Techno-Science Research Journal V14 (Special Issue) (2026) P 78-89

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

An Explainable Hybrid Machine Learning Framework for Student Profiling and Feature Attribution in Mathematics Achievement: Analysis of Cambodian High Schools

Tong Ly[1*], Sokkhey Phauk[1], Sothea Has[1], Dona Valy[2]

[1] *Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

[2] *Department of Information and Communication Engineering, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

Received: 31 March 2026; Revised: 05 July 2026; Accepted: 06 July 2026; Available online: 31 July 2026

Abstract: *In recent years, the application of machine learning in education has gained significant attention for its potential to uncover hidden patterns and improve student learning outcomes. This study employs a hybrid unsupervised-supervised machine learning approach to achieve two primary objectives: first, to uncover hidden learning profiles of Cambodian high school students, and second, to identify and validate the most significant factors contributing to their mathematics achievement. Using a dataset of 7,188 students, a two-phase analytical pipeline integrating unsupervised and supervised learning was developed. In the first phase, k-modes and k-prototypes clustering were applied with multiple internal validation and stability metrics to analyze latent student profiles. This exploratory phase informed a the second phase using CatBoost, Decision Tree, and Random Forest classifiers, with feature significance determined through a consensus of Mean Decrease in Impurity (MDI), Permutation Feature Importance (PFI), and Recursive Feature Elimination (RFE) metrics. Results indicate that k-prototypes produced a superior two-cluster solution (stability score: 0.9941) compared to k-modes (0.1181), segmenting students primarily by math anxiety, grade level. Supervised modeling identified previous semester math achievement, anxiety, and grade level as the most stable predictive features, with the core feature set achieving a cross-validation accuracy of 0.7068. The integrated analytical framework successfully translated computational metrics into actionable educational insights, providing a validated, data-driven foundation for targeted intervention strategies to support mathematics learning in Cambodian high schools.*

Keywords: educational data mining; feature importance; k-prototypes; machine learning; students' profiling

1. INTRODUCTION

1.1. Research background

In Cambodia, the focus on STEM education has spanned for a decade [1]; however, students' interest and academic performance in these subject areas are not satisfactory [2]. According to [3], for example, despite the high demand for human resources in science and engineering subjects, students' enrollment in these subjects is limited. At the high school level, students' enrollment has declined and many more students have shifted from the pure science to the social science track [4]. Various factors contribute either directly or indirectly to students' learning performance in general [5]. These factors stem from students' mental and psychological well-being [6–7], gender, the amount of time the students spend studying learning materials, and students' prior academic performance [8]. Self efficacy in technology, knowledge sharing, social outcome and their enjoyment in learning etc. have been found among those major factors [5]. Extra-curricular activities, students' hobbies and their behaviors have also been identified as significant factors. In the age of technology, students' engagement with digital devices is another critical aspect found to have affected students' learning in general [5, 9–11].

[1]Corresponding author: *Tong Ly*
E-mail: lytongvps2024@gmail.com

In Cambodia, issues surrounding mathematics education have been widely discussed; however, those inferences touch only on a general overview and often rely on personal conjectures or speculative perceptions rather than empirical data. This study addresses this gap by focusing on educational data mining in mathematics education to establish concrete, evidence-based student profiles. Specifically, the research aims to investigate features characterizing students with different performance levels—for example at-risk/poor, satisfactory and high-performing groups. By applying both unsupervised and supervised machine learning techniques, the study investigates the distinct patterns and features associated with students having different mathematics achievement levels. This student profiling study offers a framework for educators to comprehensively understand their students, which can guide pedagogy and future policy decisions within Cambodian secondary school systems.

1.2. Research problem

Determining factors contributing to students' learning in mathematics has attracted attention among educators and researchers across the globe; however, those existing research in Cambodia remains constrained both methodologically and in the aspect of the students' attributes. Few studies, for example [12–13], explored these issues with a small range of factors, yet these studies utilized simple statistical methods such as t-test, anova or from qualitative approach, limiting their capacity to uncover the patterns or complex multivariate relationships. Methodologically, there exist several studies that have explored these issues using machine learning and deep learning methods, especially in predicting the students' learning in mathematics [14–19]. These studies showed promising results in prediction, but none provided concrete empirical findings on the profiles characterizing students in different performance groups. Consequently, the field lacks an approach that uncovers robust student groupings and identifies their key learning drivers.

In Cambodia, the application of advanced predictive methods—such as multiple linear regression, path modeling, structural equation modeling (SEM), or machine learning and deep learning techniques in educational data mining (EDM) is a challenging task, especially among educators and researchers with limited knowledge and skills in mathematics, statistics or machine learning. To bridge this gap, the present study proposes an explainable hybrid pipeline for student profiling. The framework first uncovers students' learning profiles through two unsupervised clustering techniques: k-modes and k-prototypes with multi-metric cluster validation. It then identifies the most important predictors of the students' mathematics achievement using explainable base and ensemble supervised models integrated with model-agnostic feature importance. This dual-phase design not only addresses the local methodological gap—where analyses have been restricted to t-tests and ANOVA [12–13]—but also tackles the student profiling task which has received limited attention in the previous studies [14–19]. By integrating the rigorous clustering techniques and explainable AI, the study thereby delivers both methodological advancement and actionable insights for Cambodian mathematics education.

1.3. Research objectives

Utilizing the power of machine learning approaches, this study has the following two objectives:

- To explore the hidden learning profiles, in the context of mathematics learning, of Cambodian high school students using unsupervised machine learning techniques.
- To identify and validate the significant factors contributing to high school students' mathematics achievement using supervised learning models combined with model-agnostic feature importance techniques.

1.4. Significance of the study

This study provides significant contributions in both methodological and practical aspects by developing an explainable and interpretable, data-driven framework that integrates unsupervised clustering and supervised machine learning techniques to understand the features contributing to students' learning in mathematics. Methodologically, the hybrid pipeline moves beyond the black-box prediction or classification tasks by thoroughly validating student profiles through clustering techniques with k-modes and k-prototypes and identifying their key drivers using model-agnostic, explainable AI. Practically, the proposed study framework translates complex multivariate relationships into clear, actionable student profiles, providing teachers and educators with an evidence-based tool for targeting low-performing learners and at-risk students. These evidence-based findings help them with early intervention and strategic educational decision-making to improve the students' mathematics learning and beyond.

2. LITERATURE REVIEW

In the educational landscape, the practice of data science methods as well as Artificial Intelligence (AI) to understand related educational issues has become a transformative approach to enhancing academic outcomes [20]. The application of these advanced techniques has also become increasingly prominent when the volume of educational data is getting larger, and when the objectives of the study involve more complex data structures. In education, the utilization of data science ranges from predicting students' learning,

identification of at-risk students, identifying features contributing to the students' learning as well as developing intelligent systems with actionable AI and recommendation frameworks.

The educational data mining (EDM) community has increasingly employed machine learning to uncover complex patterns in student learning, moving beyond traditional statistical methods [21–22]. However, two parallel strands of research—unsupervised profiling of student subgroups and supervised prediction of achievement—have largely developed in isolation, with significant methodological limitations that the present study explicitly addresses.

2.1. Statistical methods in educational data mining

In educational data mining (EDM), statistical methods have been extensively used and have become an effective foundation for studying educational issues. Descriptive statistical techniques such as data visualization and numerical measures, as well as inferential techniques such as t-tests, ANOVA, chi-square, correlation, and regression have been used to study relationships or the effects of related factors surrounding students' learning [12, 22–24]. These foundational approaches are essential for initial data exploration and hypothesis testing, allowing researchers to identify significant patterns in areas like academic performance and engagement. Advanced statistical methods, such as multiple linear regression, path modeling and structural equation modeling (SEM), etc., provide superior advantages in studying latent variables and detecting underlying relationships in educational data [25–26].

2.2. Machine learning in educational data mining

With the increase in data volume and the emergence of machine learning, the application of machine learning models has evolved significantly. Models such as K-Nearest Neighbors (KNN), Naive Bayes (NB), and decision trees have become popular tools. Additionally, ensemble methods, such as random forest and XGBoost, possess superior ability and provide insight into feature importance [15–19, 27–30]. Decision trees, for example, provide intuitive if-then rules but are prone to overfitting. Random forest [31], on the other hand, mitigates this issue by aggregating many trees and delivers superior performance. Advanced boosting algorithms, such as CatBoost [32], handle categorical features more effectively through ordered target statistics and symmetric tree structures.

Tree-based machine learning algorithms provide robust explainability through model-based characteristics, integrated post-hoc techniques, and feature elimination workflows. They utilize built-in metrics like Mean Decrease in Impurity (MDI) feature importance [31], alongside external frameworks like Permutation Feature Importance (PFI) [33]. For local and global explainability, metrics such as Local Interpretable Model-agnostic Explanations (LIME) [34] and SHapley Additive exPlanations (SHAP) [35] are becoming very popular in EDM. While MDI suffers from high-cardinality bias and inconsistency, PFI and SHAP address these limitations; specifically, SHAP isolates each feature's marginal contribution across all combinations to handle correlated predictors [36]. Furthermore, these metrics are frequently paired with iterative wrapper methods like Recursive Feature Elimination (RFE) [37] to systematically prune redundant variables and optimize model parsimony.

In addition, to identify latent student profiles from survey data without labeling, unsupervised learning techniques are becoming comparatively important [22, 38]. Algorithms such as k-means, k-modes, hierarchical clustering, PCA, and t-SNE [38–40], group students based on survey responses and behavioral data. For example, k-means groups students by minimizing within-cluster variance using Euclidean distance [29] in numerical data. Conversely, k-modes is used to handle non-numeric attributes by measuring dissimilarity based on simple matching coefficients using modes [41]. To address mixed-type datasets, k-prototypes combines these approaches by leveraging a weighted sum of Euclidean distances for numerical features and matching coefficients for categorical variables [29]. Furthermore, advanced methods that incorporate dimensionality reduction techniques—such as Principal Component Analysis (PCA) for continuous variables, Multiple Correspondence Analysis (MCA) for categorical variables, and Factor Analysis of Mixed Data (FAMD) for mixed-type datasets—compress high-dimensional student attributes into a lower-dimensional subspace before applying clustering techniques.

3. METHODOLOGY

3.1. Research processes

The overall research design and operational steps employed in this study are systematically outlined in the researcg flowchart presented in Fig 1 below.

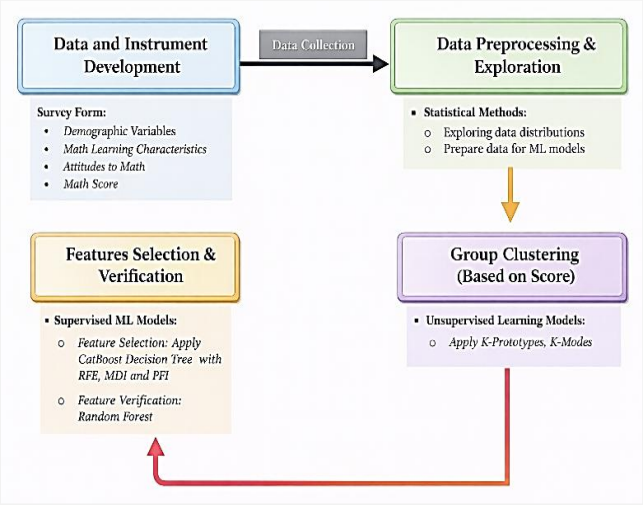

Fig 1: The overall research processes

The overall workflow outlined in Figure 1 is structured into three primary phases:

- Data preparation: Data was collected using a structured survey method and processed using Exploratory Data Analysis (EDA) to detect outliers, address missing values through imputation, and understand feature distributions.
- Unsupervised profiling: Unsupervised clustering algorithms were applied to establish the optimal cluster count (k) and map out hidden, latent sub-profiles among the student performance groups.
- Supervised verification: Supervised learning algorithms were applied to detect significant features contributing to the students' math achievement as well as to validate the key variables driving the cluster formations.

3.2. *Research dataset*

The dataset consists of 7,188 samples and 13 variables, with features comprising students' general information, math learning characteristics, and math anxiety. The target variable is math score, recorded on a total scores of 100, 125 and 150 depending on grade level. For uniform measurement, the score was normalized using a proportional method in which the decimal value multiplied by 100 was taken as the raw score. The general characteristics of these varables are presented in Table 1 below:

Table 1. *Measurement scales of each feature in the dataset*

Attribute	Attribute	Type
Features	Age	Scale
	Gender	Nominal
	Grade	Nominal
	School location	Nominal
	Type of school	Nominal
	Extra-paid math (last year)	Binary
	Extra-paid math (this year)	Binary
	Previous Sememster Score	Nominal
	Member good at math	Binary
	Class attendance	Ordinal
	Math self-learning time/week	Ordinal
	Math Anxiety	Scale
Target	Math Score	Scale

3.3. *Proposed models in the study*

To achieve the two objectives, the unsupervised phase employed k-modes and k-prototypes to explore hidden learning profiles among Cambodian high school students, while the supervised phase utilized CatBoost, decision tree, and random forest—combined with model-agnostic feature importance techniques—to identify and validate the significant factors contributing to mathematics achievement, as detailed in Table 2.

Table 2. Models used and metrics used in the study

Model and Methods	Metrics
A. Unsupervised Models: ▪ K-Modes ▪ K-Prototypes	- Silhouette Score, Davies-Bouldin Index, Dunn Index, Elbow Method, Stability Measures
B. Supervised Models (Classification): ▪ CatBoost ▪ Decision Tree ▪ Random Forest - RFE, MDI and PFI, CV and Stability analysis	- Accuracy, F1-Score, Recall, Precision

The unsupervised algorithms k-modes and k-prototypes are specifically used to handle categorical and mixed data respectively [29]. K-modes uses cluster centroids based on modes—the most frequent values rather than the mean values employed by k-means. Following [29], the technique applies a simple mismatch-based dissimilarity measure:

$$d(X_i, Q_l) = \sum_{j=1}^{m} \delta(x_{i,j}, q_{l,j}) \qquad (Eqn\ 1)$$

where $\delta(p, q) = 0$ if $p = q$ and $\delta(p, q) = 1$ if $p \neq q$.

A notable limitation of k-modes is its sensitivity to the initial mode selection, which can cause local optima and cluster instability across different initializations.

In contrast, the k-prototypes offers a hybrid approach integrating k-means for numerical attributes and k-modes for categorical attributes within a unified cost function [29]. The algorithm integrates squared Euclidean distance for numerical

attributes and simple matching distance for categorical attributes as shown in Eqn 2.

$$d(X_i, Q_l) = \sum_{j=1}^{p} \left(x_{i,j} - q_{l,j}\right)^2 + \gamma \sum_{j=1}^{m} \delta\left(x_{i,j}, q_{l,j}\right) \quad (Eqn\ 2)$$

where γ balances the influence of both attribute types.

K-prototypes can handles mixed-type data; however, appropriate weighting of numerical and categorical attributes remains challenging for optimal cluster quality.

For labelled data, especially predicting or classification tasks, some tree-based models such as decision tree, CatBoost, random forest, xgboost etc. were proven to provide significant insights with feature importance for explanability. CatBoost, for example, has superior power for categorical features through ordered target statistics and symmetric tree structures [32].

CatBoost is a gradient boosting technique using binary decision trees (*Eqn 3*) as base predictors. The decision trees algorithm divides the feature space $\mathcal{R}^m$ into several disjoint regions (tree nodes) according to the values of some splitting attributes $a = 1_{\{x^k > t\}}$ where the comparison of the splitting attributes is against the threshold value t.

$$h(x) = \sum_{j=1}^{J} b_j \mathbf{1}_{\{x \in \mathcal{R}_j\}} \quad (Eqn\ 3)$$

where $\mathcal{R}_j$ are disjoint regions corresponding to the leaves of the tree.

The decision tree is a supervised model that recursively splits data into homogeneous subsets based on the values of different attributes (if-then-else rules) until a stopping criterion is met. The result is a tree-like structure, where each node represents a split or decision based on a specific attribute [42]. Decision trees are sensitive to overfitting, resulting in instability compared to ensemble methods. Random forest, on the other hand, is an ensemble method capable of dealing with the issues in decision trees. The model addresses the overfitting issue through the construction of multiple decision trees on random bootstrap samples of data and random feature subsets, combining predictions through majority voting or averaging to improve accuracy and control overfitting [31].

3.4. Metrics and feature importance

In clustering validation, different metrics have been used to assess cluster quality and stability. These metrics evaluate clustering results from complementary perspectives, including compactness, separation, and reproducibility. The Silhouette Score, for example, measures how closely an object or feature is connected relative to neighboring clusters with higher value indicating well-separated groupings [43]. The Davies–Bouldin Index evaluates the average similarity between clusters normalized by within-cluster compactness with the lower values reflecting better separation among clusters [44]. The Calinski–Harabasz Index [45] measures the variance ratio comparing between-cluster dispersion to within-cluster dispersion with larger values suggesting more distinct clusters. The Dunn Index measures the balance of inter-cluster distance against intra-cluster diameter, with higher values indicating better balance [46]. Stability Measures are metrics used to assess clustering consistency across data variations or initializations, where higher value indicates the greater reproducibility [47]. Cluster Balance [43], also known as the coefficient of variation of cluster sizes (CV) and the Separation Quality [48] are often used in the silhouette-based calculation. CV and Separation Quality provide additional descriptive insights into the quality of the separation based on the Silhouette Score.

Table 3. Metric and criteria for clustering comparison

Metrics	Value	Preferable value
Silhouette Score	-1 to 1	1(perfect), 0(overlap) and -1(misclassified)
Davies-Bouldin	0 to ∞	Closer to 0 (better)
Calinski-Harabasz	0 to ∞	Higer is better (>300)[2]
Dunn Index	0 to ∞	Higer is better (>0.7)[3]
Stability Score	0 to 1	Higer is better (>0.5)
Cluster Balance CV)	0 to ∞	Closer to 0 (more balance cluster)
Separation Quality	-1 to 1	Higher better

Note: These metrics are intended for comparison of K-modes and K-prototypes for stable model in identifying the optimal 'k'.

Tree-based supervised learning models are valued for their explainability through feature importance integration [49]. These techniques assign scores to input features based on their contribution to predictions (target), thus enhancing the model interpretability, guiding feature selection, and improving predictive performance of the models [31, 50]. These methods are broadly categorized as model-specific such as Gini importance in tree-based models or coefficient magnitudes in linear regression [51] or model-agnostic, like permutation importance, which shuffles the feature values to assess the performance changes [37, 49, 52]. In addition, Altmann et al. [33] demonstrated that permutation importance outperforms Mean Decrease in Impurity (MDI) in reliability. A more advanced technique, SHAP (SHapley Additive exPlanations), works to quantify each feature's marginal contribution to individual predictions by considering all possible feature combinations [35]. In this

[2] The value Calinski-Harabasz >300 is generally accepted for general data but not theorethically stated. The value may reach thousand or can be 200 for high dimensional and complex data.

[3] The value Dunn Index >0.7 is commonly accepted but not theorethically fixed. The value may vary depending on dimensions or complexity of the data.

case, SHAP values ensure fair distribution of the model's prediction outcomes, balance contributions among correlated features, and provide local accuracy and consistency, reflecting the model's true behavior [35, 53].

3.5. *Data preprocessing and exploration*

Data preprocessing revealed 137 duplicates, with less than 5% of missing values in each column. Duplications were removed to ensure data integrity. Missing values for the numerical variables, Anxiety and Score, were imputed using median substitution due to non-normal distributions. Missing values for categorical variables (<=0.1%) were removed due to their minimal proportions. The attribute 'SchType' was excluded due to sample proportion imbalance among the different levels in the variable. The Variable 'Age' was categorized into binary groups while 'SelfLearnPerWeek' was regrouped into two levels ("*<3 hours" and* "≥3 hours"). The feature 'Attendance' was consolidated into three categories ("*Always*", "*Often*", "*Less Frequent*"). A small number of outliers were detected for the target variable 'Score', but they were retained because sensitivity analysis showed negligible impact on model performance.

3.6. *Model's applications and experiments*

The applications of the unsupervised and supervised models were conducted sequentially to form an integrated analytical pipeline. Initially, unsupervised models (K-Modes and K-Prototypes) were utilized to explore the general learning characteristics of students based on their mathematics achievement scores. Subsequently, supervised models (CatBoost, Decision Tree, and Random Forest) built upon these foundational insights to identify and validate the specific features that significantly contribute to academic performance, thereby providing a complementary layer of analytical depth. The execution of these proposed models was structured according to the following experimental setup:

- Stage 1: Unsupervised clustering
  First, student data were partitioned using K-Modes and K-Prototypes algorithms. To optimize the clustering process, the models were iteratively executed across a range of cluster sizes to determine the optimal *k*value based on diverse cluster evaluation metrics. Following this evaluation, the more robust model was deployed with a fixed cluster configuration (specifically, k-Prototypes with $k = 2$) to isolate and identify the distinct profiles characterizing each student group, utilizing their math performance as a baseline reference.
- Stage 2: Supervised classification
  Following the unsupervised stage, explicit feature sets were defined for experimental evaluation, and the continuous target variable ('Score') was binarized into two distinct groups ($k = 2$) using the overall mean score as the classification threshold:

$$Score\ Group = \begin{cases} Group\ 1\ (High), \text{ if } Score \geq 76.85 \\ Group\ 2\ (Low), \text{ if } Score < 76.85 \end{cases}$$

The categorized dataset was subsequently partitioned into training and testing sets using an 80:20 split ratio. Finally, three machine learning classifiers—CatBoost, Decision Tree, and Random Forest—were trained and validated using multiple feature importance techniques, including Mean Decrease in Impurity (MDI), Permutation Feature Importance (PFI), and Recursive Feature Elimination (RFE), to isolate the primary features driving math achievement.

4. ANALYSES AND RESULTS

In this section, the results are presented in two parts based on the experiments and analyses of unsupervised and supervised learning models.

4.1. *Results of unsupervised learning models*

Table 4 reveals the optimal values of '*k*' from k-modes and k-prototypes. K-modes grouped the data into 2 and 4 clusters *(k=2, k=4)*, considering different metrics while k-prototypes detected only 2 clusters (*k=2*) across all metrics. These findings are evidenced by Figure 2 to Figure 5, which follow the table.

Table 4. The optimal values of '*k*' identified by k-modes and k-prototypes

Method	Optimal values for 'k'	
	k-modes	k-prototypes
Silhouette Score	2	2
Davies-Bouldin	4	2
Calinski-Harabasz	2	2
Combined Metrics	4	2

Figure 2 to Figure 5 reveals the identification of the optimum value '*k*'for k-modes and k-prototypes across the four different techniques and metrics.

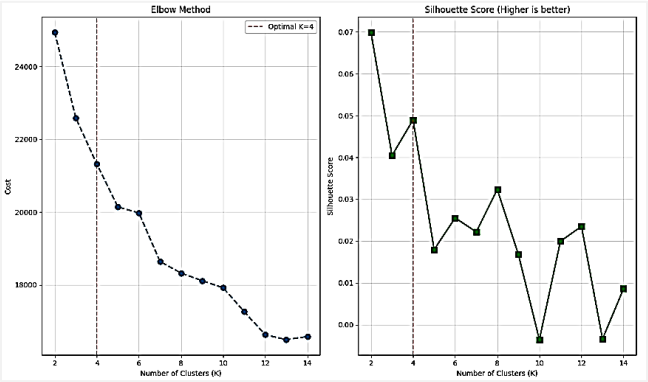

Fig 2. Cluster separation (k-modes) based on Elbow and Silhouette Score

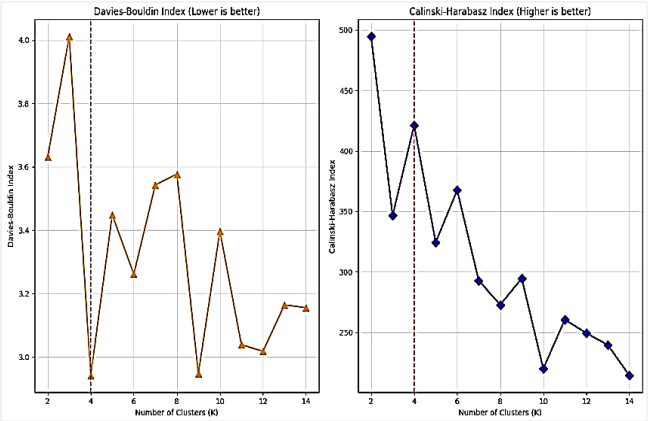

Fig 3. Cluster separation (k-modes) based on Davies-Bouldin and Calinski-Harabasz

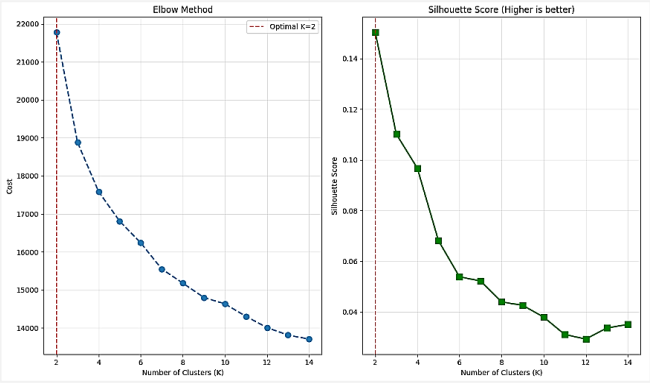

Fig 4. Cluster separation (k-prototypes) based on Elbow and Silhouette Score

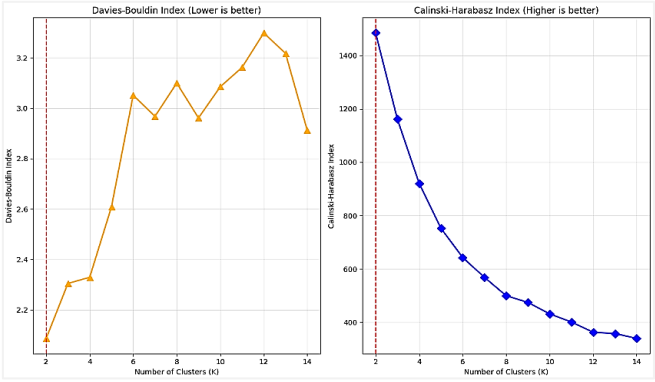

Fig 5. Cluster separation (k-prototypes) based on Davies-Bouldin and Calinski-Harabasz

Table 5 presents the comparative performance statistics of the clustering models. K-prototypes outperformed k-modes across the evaluated metrics.

Table 5. Comparative clustering statistics from k-modes and k-prototypes

Method	Optimal values for '*k*'	
	k-modes	k-prototypes
Silhouette Score	0.0549	0.1504
Davies-Bouldin	3.4460	2.0860
Calinski-Harabasz	485.9153	1483.9919
Dunn Index	0.1543	0.0195
Stability Score	0.1181 ± 0.0772	0.9941 ± 0.0063
Cluster Balance	0.3154	0.1235
Separation Quality	0.0549	0.1504

Note: For '*Stability Score*', higher is better while '*Cluster Balance (CV)*', lower is better (see Table 3 for details).

A comparative analysis indicates that k-prototypes offers a more viable clustering solution than k-modes due to its greater stability and more balanced cluster distribution. Both algorithms exhibited weak point-level separation, as reflected by low silhouette scores (*k-modes: 0.0549; k-prototypes: 0.1504*) and unfavorable Davies–Bouldin scores (*k-modes: 3.4460; k-prototypes: 2.0860*), although k-prototypes performed better on both metrics. The superiority of k-prototypes over k-modes was further evident in its higher stability score (0.9941) and substantially larger Calinski–Harabasz index (1483.992 compared to 485.915 for k-modes), demonstrating its suitability for handling mixed data types.

Figure 6 reveals the two key characteristics of the numerical features of students profiles separating the two student clusters.

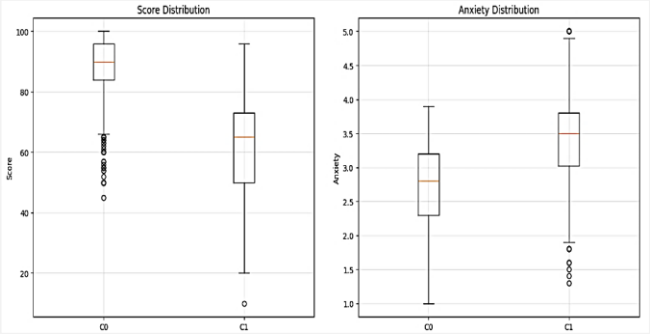

Fig 6. The separation of two numerical variables for each cluster

Two numerical variables played a key role in the cluster separation produced by k-prototypes. A holistic analysis of the key features contributing to each cluster is presented in Table 6. The results showed that the two clusters shared the majority of common features, with the exception of 'Anxiety', 'SchLocat', and Grade. Differences in these three features signified the key distinction between the two performance groups.

Table 6. Characteristics of key features contributed to 'Score' for the identified clusters (k-prototypes)

Cluster	Sample	Avg.Score	Feature separation (% in cluster)
1	3899 (56.2%)	88.68	▪ *Anxiety (<2.72)* ▪ Age <18 (76.5%) ▪ Female (66.0%) ▪ *Grade 10 (42.3%)* ▪ *Districts' (46.6%)* ▪ ExtraPreYear: 'Yes' (77.6%)
2	3042 (43.8%)	61.68	▪ *Anxiety (>3.42)* ▪ Age<18 (59.2%) ▪ Female (64.6%) ▪ *Grade 12 (43.6%)* ▪ *Phnom Penh (51.4%)* ▪ ExtraPreYear: 'Yes' (60.1%)

The two identified clusters reveal two contrasting performance groups—high and low performers—based on anxiety and grade level, which subsequently provide feature segmentation to inform the supervised learning applications that follow.

4.2. Results of supervised learning models

Figures 7 and Figure 8 present a comparison of feature importance derived from the two supervised models, CatBoost and Decision Tree. A threshold value of 5% was applied to determine the significant contribution of each feature to the target variable. The significance of each feature was assessed based on consistency across three importance metrics: Mean Decrease in Impurity (MDI), Permutation Feature Importance (PFI), and Recursive Feature Elimination (RFE).

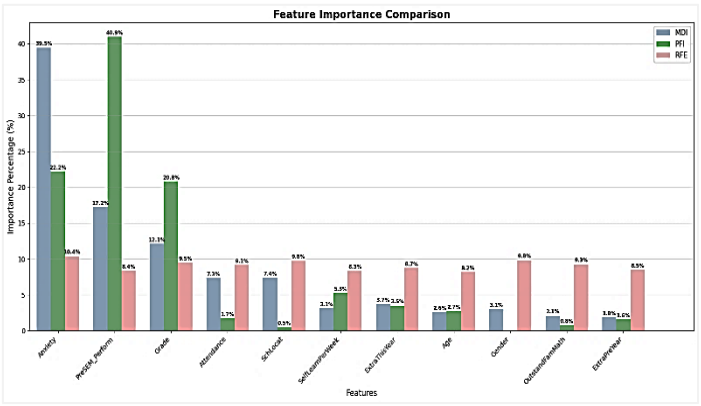

Fig 7. Feature importance from CatBoost (Acc. = 0.7221)

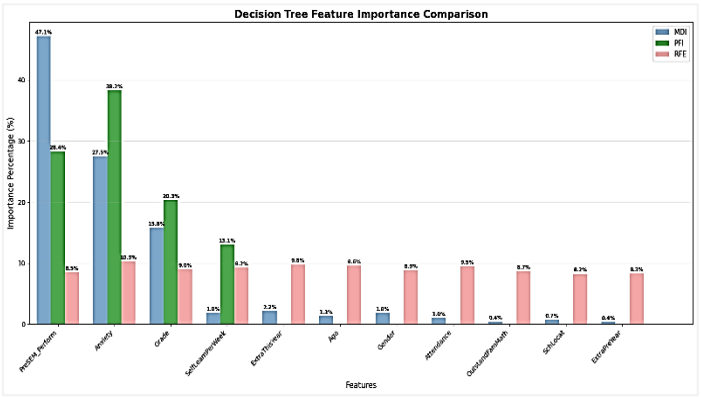

Fig 8. Feature importance from decision tree (Acc. = 0.7048)

According to [33], PFI is considered a more stable and reliable measure of feature importance than MDI because its calculation utilizes out-of-bag samples and is less biased by the scale or number of categories in the features. Thus, features with PFI > 5% in both CatBoost and Decision Tree were considered significant. Three significant features—'PreSEM_Perform', 'Anxiety', and 'Grade'—were identified. Some features exhibited inconsistencies across RFE, MDI, and PFI; however, those with PFI > 5%, even if scoring lower on other metrics, were retained and included in the validation phase with the random forest. Other features, such as 'SelfLearnPerWeek', 'SchLocat', and 'Attendance', were added at the validation stage based on domain knowledge. The inclusion of these features led to the following model experiments based on feature sets.

- Core: a set of significant features identified from CatBoost and decision tree (PFI >5%). This set contains three feature variables, PreSEM_Perform, Anxiety and Grade.

- Set1, Set2 and Set3, the increments of features, '*SelfLearnPerWeek*', '*SchLocat*' and '*Attendance*' without removing.

Table 8 presents the validation results from the random forest, which were evaluated based on feature stability rather than predictive performance (accuracy). The results indicate that the Core feature set yielded better performance compared to the incremental feature sets (Set 1, Set 2, and Set 3).

Table 7: Comparisions of Cross-Validation Accuracy across different feature sets using random forest

Feature Set	CV Acc	CV Std	Acc	ROC-AUC	F1-Score
Core	***0.7068***	***0.0108***	***0.5045***	***0.4894***	***0.5682***
Set1	0.7005	0.0118	0.5063	0.4914	0.5695
Set2	0.6816	0.0076	0.4992	0.4889	0.5580
Set3	0.6692	0.0072	0.4963	0.4878	0.5585

Note: the higher values of CV Accuracy within one CV Std indicate better performance

The findings confirmed two features from the Core Set—'Anxiety' and 'Grade'—which were also identified by k-prototypes, while 'PreSEMPerform' emerged as an additional feature identified by the supervised learning models. A progressive decline in predictive performance was observed as additional features were introduced. For the Core Set, the deviation between the CV accuracy (~70%) and the test accuracy (~50%), coupled with a ROC-AUC near 0.5, suggests that while the core features capture internal patterns, the model currently struggles with generalization. The finding also indicates that the added features, such as 'Attendance' and 'School Location', may introduce noise or multicollinearity rather than enhancing the model's discriminative power.

5. DISCUSSION AND LIMITATIONS

The findings reveal the superiority of k-prototypes over k-modes based on the stability score and Calinski–Harabasz index. Although the majority of the variables in the dataset are categorical, k-prototypes performed effectively in handling mixed data types. This finding supports the evidence of the superior performance of k-prototypes in handling mixed-type data reported by [29]. However, despite the robust segmentation, the low Silhouette Score (0.1504) and Dunn Index (0.0195) reveal significant overlap at cluster boundaries. The low Dunn Index value indicates that the intra-cluster diameter (the spread of data within a group) is disproportionately large compared to the inter-cluster separation. This reflects the complex, non-linear nature of behavioral data, where distinct categories often share dense, interconnected boundaries rather than clear, isolated gaps. This finding aligns with empirical evidence on cluster separation presented by [55], who noted that perfect separation is rare in such contexts. Nonetheless, the findings identified three key features—math anxiety, school location, and grade level—that signify the distinction between high- and low-performing learner groups.

The findings from the supervised models confirmed two overlapping features—math anxiety and grade level—which were also identified by k-prototypes; however, the supervised model experiments suggested an additional feature, 'PreSEM_Perform' (i.e., students' prior semester mathematics outcome), as a key feature contributing to their math score. The inclusion of the variable 'SchLocat', which was identified by k-prototypes, added no significant improvement to the supervised model performance.

This hybrid framework, although it yielded two inconsistent features—'SchLocat' and 'PreSEM_Perform'—provides significant contributions to methodological insights by offering a framework for understanding student profiling. The findings also fill the gap where simple statistical methods were previously used to understand students' mathematics learning [12], and extend beyond the prediction-focused approaches employed in previous studies [12, 16-17, 19].

6. CONCLUSION

The study successfully achieved its objectives by applying a hybrid data science pipeline that integrated unsupervised and supervised machine learning to uncover hidden learning profiles distinguishing students based on their math achievement. The experiments with two unsupervised models, k-modes and k-prototypes, demonstrated that k-prototypes performed better in grouping students for the provided mixed-type dataset. The algorithms consistently revealed two clusters, while the validation from k-prototypes revealed two clusters based on math performance (Score) that had two key features: 'math anxiety' and 'grade level'. Other features were common across the two clusters. The initial feature selection using two supervised models, CatBoost and decision trees, added the variable related to the students' previous semester math score to the list from the unsupervised models. The feature validation using random forest with different feature sets revealed that previous semester math achivement', 'math anxiety' , and 'grade level' were key determinants of the students' math performance. The present study contributes to educational data mining and Cambodian mathematics education by demonstrating a reproducible, end-to-end framework using advanced techniques in machine learning with interpretability and actionable insights. The study offers educators and policymakers a strong foundation for designing intervention programs such as math anxiety-reducing teaching method and math skill-building measures.

REFERENCES

[1] S. Un, N. Sin, and R. Chhim, "Improved STEM Education in Cambodia through Spatial Analysis of

Secondary Resource Schools," Cambodia Educ. Rev., vol. 4, no. 1, pp. 52–86, 2021.
[2] S. Kao and S. Kinya, "A Review on STEM Enrollment in Higher Education of Cambodia: Current Status, Issues, and Implications of Initiatives," Int. Dev. Coop., vol. 26, no. 1&2, pp. 123–134, 2020, [Online]. Available: https://www.researchgate.net/publication/339815574_A_Review_on_STEM_Enrollment_in_Higher_Education_of_Cambodia_Current_Status_Issues_and_Implications_of_Initiatives
[3] JICA, "Data Collection Survey on Human Resource Development for Industrialisation in the Education Sector in the Kingdom of Cambodia. Final Report," 2016.
[4] S. KAO and S. KINYA, "Factors affecting students' choice of science and engineering majors in higher education of Cambodia," Int. J. Curric. Dev. Pract., vol. 21, no. 1, pp. 69–82, 2019, doi: 10.18993/jcrdaen.21.1.
[5] M. Alsarayreh and A. Aljaafreh, "Factors influencing students ' academic performance in universities : Mediated by knowledge sharing behavior," Knowl. Manag. E-Learning, vol. 15, no. 4, pp. 643–671, 2023, doi: https://doi.org/10.34105/j.kmel.2023.15.036.
[6] N. Tape et al., "The impact of psychological well-being and ill- being on academic performance : a longitudinal and cross-sectional study," Educ. Dev. Psychol., vol. 38, no. 2, pp. 206–214, 2021, doi: 10.1080/20590776.2021.1986356.
[7] M. B. Mustafa et al., "The Relationship between Psychological Well-Being and University Students Academic Achievement The Relationship between Psychological Well-Being and University Students Academic Achievement," vol. 1, no. 7, pp. 518–525, 2020, doi: 10.6007/IJARBSS/v10-i7/7454.
[8] L. Al-Alawi, J. Al Shaqsi, A. Tarhini, and A. S. Al-Busaidi, "Using machine learning to predict factors affecting academic performance: the case of college students on academic probation," Educ. Inf. Technol., vol. 28, no. 10, pp. 12407–12432, 2023, doi: 10.1007/s10639-023-11700-0.
[9] C. I. Espinoza Lema, Y. D. Alvarado Cabrera, G. I. Sagnay Andrade, and S. A. Villaprado Véler, "Effective use of digital tools to enhance teaching and Learning," Rev. Estud. Gen., vol. 4, no. 2, pp. 393 – 409, 2025, doi: 10.70577/reg.v4i2.100.
[10] S. Gopinathan, A. H. Kaur, S. Veeraya, and M. Raman, "The Role of Digital Collaboration in Student Engagement towards Enhancing Student Participation during COVID-19," Sustainability, vol. 14, no. 11, p. 6844, 2022, doi: https://www.mdpi.com/2071-1050/14/11/6844.
[11] M. Tania, "The Role of Digital Learning Tools in Improving Students ' English Skills : A Systematic Literature Review," vol. 4, no. 2, pp. 702–711, 2025.
[12] T. Ly, S. Phauk, and S. Chea, "Relationship between Students' Attitudes toward Mathematics and Their Achievement in Probabilities and Statistics: A Case of Engineering Students at ITC," Cambodian J. Humanit. Soc. Sci., vol. 1, no. 1, pp. 36–53, 2022.
[13] T. Ly and B. Takuya, "Analysis of Cambodian Students' Errors in Solving Quadratic Function Problems Using Newman's Error Analysis," Cambodian J. Humanit. Soc. Sci., vol. 2, no. 1, pp. 32–44, 2023.
[14] S. Phauk and T. Okazaki, "Integration of Educational Data Mining Models to a Web-Based Support System for Predicting High School Student Performance," Int. J. Comput. …, vol. 15, no. 2, pp. 131–144, 2021, [Online]. Available: https://www.researchgate.net/profile/Sokkhey-Phauk/publication/349226410_Integration_of_Educational_Data_Mining_Models_to_a_Web-Based_Support_System_for_Predicting_High_School_Student_Performance/links/6025c8a5299bf1cc26bcd5e7/Integration-of-Educational-D
[15] S. Phauk, N. Sin, T. Ly, and O. Takeo, "Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in cambodia," IEIE Trans. Smart Process. Comput., vol. 9, no. 3, pp. 217–229, 2020, doi: 10.5573/IEIESPC.2020.9.3.185.
[16] S. Phauk and O. Takeo, "Hybrid machine learning algorithms for predicting academic performance," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 32–41, 2020, doi: 10.14569/ijacsa.2020.0110104.
[17] S. Phauk and O. Takeo, "Comparative Study of Prediction Models for High School Student Performance in Mathematics," IEIE Trans. Smart Process. Comput., vol. 8, no. 5, pp. 394–404, 2019, doi: 10.5573/IEIESPC.2019.8.5.394.
[18] S. Phauk and O. Takeo, "Development and optimization of deep belief networks applied for academic performance prediction with larger datasets," IEIE Trans. Smart Process. Comput., vol. 9, no. 4, pp. 298–311, 2020, doi: 10.5573/IEIESPC.2020.9.4.298.
[19] S. Phauk and O. Takeo, "Study on dominant factor for academic performance prediction using feature selection methods," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 492–502, 2020, doi: 10.14569/IJACSA.2020.0110862.
[20] R. Abdrakhmanov, A. Zhaxanova, M. Karatayeva, G. Z. Niyazova, K. Berkimbayev, and A. Tuimebayev, "Development of a Framework for Predicting Students' Academic Performance in STEM Education using Machine Learning Methods," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 1, pp. 38–46, 2024, doi: 10.14569/IJACSA.2024.0150105.
[21] C. Romero and S. Ventura, "Educational data mining and learning analytics : An updated survey," Data Min. Knowl. Discov., vol. 10, no. 3, pp. 1–21, 2020, doi:

10.1002/widm.1355.
[22] R. S. Baker and P. S. Inventado, “Educational Data Mining and Learning Analytics,” in Research to Practice, New York: Springer Science+Business Media, 2014, pp. 61–75. doi: 10.1007/978-1-4614-3305-7.
[23] M. D. Rahul, “A Review of Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences (3rd ed .),” J. Educ. Behav. Stat., vol. 30, no. 2, pp. 227–229, 2005.
[24] S. Slater, S. Joksimovic, V. Kovanovic, and R. Baker, “Tools for educational data mining: a review,” J. Educ. Behav. Stat., vol. 42, no. 1, pp. 85–106, 2017, doi: 10.3102/1076998616666808.
[25] M. S. Khine, “Structural Equation Modeling Approaches in Educational Research and Practice,” in Application of Structural Equation Modeling in Educational Research and Practice, M. S. Khine, Ed., 2013, pp. 279–283.
[26] T. Teo, L. T. Tsai, and C.-C. Yang, “Applying Structural Equation Modeling (SEM) in Educational Research: An Introduction,” in Application of Structural Equation Modeling in Educational Research and Practice, M. S. Khine, Ed., 2013, pp. 3–21. doi: 10.1007/978-94-6209-332-4_1.
[27] A. Almalawi, B. Soh, A. Li, and H. Samra, “Predictive Models for Educational Purposes: A Systematic Review,” Big Data Cogn. Comput., vol. 8, no. 187, pp. 1–42, 2024, doi: https://doi.org/10.3390/ bdcc8120187 Academic.
[28] C. Ganesh and E. K. Reddy, “Overview of the Predictive Data Overview of the Predictive Data Mining Techniques,” Int. J. Comput. Sci. Eng., vol. 10, no. 1, pp. 28–36, 2022, doi: https://doi.org/10.26438/ijcse/v10i1.2836.
[29] Z. Huang, “Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values,” Data Min. Knowl. Discov., vol. 2, no. 3, pp. 283–304, 1998.
[30] N. Nurul, H. Mat, S. Deris, A. Mat, M. S. Mohamad, and S. S. Safaai, “Review on Predictive Modelling Techniques for Identifying Students at Risk in University Environment,” MATEC Web Conf., vol. 255, no. 2019, 2019, doi: https://doi.org/10.1051/matecconf/201925503002.
[31] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.
[32] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost : unbiased boosting with categorical features,” in Advances in Neural Information Processing Systems, 2018, pp. 6638–6648.
[33] A. Altmann, L. Tolo, O. Sander, and T. Lengauer, “Permutation importance : a corrected feature importance measure,” BIOINFORMATICS, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.
[34] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘ 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, San Francisco: ACM, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
[35] S. M. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” in 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017.
[36] A. Fisher, C. Rudin, and F. Dominici, “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously,” J. Mach. Learn. Res., vol. 20, no. 117, pp. 1–81, 2019.
[37] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” Mach. Learn., vol. 46, no. 2002, pp. 389–422, 2002.
[38] I. T. Jolliffe and J. Cadima, “Principal component analysis : a review and recent developments,” Phil.Trans.R.Soc.A, 2016, doi: http://dx.doi.org/10.1098/rsta.2015.0202.
[39] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008, [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html
[40] C. Márquez-vera, A. Cano, C. Romero, and S. Ventura, “Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data,” Appl. Intell., 2013, doi: 10.1007/s10489-012-0374-8.
[41] A. Chaturvedi, P. E. Green, and J. D. Caroll, “K-modes Clustering,” J. Classif., vol. 18, pp. 33–55, 2001, doi: 10.1007/s00357-001-0004-3.
[42] R. Rivera-lopez, J. Canul-reich, E. Mezura-montes, and M. A. Cruz-chávez, “Induction of decision trees as classification models through metaheuristics,” Swarm Evol. Comput., vol. 69, no. March 2022, p. 101006, 2022, doi: 10.1016/j.swevo.2021.101006.
[43] P. J. Rousseeuw, “Silhouettes : a graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, no. 1987, pp. 53–65, 1987, doi: https://doi.org/10.1016/0377-0427(87)90125-7.
[44] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 1, no. 1979, pp. 224–227, 1979, doi: http://dx.doi.org/10.1109/TPAMI.1979.4766909.
[45] T. Caliński and J. Harabasz, “Communications in Statistics A dendrite method for cluster analysis,” Commun. Stat., vol. 3, no. 1, pp. 1–27, 1974, doi: http://dx.doi.org/10.1080/03610927408827101.
[46] J. C. Dunn, “A Fuzzy Relative of the ISODATA

Process and Its Use in Detecting Compact Well-Separated Clusters," J. Cybern., vol. 3, no. 3, pp. 32–57, 2008, doi: http://dx.doi.org/10.1080/01969727308546046.
[47] U. von Luxburg, "Clustering Stability : An Overview," Found. Trends Mach. Learn., vol. 2, no. 3, pp. 235–274, 2009, doi: 10.1561/2200000008.
[48] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. A JOHN WILEY & SONS, INC., PUBLICATION, 1990.
[49] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub, 2022. [Online]. Available: https://leanpub.com/interpretable-machine-learning
[50] L. Breiman, "Using Iterated Bagging to Debias Regressions," Mach. Learn., vol. 45, pp. 261–277, 2001.
[51] M. Nalluri, M. Pentela, and N. R. Eluri, "A Scalable Tree Boosting System : XG Boost," Int. J. Res. Stud. Sci. Eng. Technol., vol. 7, no. 12, pp. 36–51, 2020, doi: https://doi.org/10.22259/2349-476X.0712005.
[52] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., vol. 3, no. 2003, pp. 1157–1182, 2003.
[53] A. Redelmeier, M. Jullum, and K. Aas, "Explaining Predictive Models with Mixed Features Using Shapley Values and Conditional Inference Trees," in 4th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE), 2021, pp. 117–137. doi: 10.1007/978-3-030-57321-8_7.
[54] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation importance : a corrected feature importance measure," BIOINFORMATICS, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.
[55] A. K. Jain, "Data clustering : 50 years beyond K-means q," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.