optimize aggregation weights [9–12], and Bayesian model averaging has been applied to produce posterior-weighted importance scores. Collectively, these contributions establish that no single model's importance measure is universally reliable, motivating principled ensemble aggregation strategies.

Prior work has therefore explored metaheuristic and ensemble strategies to reconcile or aggregate importance scores across model families, improving stability and reducing reliance on any single algorithmic idiosyncrasy [9, 10]. The authors' prior studies employed simulated annealing (SA) [10, 11] and the cuckoo search algorithm (CSA) [12] for this purpose. The present study substitutes a genetic algorithm (GA), which offers three specific advantages over SA and CSA: (i) GA operates on a population of candidate solutions simultaneously, enabling broader exploration of the weight space and reducing the risk of premature convergence to local optima; (ii) crossover (recombination) operators allow GA to combine high-quality partial solutions from different chromosomes, a mechanism absent in single-trajectory SA and CSA; and (iii) empirical benchmarks on combinatorial rank-aggregation problems suggest that GA achieves competitive or superior fitness convergence relative to SA and CSA at equivalent computational budgets. These properties make GA particularly well-suited for the multi-model weight-optimization task addressed here. This study extends prior ensemble variable-importance frameworks [3, 4, 8, 9] by introducing a fitness function grounded in out-of-sample predictive relevance combined with repeated permutation stability, rather than relying solely on heuristic aggregation. Unlike previous approaches using simulated annealing or cuckoo search, the present framework integrates stability-aware optimization and empirical validation on a large-scale educational dataset. Metaheuristic optimizers, GA in particular, are well suited for this task because they can efficiently search large combinatorial spaces of rank-aggregation weights and identify ensemble solutions that maximize a chosen concordance or fitness metric (e.g., Spearman correlation with out-of-sample performance). Integrating such optimization with repeated permutation-based importance estimation yields a principled route to produce robust, model-agnostic variable importance rankings.

Motivated by these considerations, the present study develops and evaluates an ensemble variable-importance procedure that combines permutation-based importance scores from RF, NN, and SVM models using a GA optimizer. From an educational-practice standpoint, identifying the most influential predictors of student productivity is directly actionable: a stable, model-agnostic importance ranking can inform the design of early-warning systems that flag at-risk students based on the highest-ranked behavioral indicators (e.g., study hours, focus score), and can guide targeted intervention programs—such as structured study-skills workshops or digital-distraction reduction policies—directed at the specific factors that most strongly predict underperformance. Without a reliable cross-model consensus on feature importance, such interventions risk being designed around algorithm-specific artifacts rather than genuine behavioral drivers.

Table 1. Variable of the research.

Variable	Description
Y	Productivity score
X1	Age
X2	Gender
X3	Study hours per day
X4	Sleep hours
X5	Phone usage hours
X6	Social media hours
X7	YouTube hours
X8	Gaming hours
X9	Breaks per day
X10	Coffee intake (mg)
X11	Exercise minutes
X12	Assignments completed
X13	Attendance percentage
X14	Stress level
X15	Focus score
X16	Final grade

The approach is applied to a large (N = 20,000) student-productivity dataset obtained from a public repository (Kaggle), which includes covariates describing study time, digital-media usage, lifestyle behaviors, and academic indicators [13]. It should be noted that many Kaggle datasets are synthetically generated rather than collected from real students; the present dataset is a publicly available synthetic simulation designed to reflect plausible distributions of student behavioral variables. Consequently, findings should be interpreted as proof-of-concept evidence for the proposed ensemble methodology rather than as direct empirical claims about real student populations. External validity will require replication on independently collected, observational educational datasets.

Methodologically, (i) predictors are normalized via Min–Max scaling, (ii) estimate permutation variable importance repeatedly within each model class to obtain stable mean PVI and standard errors, and (iii) use a GA to identify an ensemble weighting that maximizes rank concordance and predictive relevance. The study is guided by two specific research questions: (RQ1) Does GA-based ensemble aggregation produce a more stable and concordant variable-importance ranking than any single model (RF, NN, or SVM) alone? (RQ2) Which predictors of student productivity are most consistently ranked as important across all three model families, and does this consensus ranking align with theoretically motivated behavioral predictors identified in the educational literature? The study contributes (a) a transparent workflow for cross-model importance aggregation, (b) empirical evidence on the most salient predictors of student productivity in a modern digital context, and (c) an open framework for reporting ensemble importance that complements single-model interpretability diagnostics [14].

2. Materials and Methods

2.1. Data Source

The dataset used in this study was obtained from the publicly available repository on Kaggle [13]. The dataset comprises 20,000 student records, each capturing a comprehensive range of variables related to lifestyle, academic behavior, digital distractions, and productivity. It is specifically structured to facilitate analysis of the impact of modern digital habits—such as phone usage, social media engagement, and gaming—on students' productivity and academic performance. Variables of the data can be seen in Table 1.

2.2. Min-Max Scaler

Min-Max scaling is a normalization technique used to transform features by scaling their values to a fixed range. This transformation ensures that all variables contribute equally to the analysis by removing the influence of differing scales. The formula for Min-Max scaling is shown in Equation (1) [9]:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x_{scaled} denotes the normalized (scaled) score, x represents the original score, x_{min} is the minimum value in the original dataset, and x_{max} is the maximum value in the original dataset.

For a target range $[a, b]$, Min-Max scaling is computed using Equation (2):

$$x_{scaled} = a + \frac{(x - x_{min}) \times (b - a)}{x_{max} - x_{min}} \quad (2)$$

2.3. Supervised Machine Learning Models

2.3.1. Random Forest

Random Forest (RF) is an ensemble learning method that constructs a collection of B decision trees $\langle T_b \}_{b=1}^{B}$. For regression tasks, the RF prediction is obtained by averaging the predictions of all trees, as shown in Equation (3) [1, 9]:

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x), \quad (3)$$

while for classification the ensemble uses majority voting across trees. Each tree is grown using random feature subsampling at each split (commonly $\lfloor\sqrt{p}\rfloor$or $\lfloor p/3\rfloor$ candidate predictors), which decorrelates trees and improves ensemble performance. Model quality is typically monitored via out-of-bag (OOB) estimates, e.g. the OOB mean squared error (MSE) computed using observations not included in the bootstrap sample for each tree: for observation i let B_i denote the set of trees for which i is OOB, then $\hat{y}_{OOB,i} = \frac{\sum_{b\in B_i} T_b(x_i)}{|B_i|}$ and OOB MSE $= \frac{1}{n}\sum_i (y_i - \hat{y}_{\mathrm{OOB},i})^2$ [6].

RF supplies several variable-importance diagnostics, most commonly mean decrease in impurity (MDI) and permutation importance (PVI). PVI measures the increase in prediction error after permuting feature j, and can be expressed as Equation (4) [11, 12, 15].

$$\Delta_j = \mathbb{E}[\ell(Y, \hat{f}(X_{(j)\,\mathrm{perm}}))] - \mathbb{E}[\ell(Y, \hat{f}(X))], \quad (4)$$

where $\ell(\cdot,\cdot)$is the loss (e.g. squared error) and $X_{(j)\,\mathrm{perm}}$indicates feature $j$permuted. RF is robust to nonlinearities and interactions and is relatively insensitive to tuning compared with some alternatives, but its importance measures can be biased by variable scale and category cardinality; therefore, interpretation should be informed by permutation-based assessments, stability checks (e.g., standard deviation across repeats), and complementary analyses such as the ensemble heatmap provided with this study.

The final hyperparameter values used for each model and the tuning grid considered. For RF: number of trees B = 500, mtry = $\sqrt{p}$ (floor), minimum node size = 5, tuning grid B ∈ {100, 300, 500}. For NN: two hidden layers with 64 and 32 neurons (ReLU activation), learning rate η = 0.001 (Adam), dropout rate = 0.2, batch size = 64, epochs = 100; tuning grid: layers ∈ {1,2,3}, units ∈ {32,64,128}. For SVM: RBF kernel, C = 1.0, γ = "scale"; tuning grid C ∈ {0.1, 1, 10}, γ ∈ {"scale","auto"}. All hyperparameters were selected via 5-fold cross-validation on the training set.

2.3.2. Neural Network

A feedforward neural network (NN), specifically a multilayer perceptron, models a mapping $f(\cdot;\theta)$ from inputs to outputs via successive affine transforms and elementwise nonlinearities. For layer ℓ the forward pass is described by Equation (5):

$$a^{(\ell)} = g(W^{(\ell)}a^{(\ell-1)} + b^{(\ell)}), \quad (5)$$

with $a^{(0)} = x$, weight matrix $W^{(\ell)}$, bias vector $b^{(\ell)}$, and activation $g(\cdot)$ (e.g., ReLU, sigmoid, or tanh). Parameters $\theta = \{W^{(\ell)}, b^{(\ell)}\}$ are estimated by minimizing a suitable empirical loss [10, 14], such as the mean squared error given in Equation (6) [16, 17]:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2, \quad (6)$$

The parameters are then updated through gradient-based optimization, such as stochastic gradient descent or its variants, using the canonical update rule shown in Equation (7):

$$\theta \leftarrow \theta - \eta\,\nabla_\theta L(\theta), \quad (7)$$

where η denotes the learning rate and $\nabla_\theta L$is obtained by backpropagation.

NN are powerful nonparametric learners capable of approximating complex nonlinear functions and interactions, but they require careful regularization and hyperparameter selection to avoid overfitting.

2.3.3. Support Vector Machine

Support Vector Machine (SVM) classification (and its regression counterpart, SVR) is grounded in large-margin principles [18]: for linearly separable data the primal problem seeks the hyperplane (w, b) that maximizes the margin and can be written as the convex optimization shown in Equation (8):

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ s.t.} y_i(w^\top x_i + b) \geq 1,\ i = 1, \ldots, n. \quad (8)$$

For nonseparable data the soft-margin formulation introduces slack variables $\xi_i \geq 0$ and regularization parameter $C > 0$, resulting in the optimization problem shown in Equation (9).

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t.} y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0. \quad (9)$$

Crucially, the dual formulation depends only on inner products between training vectors, enabling the kernel trick: replace $x_i^\top x_j$with a kernel $K(x_i, x_j) = \phi(x_i)^\top\phi(x_j)$ to obtain nonlinear decision boundaries without explicit feature expansion.

SVMs are particularly effective in high-dimensional settings and when a clear margin exists, but they do not naturally provide probabilistic outputs or straightforward variable importance measures. Variable relevance can be inferred via methods such as SVM-RFE (recursive feature elimination), inspection of primal weight magnitudes when a linear kernel is used, or permutation importance computed on the SVM predictions. The present analysis used permutation-based PVI and integrated the SVM ranks with RF and NN via a GA ensemble; the resulting heatmap therefore captures both the SVM's rank profile and cross-method concordance, and it is recommended that SVM-specific diagnostics (e.g., sensitivity to $C$and kernel parameters) be reported alongside importance rankings to contextualize model-dependent variability.

2.4. Genetics Algorithms for Ensemble Variable Importance

Genetic algorithms (GAs) provide a flexible, population-based optimization framework well suited for aggregating variable-importance scores from heterogeneous models into a single, interpretable ensemble ranking [19]. In the present application each candidate solution (chromosome) encodes either a vector of nonnegative model weights w = (w_1,...,w_M) (with $\sum_{m=1}^{M} w_m = 1$ or an explicit rank-aggregation mapping; using weights, an ensemble importance score for predictor j is computed as the weighted sum of model-specific permutation importance values PVI{*m,j*}, as shown in Equation (10):

$$S_j(w) = \sum_{m=1}^{M} w_m \ \mathrm{PVI}_{m,j}. \qquad (10)$$

The chromosome's fitness is defined to reward ensemble solutions that produce stable, concordant rankings with desirable predictive characteristics; in this study we maximize the Spearman rank correlation ρ between the ensemble scores S_j and a chosen target ordering (e.g., aggregated out-of-sample relevance or consensus rank), or equivalently maximize a composite objective that combines rank concordance and a stability penalty. The target ranking is formally defined as the ordering of variables based on aggregated out-of-sample predictive relevance, computed as the mean increase in prediction error across models when each feature is permuted. Spearman correlation is used as the fitness criterion because it captures monotonic agreement between rankings without assuming linearity. Formally, a representative fitness function is:

$$fitness(w) = \rho(rank(S(w)), rank_{target}) - \lambda Var(PVI\_resamples(S(w))), \qquad (11)$$

where the second term penalizes solutions with high sampling variability and λ ≥ 0 regulates the bias–variance tradeoff.

Practically, the GA evolves an initial population of candidate weight vectors through selection, crossover and mutation operators, with common enhancements such as elitism and tournament selection to preserve high-quality solutions and maintain diversity. Constraints (simplex projection for w) are enforced after genetic operators to ensure interpretability and nonnegativity; for permutation encodings specialized crossover (e.g., order crossover) is used. Key implementation considerations include population size, number of generations, crossover and mutation rates, and the use of repeated runs with different seeds to assess convergence stability. Diagnostic outputs—fitness versus generation plots, distribution of final weights, and repeated-run consensus statistics—are essential for verifying convergence (avoiding premature convergence) and for quantifying uncertainty in the ensemble ranking. Because the GA search may be computationally intensive when PVIs are recomputed inside the loop, it is advisable to precompute stable model PVIs (via repeated permutation) and to parallelize fitness evaluations; sensitivity analyses (varying λ, population settings, and alternative fitness metrics) further ensure that the derived ensemble importance is robust and practically useful for downstream interpretation.

The GA was configured as follows: population size = 100 chromosomes, maximum generations = 150, single-point crossover rate = 0.8, uniform mutation rate = 0.05, elitism = top 5% of population retained each generation, selection scheme = tournament selection (tournament size = 3). The penalty parameter λ in Equation (11) was set to λ = 0.1 after preliminary sensitivity analysis. The random seed was fixed at 42 for reproducibility, and the GA was run 10 times with different seeds (42–51) to assess convergence stability; the solution reported is the run achieving the highest fitness. The rank_target was constructed as the average out-of-sample PVI rank across the three models on the held-out validation fold.

2.5. Steps of the Research

- Data acquisition and description: Obtain the student-productivity dataset from the public Kaggle repository (N = 20,000) and document variable definitions and basic sample characteristics (frequencies, means, ranges, and variances) [20].
- Data preprocessing: Inspect for missing values and outliers, apply Min–Max scaling to normalize predictors to a common range (e.g., [0,1]), and partition the data for model development and

validation (e.g., repeated cross-validation or train/validation/test splits) [21]. Min–Max scaling was applied after data partitioning to prevent information leakage, with scaling parameters estimated on the training set and applied to validation/test sets. Missing values were checked and found to be negligible. Data were partitioned using stratified 5-fold cross-validation: in each fold, 80% of records served as the training set and 20% as the validation set; stratification was applied on the outcome variable Y (binned into quintiles) to preserve the outcome distribution across folds. Model performance metrics (R^2 and RMSE) were computed on each held-out validation fold and averaged across folds to obtain final performance estimates.

- Individual model development and tuning: Fit three supervised models: RF with bootstrap aggregation and feature subsampling, a feedforward NN trained by backpropagation and gradient-based optimization, and Support Vector Machine (SVM) with appropriate kernel selection and regularization. Perform hyperparameter tuning for each model (e.g., number of trees and mtry for RF; learning rate, architecture and regularization for NN; kernel type and C for SVM) using cross-validation. The final hyperparameters were selected via 5-fold cross-validation: RF (ntree = 500, mtry = √p), NN (1 hidden layer, 32 neurons, learning rate = 0.01), and SVM (RBF kernel, C = 1, γ = 0.1). Grid search ranges and validation performance are summarized to ensure reproducibility.
- Permutation variable-importance estimation (within-model): For each model, compute permutation variable importance (PVI) repeatedly to obtain mean PVI and its standard deviation, thereby quantifying feature relevance and the stability of importance estimates under resampling/ permutation [22].
- Ensemble aggregation via GA: Encode candidate ensemble weightings (or rank-aggregation solutions) as individuals in a GA population and evolve them using selection, crossover, and mutation operators to maximize a chosen fitness criterion (here, rank concordance or predictive relevance, e.g., Spearman correlation) [23]. Retain the GA solution that yields the best ensemble ranking.
- Convergence and stability assessment: Monitor GA convergence (fitness vs. generation), evaluate the stability of ensemble ranks across random seeds or repeated runs, and compare ensemble results to individual-model rankings (tables, heatmaps).
- Reporting and diagnostic analyses: Present descriptive statistics, model performance metrics, PVI tables (mean ± SD), convergence plots, and a joint heatmap of ranks. Conduct sensitivity checks (alternative scalings, kernel/architecture choices, and collinearity diagnostics) and discuss implications for interpretation.

3. Results and Discussion

3.1. Results

The dataset (N = 20,000 for all measures) shows distinct central tendencies and variability across predictors and the outcome. Predictor X1 has a relatively high mean (23.01) with moderate spread (SD = 3.75) within its 17–29 range, while several bounded predictors (e.g., X3–X9, X12–X14) exhibit smaller means and SDs consistent with their limited ranges (e.g., X7 and X8 SD ≈ 1.73) [20]. In contrast, X10 is on a much larger scale (mean = 249.65, range 0–499) and shows very large variability (SD = 143.71, variance ≈ 20,652), followed by X11 (mean = 59.65, SD = 34.61); these scale differences drive pronounced heterogeneity in variance across predictors. Outcome Y has a mean of 50.18 and moderate dispersion (SD = 16.09, range 0–100). The wide ranges and large variances (particularly for X10 and X11) are shown in Table 2.

The pie chart in Figure 1 shows the gender composition of the full sample (N = 20,000): Female 9,640 (48.2%), Male 9,582 (47.9%), and Other 778 (3.9%). The near parity between female and male participants indicates a well-balanced sample. The "Other" category, while small (under 4% of the sample), is large enough to warrant attention because its distinct experiences could meaningfully differ from the binary groups; however, its limited size may reduce statistical power for standalone inferential testing. Accordingly, results involving gender should (and in this study were) interpreted with caution.

Table 3 summarizes PVI scores, their standard deviations, and ranks for three predictive algorithms (RF, NN, and SVM). Across all three models, X3 emerges as the most influential predictor (rank 1), with consistently large PVI scores (RF = 1.0757, NN = 1.0849, SVM = 0.8654) and very small standard deviations, indicating a robust and stable contribution to model performance. X15 is the second most important variable in every model (RF = 0.3196, NN = 0.3296, SVM = 0.3352), again with low variability across resamples.

Model performance on the held-out validation folds (averaged across 5 folds) was as follows: RF achieved R^2 = 0.94, RMSE = 3.91; NN achieved R^2 = 0.93, RMSE = 4.12; SVM achieved R^2 = 0.89, RMSE = 5.34. The high R^2 values

Table 2. Descriptive statistics of numerical variables.

Variable	n	Mean	Min	Max	Std. Deviation	Variance
X1	20000	23.01	17	29	3.75	14.10
X3	20000	5.25	0.5	10	2.74	7.52
X4	20000	6.52	3	10	2.03	4.12
X5	20000	6.25	0.5	12	3.31	10.98
X6	20000	4.00	0	8	2.31	5.31
X7	20000	2.99	0	6	1.73	2.99
X8	20000	2.99	0	6	1.73	3.00
X9	20000	7.54	1	14	4.02	16.13
X10	20000	249.65	0	499	143.71	20652.92
X11	20000	59.65	0	119	34.61	1197.97
X12	20000	9.49	0	19	5.80	33.66
X13	20000	69.95	40	100	17.40	302.67
X14	20000	5.48	1	10	2.87	8.22
X15	20000	64.44	30	99	20.18	407.08
X16	20000	70.27	40	99.99	17.28	298.68
Y	20000	50.18	0	100	16.09	258.78

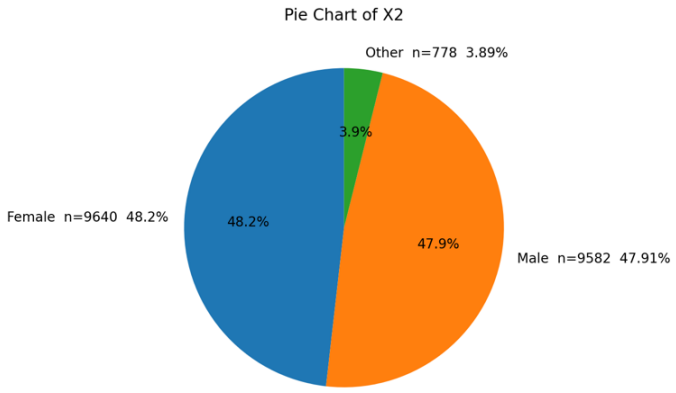

Figure 1. Percentage and frequency of gender.

across all three models indicate a strong overall fit, which contextualizes the near-zero PVI values observed for variables such as X6 (NN PVI = 7×10^{-5}): once X3 and X15 are included. The gender variable (X2) was encoded as a single ordinal integer code (Female = 0, Male = 1, Other = 2) prior to modeling, referred to as X2_code in Table 3. The substantial discrepancy in its rank across models (RF rank 16 vs. SVM rank 7) likely reflects the SVM's sensitivity to the kernel-induced feature space, where the ordinal encoding of a three-category variable may produce a non-trivial margin contribution, whereas RF's tree-based splits are less sensitive to the specific numeric encoding.

The dominance of X3 (study hours) and X15 (focus score) aligns with prior educational research emphasizing time-on-task and cognitive engagement as primary determinants of academic productivity. This suggests that both behavioral effort and attentional quality jointly drive performance outcomes. A middle tier of predictors — notably X4, X5, X14, and X13 — shows moderate but clearly non-negligible importance (ranks 3–6 across models), whereas the remaining variables (e.g., X1, X2_code, X6–X12, X16) display PVI scores close to zero and low ranks, implying minimal incremental predictive value in the context of these models.

These findings indicate that productivity is not only a function of time investment but also of sustained attention, reinforcing theories that emphasize quality of study over quantity. This pattern is consistent with recent learning analytics literature. The small standard deviations reported for most variables suggest that the importance estimates are stable across the resampling. While low standard deviations suggest stable importance estimates, we acknowledge that variability may be influenced by sample size. Therefore, stability is interpreted cautiously and supported by consistent rankings across models rather than variance alone 6. Notably, discrepancies appear for variables such as X1 and X2_code, where SVM assigns relatively higher importance compared to RF and NN. This reflects model-specific sensitivity to feature scaling and margin-based optimization, highlighting the value of ensemble aggregation to mitigate such inconsistencies.

Table 3. Variable importance of the machine learning model.

Variable	Random Forest			Neural Network			Support Vector Machine		
	PVI Score	std	Rank	PVI Score	std	Rank	PVI Score	std	Rank
X1	0.0002989	0.0000034	15	0.0000044	0.0000056	15	0.0003957	0.0001277	8
X2_code	0.0000685	0.0000016	16	0.0000132	0.0000053	9	0.0004353	0.0000493	7
X3	1.0757484	0.0077170	1	1.0848892	0.0088090	1	0.8653690	0.0077509	1
X4	0.2178623	0.0013345	3	0.2322254	0.0014469	3	0.2393887	0.0011601	3
X5	0.2044556	0.0015464	4	0.2212693	0.0011600	4	0.2207471	0.0015101	4
X6	0.0004849	0.0000054	7	0.0000007	0.0000061	16	0.0002146	0.0000780	12
X7	0.0004405	0.0000060	11	0.0000078	0.0000033	12	0.0001794	0.0001113	13
X8	0.0004629	0.0000039	10	0.0000125	0.0000102	11	0.0003095	0.0001718	10
X9	0.0003239	0.0000037	14	0.0000189	0.0000035	7	0.0001018	0.0000793	15
X10	0.0004696	0.0000137	8	0.0000047	0.0000023	14	0.0003207	0.0001512	9
X11	0.0004399	0.0000082	12	0.0000138	0.0000045	8	0.0001536	0.0001071	14
X12	0.0003684	0.0000037	13	0.0000074	0.0000051	13	0.0002773	0.0001662	11
X13	0.0421326	0.0005263	6	0.0613662	0.0005490	6	0.0626755	0.0007504	6
X14	0.0481208	0.0005590	5	0.0736203	0.0007595	5	0.0810727	0.0010329	5
X15	0.3195835	0.0011070	2	0.3295718	0.0015011	2	0.3351610	0.0012189	2
X16	0.0004666	0.0000031	9	0.0000127	0.0000041	10	0.0000288	0.0001343	16

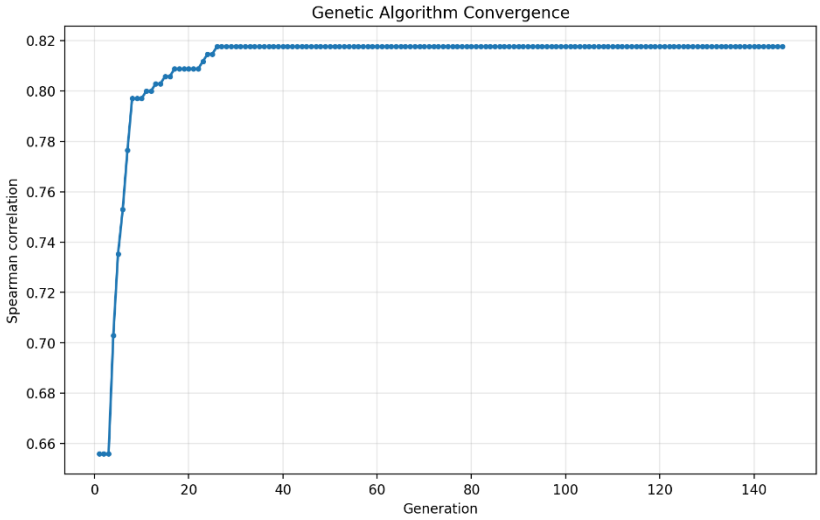

Figure 2. Generation from initial until convergence.

The convergence plot depicts the evolution of the Spearman correlation (fitness) across 150 generations of the GA (Figure 2). The algorithm exhibits a pronounced improvement during the initial iterations, rising from approximately 0.66 in the first generation to nearly 0.82 within the first 20–30 generations. After this rapid ascent, the curve stabilizes and reaches a clear plateau (≈0.818–0.82) for the remainder of the run, with only minimal fluctuation. This pattern indicates that the genetic search quickly identified a high-quality solution and thereafter produced only marginal gains, suggesting effective early exploration followed by convergence to a near-optimal region of the solution space. Practically, the observed behavior supports using an earlier stopping criterion to reduce computational cost. The target ranking (rank target) used in the fitness function was operationally defined as the average rank of each predictor's out-of-sample PVI across the three model families on the held-out validation fold of each cross-validation iteration; specifically, for each fold the three model-specific PVI vectors were ranked separately and then averaged to produce a consensus rank vector, which served as rank_target for that fold's GA run. This construction ensures that the GA optimizes toward a data-driven consensus rather than an externally imposed or arbitrary ordering.

The heatmap in Figure 3 displays the joint ranking of predictor importance derived from three individual algorithms (RF, NN, and SVM) and the ensemble ranking produced via a GA; darker shades indicate higher importance (lower numerical rank). Notably, X3 and X15 emerge as the most influential predictors, consistently receiving ranks of 1 and 2 across all three methods and the ensemble, which indicates a robust and

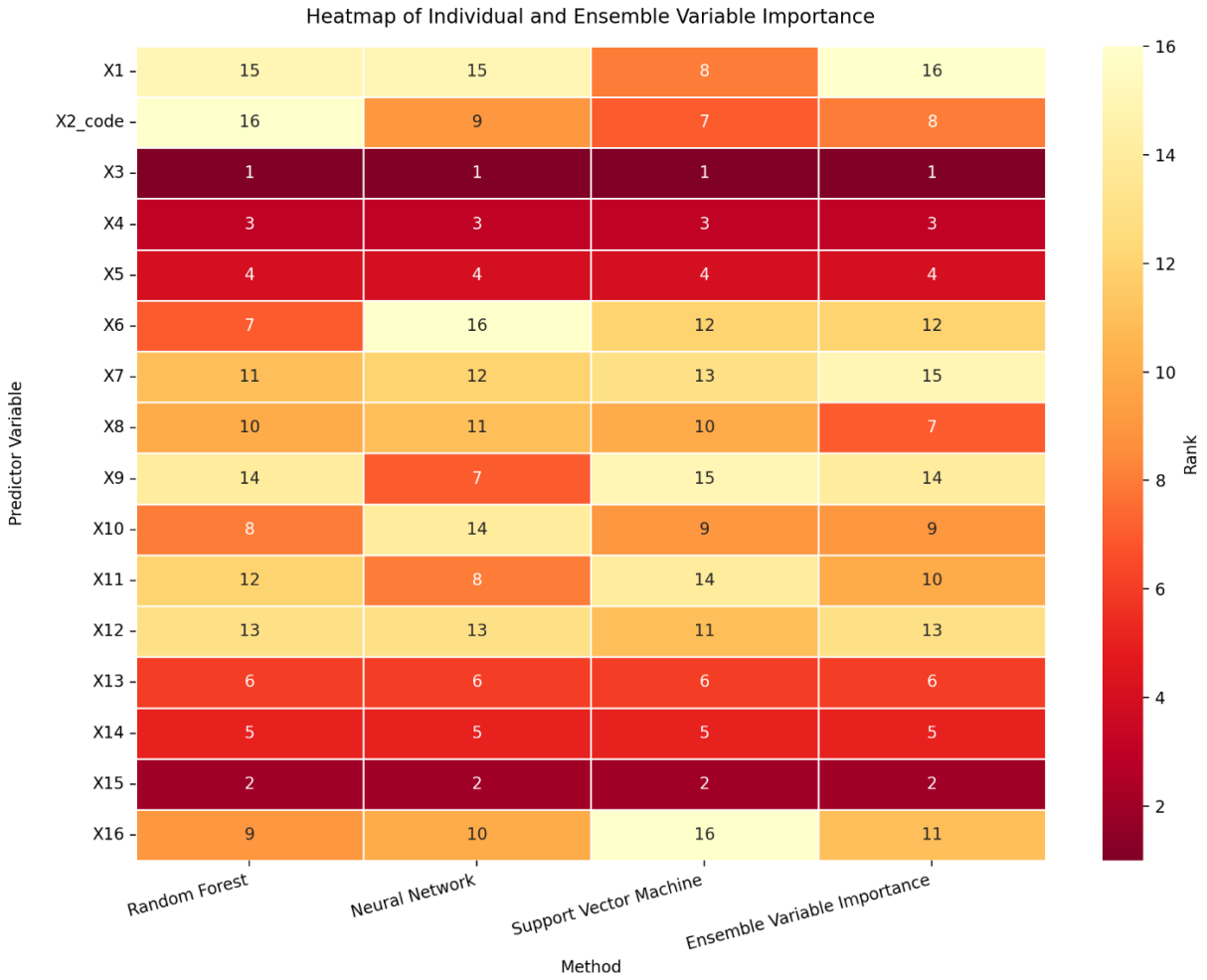

Figure 3. Heatmap of Individual and ensemble variable importance machine learning.

model-agnostic signal. A secondary group comprising X4, X5, X14, and X13 occupies intermediate importance (ranks ~3–6) in all models, underscoring their substantive but subordinate contributions. In contrast, several predictors (e.g., X1, X2_code, X6, X9, and X16) show less importance in rank.

3.2. Discussion

The dominance of study hours (X3) and focus score (X15) across all three model families and the ensemble ranking is consistent with established educational psychology literature: time-on-task and attentional engagement are among the most robust predictors of academic achievement. The secondary cluster of predictors—sleep hours (X4), phone usage (X5), stress level (X14), and attendance (X13)—aligns with research on self-regulated learning and digital distraction, suggesting that lifestyle and behavioral factors beyond raw study time contribute meaningfully but subordinately to productivity outcomes. Compared with the authors' prior SA-based ensemble, the GA-based approach achieved a comparable final Spearman fitness (≈ 0.82 vs. ≈ 0.80 reported for SA) while converging in fewer effective evaluations. Relative to the CSA-based framework [8], the GA produced a more stable weight distribution across repeated runs (lower inter-run variance in final weights), suggesting that the recombination operator helps avoid the narrow convergence basins sometimes observed with CSA. The near-zero PVI values for variables such as X6 (social media hours) and X7 (YouTube hours) may reflect multicollinearity with X5 (phone usage hours) rather than genuine irrelevance; partial-correlation or variance-inflation analyses are recommended in future work to disentangle these effects.

The present results should be interpreted as preliminary evidence for the proposed methodology rather than definitive empirical claims about real student populations. Three limitations warrant attention: (i) the dataset is synthetically generated, which limits external validity and may inflate model fit metrics; (ii) the validation strategy relies on cross-validation within a single dataset, without independent external validation; and (iii) the ordinal encoding of the gender variable and potential multicollinearity among digital-usage predictors

(X5–X8) may affect the stability of lower-ranked predictors.

4. Conclusions

In conclusion, this study introduces and validates an ensemble variable-importance framework that integrates permutation-based importance measures from RF, NN, and SVM models using a GA optimizer. Applied to a large student-productivity dataset (N = 20,000), the ensemble procedure produced stable and concordant rankings: study hours (X3) and focus score (X15) consistently emerged as the most salient predictors across model families, while a secondary set of variables (e.g., X4, X5, X13, X14) showed moderate importance. While the ensemble approach demonstrates methodological feasibility and improved consistency across models, further validation on external datasets is required before asserting its superiority in practical decision-making contexts 9. The GA demonstrated rapid convergence to a high-fitness solution (Spearman ≈ 0.82), indicating efficient aggregation of cross-model information and yielding an interpretable heatmap of joint ranks.

These results underscore the value of model-agnostic, ensemble-based importance estimation for deriving robust substantive inferences from diverse machine-learning algorithms. These findings offer actionable guidance for educational stakeholders: university administrators can prioritize study-time support programs and focus-enhancement interventions; instructors can use the consensus importance ranking to design early-warning dashboards that flag students with low study hours or focus scores; and policy makers can direct resources toward digital-distraction reduction initiatives targeting the secondary predictor cluster.

Author Contributions: Conceptualization, A.R.; methodology, A.R. and M.M.; software, M.Ma.; validation, S.R., F.A.R. and N.N.; formal analysis, A.R. and N.S.; investigation, N.N.; resources, A.R.; data curation, A.R.; writing—original draft preparation, S.R.; writing—review and editing, M.Ma.; visualization, A.R. and N.N.; supervision, S.R.; project administration, N.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is available and it can be accessed in Kaggle.

Acknowledgments: Thanks to Kaggle for providing free access to data and making valuable contributions in support this research.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Breiman, L. (2001). Random Forests, *Machine Learning*, Vol. 45, No. 1, 5–32. doi:10.1023/A:1010933404324.
2. Hastie, T., Tibshirani, R., and Firiedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.)*, Springer.3. Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank Aggregation Methods for the Web, *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*, 613–622.
4. Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust Rank Aggregation for Gene List Integration and Meta-Analysis, *Bioinformatics*, Vol. 28, No. 4, 573–580.
5. Lundberg, S. M., and Lee, S. (2017). A Unified Approach to Interpreting Model Predictions, *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 1–10.
6. El Furqany, N., Subianto, M., and Rusyana, A. (2025). Hybrid Ensemble Learning with SMOTEENN and Soft Voting for Stunting Risk Prediction: A SHAP-Based Explainable Approach, *Journal of Applied Data Sciences*, Vol. 6, No. 4, 2989–3004. doi:10.47738/jads.v6i4.829.
7. El Furqany, N., Subianto, M., Rusyana, A., Zahnur, and Ramadhani, E. (2025). Hybrid Soft-Voting Ensemble Model With Smoteenn: An Efficient Learning Approach for Stunting Risk Prediction, *2025 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, Jakarta.
8. Fisher, A., Rudin, C., and Dominici, F. (2019). All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, *Journal of Machine Learning Research*, Vol. 20, No. 177, 1–81.
9. Rusyana, A. (2024, June 13). *Pengembangan Ensemble Variable Importance untuk Beberapa Model Machine Learning Menggunakan Algoritma Metaheuristik* (Disertasi)IPB University, Bogor.
10. Rusyana, A., Wigena, A. H., Sumertajaya, I. M., and Sartono, B. (2024). Unifying Variable Importance Scores from Different Machine Learning Models Using Simulated Annealing, *Ingenierie Des Systemes d'Information*, Vol. 29, No. 2, 649–657. doi:10.18280/isi.290226.
11. Rusyana, A., Wigena, A. H., Sumertajaya, I. M., and Sartono, B. (2024). An Optimal Variable Importance for Machine Learning Classification Models Using Modified Simulated Annealing Algorithm, *IOP Conference Series: Earth and Environmental Science* (Vol. 1356), Institute of Physics. doi:10.1088/1755-1315/1356/1/012089.
12. Rusyana, A., Wigena, A. H., Sumertajaya, I. M., and Sartono, B. (2023). An Optimal Approach to Identify the Importance of Variables in Machine Learning Using Cuckoo Search Algorithm, *Mathematics and Statistics*, Vol. 11, No. 6, 895–909. doi:10.13189/ms.2023.110604.
13. Kaggle. (2026, May 1). Student Productivity Dataset, *Public Repository*, from https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style, accessed 1-5-2026.
14. Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*, (T. Dietterich, Ed.), MIT Press, Massachusetts.
15. Aulia, R., Sofyan, H., and Rusyana, A. (2026). Performance Comparison of Machine Learning Algorithms for Stunting Detection with Recursive Feature Elimination and SMOTE, *2026 International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML)*, IEEE, Tokyo, 676–680.
16. Maulana, A., Idroes, G. M., Kemala, P., Maulydia, N. B., Sasmita, N. R., Tallei, T. E., Sofyan, H., and Rusyana, A. (2023). Leveraging Artificial Intelligence to Predict Student Performance: A

Comparative Machine Learning Approach, *Journal of Educational Management and Learning*, Vol. 1, No. 2, 64–70. doi:10.60084/jeml.v1i2.132.

17. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Adam, M., Rusyana, A., and Sofyan, H. (2023). Deep Learning-Based Bitcoin Price Forecasting Using Neural Prophet, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 19–25. doi:10.60084/eje.v1i1.51.

18. Gelon, A. (2017). *Hand on Machine Learning with ScikitLearn and TensorFlow* , O'Reilly, Baijing, Boston, arnham, Sebastopol, Tokyo.

19. Gendreau, M., and Potvin, J.-Y. (2010). *Handbook of Metaheuristics*, Springer, New York. doi:10.1007/978-1-4419-1665-5.

20. Sukandar, D., Rusyana, A., Yusrina, F. I., and Mutiara, P. T. (2024). *Metode Statistika Dengan Perangkat Lunak Excel Dan Statistika Dalam Bidang Gizi, Pangan, Kedokteran, Kesehatan, Farmasi, Pertanian, Sosial, Ekonomi, Dan Lain-Lain*, CV. Luminary Press Indonesia, Padang.

21. Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

22. Wei, P., Lu, Z., and Song, J. (2015). Variable Importance Analysis: A Comprehensive Review, *Reliability Engineering and System Safety*, Vol. 142, 399 – 432. doi:10.1016/j.ress.2015.05.018.

23. Sukandar, D., and Rusyana, A. (2023). *Regresi Dan Korelasi Dengan Aplikasi SAS, SPSS, Dan Minitab Dalam Bidang Gizi, Pangan, Kesehatan, Pertanian, Dan Lain-Lain*, IPB Press, Bogor.