Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (97-108)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49369

JURNAL TEKNIK INFORMATIKA
Homepage : http://journal.uinjkt.ac.id/index.php/ti

Comparing K-Prototypes and K-Medoids with Catboost for Health Profile Clustering of Pesantren Students

Moch. Aghisna Hadzikunnuha[1*], Harits Ar Rosyid[2], M. Zainal Arifin[3]

[1,2,3]Electrical and Informatics Engineering, Faculty of Engineering, Universitas Negeri Malang
[1.2.3]Jl. Semarang 5, Kota Malang, 65145, Indonesia

Article:
Accepted: February 08, 2026
Revised: December 12, 2025
Issued: April 30, 2026

***Correspondence Address:**
moch.aghisna.2305348@students.um.ac.id

ABSTRACT

Health screening in pesantren is challenging due to communal living conditions, limited health facilities, and the need for early identification of vulnerable student groups. This study compares the performance of K-Prototypes and K-Medoids clustering for grouping student health profiles and evaluates the use of cluster labels as additional features in a CatBoost classification model. The dataset consists of 1,464 new students from Queen Al Falah Islamic Boarding School in the 2025/2026 academic year, collected through the admission system and analyzed after preprocessing. Clustering is performed using K-Prototypes and K-Medoids with three clusters to support interpretability of nutritional and health profiles. Although two clusters yield higher silhouette values, three clusters provide more meaningful distinctions for practical screening. Classification experiments use CatBoost with an 80:20 stratified train-test split, comparing baseline models and hybrid models that integrate cross-algorithm cluster features. The results show an asymmetric pattern. Adding K-Prototypes features improves K-Medoids target accuracy from 99.66 percent to 100 percent, while adding K-Medoids features slightly decreases K-Prototypes target accuracy from 98.98 percent to 98.63 percent. McNemar test results indicate that these differences are not statistically significant. Overall, the proposed framework supports reliable and interpretable health profile clustering for pesantren student monitoring.

Keywords: *CatBoost, Clustering, K-Medoids, K-Prototypes, Pesantren Students*

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (97-108)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49369

1. INTRODUCTION

The health condition of students in Islamic boarding schools (pesantren) deserves particular attention. Daily activities are dense, students live communally, and access to health facilities is often limited. In such environments, minor health issues can easily escalate, especially among new students who arrive with diverse nutritional backgrounds and medical histories. Several studies have highlighted that pesantren settings are vulnerable to the spread of infectious diseases and recurrent health disorders, driven by hygiene practices, environmental exposure, and unbalanced activity patterns [1], [2], [3]. Despite this, health screening in many pesantren is still conducted in a descriptive and fragmented manner, making it difficult to identify meaningful health patterns early.

From an analytical perspective, student health data in pesantren is not simple. It combines numerical measurements, such as height and weight, with categorical and often multi-label information, such as disease history and current illness. These characteristics reflect real-world conditions but pose challenges for many commonly used data mining methods. Algorithms that assume purely numerical input, such as K-Means, or models that rely heavily on manual feature engineering, often struggle to represent this complexity adequately. As a result, important structural differences between student health profiles may remain hidden.

Machine learning offers tools to move beyond descriptive summaries toward pattern-based analysis and prediction [4], [5]. In health-related studies, clustering is frequently used to group individuals with similar characteristics [6], while classification models are employed to assign new individuals into predefined groups [7]. Some studies have shown that combining clustering and classification can improve predictive performance [8], [9]. However, most of these approaches rely on numerical-only clustering and are not explicitly designed for mixed-type health data such as that found in pesantren student records.

K-Prototypes and K-Medoids are two clustering algorithms that are better aligned with this type of data. K-Prototypes allows numerical and categorical attributes to be processed simultaneously, making it suitable for health profiles that combine anthropometric measurements and disease-related information [10], [11], [12]. K-Medoids, especially when implemented with Gower distance, is known for its robustness to outliers and its use of real observations as cluster representatives, which improves interpretability in heterogeneous datasets [13], [14], [15]. Although both algorithms have been applied in various domains, direct comparison between K-Prototypes and K-Medoids on mixed health data remains limited, particularly in the context of pesantren students.

Another limitation in existing studies is how clustering results are used. Clusters are often treated as final outputs, not as structured information that can enrich subsequent predictive models. At the same time, modern classifiers such as CatBoost have demonstrated strong performance on tabular data with categorical features[16], [17]. Yet, the potential benefit of integrating cluster membership as an additional feature in CatBoost has rarely been explored, especially for health screening in resource-limited educational institutions.

In this study, "disease risk" does not refer to clinical diagnosis or confirmed medical conditions. Instead, it represents health-profile risk groups derived from similarities in nutritional status and disease patterns. These groups are intended to support practical decision-making, such as prioritizing nutritional interventions, identifying students who require closer monitoring, or planning preventive health programs in pesantren environments.

Based on this background, this research addresses three main questions. First, how do K-Prototypes and K-Medoids differ in clustering mixed-type health data of pesantren students? Second, to what extent can cluster membership be predicted using CatBoost as a supervised classification model? Third, does incorporating cluster information from one algorithm improve the prediction performance for the other? We hypothesize that cluster structures produced by K-Prototypes, which integrate both numerical and categorical dimensions, provide additional information that can improve the prediction of K-Medoids cluster membership when used as features in a CatBoost model.

Through this approach, the study aims to move beyond method comparison alone and demonstrate how mixed-data clustering and modern classification can be combined to

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (97-108)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49369

support more effective and realistic health risk stratification in pesantren settings.

2. METHODS

This study follows a structured and replicable analytical workflow. The research begins with the collection of new student data from the admissions administration system. The data are then preprocessed through a series of steps, including data integrity checking, cleaning, variable selection, and feature transformation. After the dataset is prepared, clustering is performed using two mixed-data algorithms, namely K-Prototypes and K-Medoids, to capture underlying health profile patterns. The resulting cluster information is then integrated as additional features in the classification stage using the CatBoost algorithm with hyperparameter optimization. The final stage involves evaluating model performance to assess the effectiveness of the proposed approach.

2.1. Data collection

The research data were obtained from the new student admission administration system of Queen Al Falah Islamic Boarding School for the 2025/2026 academic year. Data were collected through digital forms completed by prospective students and verified by the admission committee before being stored in the system. The initial dataset consisted of 1,464 student records with 44 variables. These variables included numerical attributes such as height, weight, number of siblings, and birth order, categorical attributes such as gender, registration channel, province of origin, parents' occupation, and medical history, as well as ordinal attributes such as parents' education and income levels.

From this dataset, the analysis focused on four health-related variables, namely height, weight, medical history, and current illness. These variables were selected because they are directly observable at the admission stage and provide immediate information regarding nutritional status and disease exposure. Other variables were excluded to reduce indirect socioeconomic confounding and to maintain a focused health profiling model. This choice introduces a known limitation. Age and gender may influence BMI interpretation, but they were not included as modeling variables. Nevertheless, age and gender were retained as reference information during data validation, particularly for outlier assessment.

All personal identifiers were removed prior to analysis. The dataset was anonymized, and access was restricted to the research team. Data collection was conducted as part of routine administrative procedures, with informed consent obtained from students or their guardians during the registration process.

2.2. Pre-processing

The pre-processing stage was conducted to ensure data quality and consistency before applying algorithm-based analysis. An initial data integrity check was performed to detect duplicated records that may have resulted from repeated form submissions or administrative synchronization issues. Duplicate entries were identified based on identical combinations of key attributes. As a result, two duplicated records were removed, producing a final dataset of 1,462 unique student records used in subsequent analyses.

In the medical history variable, missing values were observed to follow a non-random pattern. Students without prior health conditions tended to leave this field blank, while those with a medical history explicitly reported their conditions. Based on this pattern, blank entries were interpreted as an indication of no prior medical history. This approach follows the principle of context-based imputation for Missing Not at Random (MNAR) data [18]. It is acknowledged that this assumption may misclassify a small number of students who forgot or were unaware of their medical history.

Disease names in both medical history and current illness fields were standardized to eliminate spelling variations that referred to the same condition, such as "Magh," "Magg," and "Maag." After standardization, disease information was encoded using multi-hot or multi-label binary encoding. Each distinct disease label was transformed into a binary indicator, where a value of one indicates presence and zero indicates absence. This encoding allows multiple diseases to be represented simultaneously within a single student record. No frequency-based filtering was applied in the final encoding process.

Outliers in height and weight were handled using the Interquartile Range method. Values outside acceptable physiological ranges were examined individually. Measurements that were clearly implausible were removed, while extreme values that remained clinically plausible were retained. Although age and gender were not used as modeling variables, they were referenced during this validation process to reduce the risk of incorrect data removal.

2.3. Clustering

2.3.1. K-Prototypes

The first clustering stage employed the K-Prototypes algorithm, which is designed to handle datasets containing both numerical and categorical features. Euclidean distance was applied to numerical variables, specifically height and weight. Simple matching dissimilarity was used for categorical variables. Binary disease indicators obtained from multi-hot encoding were treated as categorical features in this algorithm.

Clustering was performed on the preprocessed dataset with the number of clusters evaluated in the range of 2 to 10. Multiple initializations were applied for each cluster configuration to improve solution stability. Internal validation indices were used to support subsequent cluster selection.

2.3.2. K-Medoids

The second clustering approach used the K-Medoids algorithm, also known as Partitioning Around Medoids. This method represents clusters using actual observations as medoids. The algorithm was implemented using Gower distance, which allows numerical and categorical features to be combined into a single dissimilarity measure. Binary disease indicators were treated as binary variables within the Gower distance computation.

To balance the contribution of different feature types, numerical and categorical feature groups were assigned equal aggregate weights. Each group contributed fifty percent to the overall distance calculation. The number of clusters was evaluated in the range of two to ten, and multiple medoid initializations were used to ensure robustness of the clustering results.

Table 1. *Parameters of clustering algorithms*

Parameter	K-Prototypes	K-Medoids
K	2 – 10	2 – 10
init	huang	-
n_init	5	5
random state	42	42
distance	mixed	gower
feature weight	-	50% numerical : 50% categorical

2.4. Feature Engineering

Feature engineering was performed by incorporating clustering results into the dataset that had undergone preprocessing. To prevent data leakage, the dataset was divided into training and testing subsets using an eighty to twenty split with stratified sampling. Clustering models were fitted exclusively on the training data. Test samples were assigned to clusters based on proximity to training-derived centroids for K-Prototypes or medoids for K-Medoids.

The resulting cluster labels were added as additional features. This allows each student to be represented not only by individual health attributes but also by structural information derived from clustering.

2.5. Classificaton

The classification stage aimed to build predictive models using original health features and clustering-derived features. The CatBoost algorithm was selected due to its ability to natively handle categorical features and its stable performance on tabular data. Two modeling scenarios were considered. The first was a baseline model without cluster features. The second was a hybrid model that incorporated clustering-based features.

Hyperparameter tuning was conducted using Grid Search on key parameters such as learning rate, tree depth, and number of iterations. The tuning process was combined with five-fold stratified cross-validation on the training set.

2.6. Evaluation

Model performance was evaluated using accuracy, precision, recall, and F1-score. In addition, the McNemar test was employed to assess whether differences in performance between baseline and hybrid models were statistically significant. This test is suitable for comparing paired classification models

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (97-108)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49369

evaluated on the same test dataset and provides an objective basis for performance comparison.

3. RESULTS AND DISCUSSION

3.1. Clustering result

This section presents the clustering results obtained using K-Prototypes and K-Medoids with K = 3. All interpretations of nutritional status follow the World Health Organization (WHO) BMI for age reference for children and adolescents aged 5–19 years. Nutritional categories are defined using age and sex adjusted z-scores, where thinness is below −2 SD, overweight is above +1 SD, and obesity is above +2 SD. Adult BMI cut-off values are not applicable to the study population.

Percentages reported for diseases represent within-cluster prevalence. "History" refers to previously diagnosed conditions reported by students (RP), while "current illness" refers to conditions reported at the time of admission (SD).

3.1.1. K-Prototypes

The K-Prototypes algorithm groups students based on a combination of anthropometric variables (height and weight) and categorical health information (medical history and current illness). Table 2 summarizes the main characteristics of each cluster.

Table 2. *Summary of k-prototypes clustering results*

Cluster	Size	Mean Height	Mean Weight	Dominant BMI for age Categories (%)	Most Frequently Conditions (%)
0	446	163.93	57.72	Normal (65.3), Overweight (26.4), Obesity (7.6)	Gastritis history (1.6), Typhoid history (1.6), Asthma history (1.1), Tonsillitis history (0.9), Current asthma (0.7)
1	620	155.46	46.50	Normal (78.7), Overweight (12.3), Thinness (5.3)	Gastritis history (3.1), Typhoid history (2.1), Acid reflux history (1.0), Asthma history (1.0), Tonsillitis history (0.5)
2	396	144.78	35.98	Normal (73.0), Thinness (9.6), Severe thinness (7.6)	Typhoid history (3.3), Gastritis history (2.0), Asthma history (1.3), Tonsillitis history (0.8), Shortness of breath history (0.8)

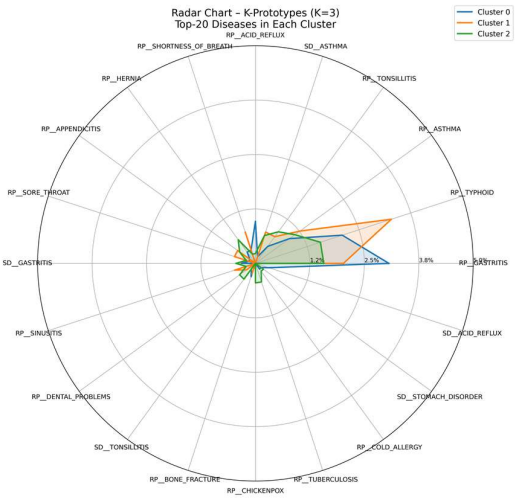

Figure 1. *Radar chart k-prototypes*

The three clusters primarily reflect differences in body-size profiles within the adolescent reference framework. All clusters are dominated by students with normal BMI-for-age. However, meaningful variation exists in the proportion of thinness, overweight, and obesity across clusters. Cluster 0 contains the highest proportion of students classified as overweight and obese, while Cluster 2 shows the highest combined proportion of thinness and severe thinness. Cluster 1 represents the most balanced nutritional profile.

Disease prevalence within clusters is generally low. Therefore, reported conditions should be interpreted as relative patterns rather than indicators of high disease burden. Across all clusters, typhoid, gastritis, and asthma consistently appear as the most frequently reported conditions, suggesting shared exposure patterns among newly enrolled students.

Figure 1 presents a radar chart visualizing the top 20 reported conditions across clusters. The radial axis represents prevalence in percent, with reference ticks shown to guide interpretation. Each line corresponds to one cluster, as indicated in the legend. The radar chart is intended for relative comparison, while

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (97-108)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49369

exact prevalence values are summarized in Table 2.

3.1.2. K-Medoids

The K-Medoids algorithm was applied using Gower distance to integrate numerical and categorical features. Unlike K-Prototypes, each cluster is represented by a medoid corresponding to an actual student record. Table 3 summarizes the clustering results.

Table 3. *Summary of k-medoids clustering results*

Cluster	Size	Mean Height	Mean Weight	Dominant BMI for age Categories (%)	Most Frequently Conditions (%)
0	469	146.28	36.55	Normal (74.0), Thinness (10.4), Severe thinness (7.9)	Typhoid history (3.20), Gastritis history (2.13), Asthma history (1.28), Current asthma (0.85), Tonsillitis history (0.85)
1	551	154.80	48.52	Normal (73.0), Overweight (17.6), Obesity (6.4)	Gastritis history (2.90), Typhoid history (2.00), Asthma history (0.54), Acid reflux history (0.54), Tonsillitis history (0.54)
2	442	165.01	56.43	Normal (72.2), Overweight (21.4), Obesity (4.1)	Gastritis history (1.81), Asthma history (1.58), Typhoid history (1.58), Hernia history (0.68), Tonsillitis history (0.68)

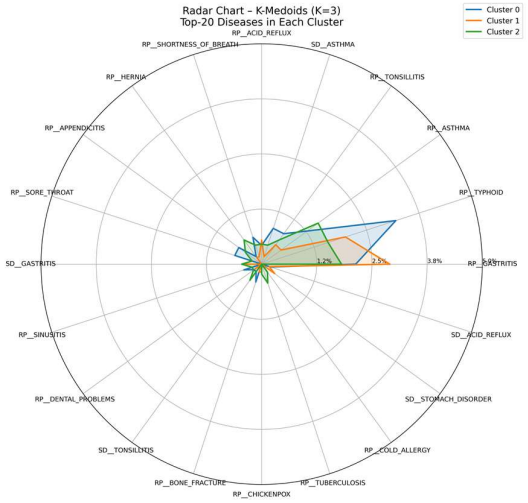

Figure 2. *Radar chart k-medoids*

The K-Medoids clusters exhibit a clear gradient of body-size profiles, from smaller to larger anthropometric characteristics. The close alignment between cluster means and medoid values indicates that the selected medoids are representative of their respective clusters. Similar to the K-Prototypes results, most students report no current illness or prior medical history. Consequently, cluster separation is driven primarily by anthropometric variation, with disease information contributing secondary differentiation.

Figure 2 presents the radar chart for K-Medoids using the same set of top 20 conditions and prevalence scale as Figure 1. This allows direct visual comparison between the two clustering approaches. As with the K-Prototypes radar chart, the figure highlights relative differences across clusters, while detailed values are provided in Table 3.

3.1.3. Comparative interpretation

Both clustering approaches consistently identify three groups aligned with nutritional heterogeneity among newly enrolled junior and senior high school students. K-Prototypes emphasizes differences in BMI-for-age distribution combined with categorical health information, while K-Medoids emphasizes anthropometric similarity through representative medoids.

Across both algorithms, typhoid, gastritis, and asthma emerge as the most frequently reported specific conditions. However, their absolute prevalences remain low, indicating that clustering primarily captures structural variation in nutritional status rather than pronounced disease stratification. Overall, the results demonstrate that mixed-data clustering can provide a meaningful framework for describing adolescent health profiles in a school-based screening context.

3.2. Silhouette coefficient evaluation

Cluster quality evaluation using Silhouette Coefficient shows K=2 produces the highest value on both algorithms (K-Prototypes: 0.443, K-Medoids: 0.444), but K=3 is selected as optimal with values 0.362 and 0.344 respectively. The selection of K=3 is based on ease of result interpretation and current research. The highest Silhouette value does not always produce the best solution in terms of practicality and result usefulness [19].

K=2 produces high cluster separation but excessive simplification eliminates important differences in students' health profiles. High Silhouette value does not guarantee meaningful clustering for further analysis [20]. K=3 produces profiles that can be clearly differentiated, with each cluster showing unique combination of nutritional status, disease types, and different needs that are more informative for decision making.

The distribution of K=3 clusters shows good balance (26.7-41.8 percent K-Medoids, 26.8-42.4 percent K-Prototypes), indicating stable clustering process and result consistency. Consistency in identifying diverse disease types across all clusters in both algorithms strengthens result validity. Both algorithms show the same pattern in identifying relationships between nutritional status and disease types, particularly high Typhoid in thin BMI groups, strengthening result credibility and ensuring K=3 is optimal for comprehensive health profile representation of the Islamic boarding school student population.

3.3. Classification

3.3.1. Catboost Model with Hyperparameter

Classification employs CatBoost to construct a predictive model that leverages basic health features of students and structural information derived from clustering outcomes. CatBoost is selected due to its capability to natively handle mixed-type data (numerical and categorical features) without requiring preprocessing encoding [21]. The dataset consists of 1,462 samples and is split into 80 percent for training (1,169 samples) and 20 percent for testing (293 samples). Stratified splitting is applied to preserve the original class distribution across train and test sets. To ensure a leak-free pipeline, the train–test split is performed before any feature transformation involving cluster labels.

Hyperparameter optimization is conducted using Grid Search with 5-fold stratified cross-validation. Model selection is based on F1-macro to ensure balanced performance across all cluster classes, avoiding bias toward dominant clusters. For the K-Prototypes target, the optimal CatBoost configuration is depth 8, learning rate 0.05, iterations 300, l2_leaf_reg 3, and border_count 64.For the K-Medoids target, the optimal configuration is depth 6, learning rate 0.01, iterations 300, l2_leaf_reg 1, and border_count 32.

3.3.2. Baseline model results

The baseline model predicts cluster labels using only the original health features, namely height, body weight, medical history, and current illness. No clustering-derived features are included at this stage. This model serves as a reference to evaluate whether additional cluster information provides meaningful predictive value.

For the K-Prototypes target, the baseline model achieves an accuracy of 98.98 percent with an F1-macro of 0.9899. Precision and recall are both above 0.99, indicating that the model is highly consistent across all classes. The overfitting gap between training and test F1 scores remains small, suggesting good generalization.

For the K-Medoids target, the baseline model achieves an accuracy of 99.66 percent and an F1-macro of 0.9967. Class-wise performance is well balanced, and only one misclassification is observed out of 293 test samples. These results confirm that both clustering algorithms produce stable and learnable structures that can be effectively predicted from basic health features alone.

3.3.3. Hybrid model results

The hybrid model adds cluster labels from one algorithm as additional features to predict cluster labels from another algorithm, implementing cross-algorithm validation strategy. This approach aligns with pseudo-labeling in semi-supervised learning where cluster membership serves as pseudo-labels for training supervised models [22]. Validation occurs implicitly through cross-predictability: high accuracy in predicting cluster algorithm B from cluster features of algorithm A provides empirical evidence that the structure of both clusters is robust and meaningful [23][24]. The hybrid model extends the baseline by incorporating cluster labels from the alternative clustering algorithm as additional input features. Specifically, K-Prototypes cluster labels are used to predict K-Medoids targets, and vice versa. To prevent label leakage, cluster labels are generated only from the training data and are then one-hot encoded separately for the training and test sets.

For the K-Prototypes target, adding K-Medoids cluster features results in a slight performance decrease. The hybrid model achieves an accuracy of 98.63 percent and an F1-macro of 0.9865, compared to 98.98 percent and 0.9899 in the baseline. Although the overfitting gap is marginally reduced, the overall classification performance does not improve. This indicates that K-Medoids cluster information does not provide additional discriminative power beyond what is already captured by the original health features and K-Prototypes structure.

In contrast, for the K-Medoids target, the hybrid model shows a consistent and meaningful improvement. Accuracy increases from 99.66 percent in the baseline to 100 percent in the hybrid model, with F1-macro, precision, and recall all reaching 1.0000. The number of misclassifications decreases from one to zero in the test set. This improvement suggests that K-Prototypes clusters capture broader health-related patterns that are informative for refining K-Medoids-based classification.

3.3.4. Comparison of baseline and hybrid

Table 4 summarizes the comparison between baseline and hybrid models for both clustering targets using the 80:20 split as the primary evaluation setting.

Table 4. *Comparison of classification performance between baseline and hybrid models*

Model	Accuracy	Precision	F1-macro	Overfitting Gap
Baseline Target K-Medoids	0.9966	0.997	0.9967	0.0080
Hybrid Target K-Medoids	1.0000	1.0000	1.0000	0.0057
(Improvement)	+0.0034	+0.0003	+0.0033	-0.0023
Baseline Target K-Prototypes	0.9898	0.9891	0.9899	0.0149
Hybrid Target K-Prototypes	0.9863	0.9851	0.9865	0.0117
(Change)	-0.0034	-0.0040	-0.0034	-0.0032

For the K-Medoids target, the hybrid model consistently outperforms the baseline across all evaluation metrics. Accuracy improves by 0.34 percentage points, and F1-macro increases from 0.9967 to 1.0000. Although the absolute numerical gain appears small, the complete elimination of misclassification errors represents a meaningful improvement in practical screening contexts.

To verify that the observed improvement for the K-Medoids hybrid model is not an artifact of a specific data split, an additional experiment using a 60:40 train–test ratio is conducted. The results confirm the same trend. The hybrid model again achieves perfect classification performance with accuracy, F1-macro, ARI, and NMI all equal to 1.0000, while the baseline model remains slightly below this level. This consistency across different split ratios supports the robustness of the observed hybrid advantage for the K-Medoids target.

For the K-Prototypes target, the hybrid model shows a small and consistent decrease across all metrics, with accuracy dropping by 0.34 percentage points and F1-macro decreasing by 0.34 percent. This symmetric decline across metrics indicates that the additional K-Medoids features introduce redundancy rather than useful complementary information.

An asymmetric pattern is therefore observed. K-Prototypes-derived features enhance the prediction of K-Medoids clusters, while K-Medoids-derived features do not improve and slightly degrade the prediction of K-Prototypes clusters. This pattern suggests an information hierarchy in which K-Prototypes captures broader and more holistic health structures, whereas K-Medoids focuses more narrowly on morphometric similarities that are already well represented by height and weight.

Overall, the classification results demonstrate that integrating K-Prototypes cluster information provides practical benefits when predicting K-Medoids clusters, while the reverse integration is not beneficial. Although statistical testing is required to assess formal significance, the consistent reduction of misclassification errors and the robustness across split ratios indicate that the hybrid strategy has clear applied value for health risk stratification in pesantren environments.

3.4. McNemar evaluation

The McNemar test evaluates statistical significance of differences between baseline and hybrid models by examining whether error rate differences between them are statistically significant or merely due to chance [25]. This

test is appropriate for comparing paired classifiers on an identical test set.

K-Medoids target demonstrates the following contingency table between baseline and hybrid predictions:

Table 5. *K-Medoids target McNemar test*

	Hybrid Correct	Hybrid Incorrect
Baseline Correct	292	0
Baseline Incorrect	1	0

For the K-Medoids target, the baseline model correctly classified 292 out of 293 test samples, while the hybrid model correctly classified all 293 samples. The contingency table shows one discordant case where the hybrid model corrected an error made by the baseline model, and no cases where the baseline model corrected an error made by the hybrid.

The McNemar test statistic for this comparison was also 1.00 with a p-value of 0.317. As with the K-Prototypes target, this result does not reach statistical significance at $\alpha = 0.05$. This outcome is expected given the very small number of discordant samples in the test set.

Despite the lack of statistical significance, the direction of the discordant pairs is informative. All discordant cases favor the hybrid model, indicating that the inclusion of K-Prototypes cluster features consistently improves error correction for the K-Medoids target.

K-Prototypes target demonstrates the following contingency table between baseline and hybrid predictions:

Table 6. *K-Prototypes target Mcnemar test*

	Hybrid Correct	Hybrid Incorrect
Baseline Correct	289	1
Baseline Incorrect	0	3

For the K-Prototypes target, the baseline model correctly classified 290 out of 293 test samples, while the hybrid model correctly classified 289 samples. The resulting contingency table shows one discordant case where the baseline model produced a correct prediction and the hybrid model produced an incorrect prediction, and no cases where the hybrid model corrected an error made by the baseline.

Based on this contingency table, the McNemar test statistic was 1.00 with a p-value of 0.317. This value exceeds the significance threshold of $\alpha = 0.05$, indicating that the observed performance difference between the baseline and hybrid models for the K-Prototypes target is not statistically significant.

These results are consistent with the classification metrics reported earlier, where the hybrid model slightly underperformed the baseline model. The absence of hybrid-only corrections confirms that adding K-Medoids cluster features does not provide additional discriminatory information for predicting K-Prototypes clusters.

Table 7. *McNemar test results for baseline vs hybrid models*

Algorithm	Baseline Acr	Hybrid Acr	Discordant_ b	Discordant_ c	McNemar Statistic	p-value	Significant
K-Prototypes	0.9898	0.9863	1	0	1.00	0.317	No
K-Medoids	0.9966	1.0000	0	1	1.00	0.317	No

Note: b denotes cases where the baseline model is correct and the hybrid model is incorrect, while c denotes cases where the hybrid model corrects errors made by the baseline.

The McNemar test results demonstrate that none of the observed performance differences between baseline and hybrid models reach statistical significance under conventional thresholds. This outcome is primarily driven by the extremely small number of misclassifications in both models, which limits the statistical power of the test.

However, from a practical perspective, the results reveal an asymmetric pattern that aligns with the classification analysis. For the K-Medoids target, the hybrid model reduces the absolute error count from one to zero, achieving perfect classification on the test set. In contrast, for the K-Prototypes target, the hybrid model introduces a small number of additional errors, confirming that K-Medoids features do not add useful information in this direction.

These findings support the interpretation that K-Prototypes captures broader health-related structure, making it informative for refining K-Medoids assignments, while K-Medoids primarily reflects morphometric information that is largely redundant with baseline features such as height and weight. Consequently, the hybrid approach is

practically beneficial only when predicting K-Medoids clusters, while the baseline model remains sufficient for K-Prototypes prediction.

CONCLUSION

This research compares the performance of K-Prototypes and K-Medoids in clustering health profile data of students at Queen Al Falah Islamic Boarding School, and integrates the clustering results as additional features in a CatBoost classification model. The results show that both clustering algorithms generate meaningful partitions, with high baseline classification accuracy of 98.98 percent for K-Prototypes targets and 99.66 percent for K-Medoids targets.

The hybrid model exhibits a clear asymmetric pattern. For the K-Medoids target, incorporating K-Prototypes cluster features improves accuracy to 100 percent, correcting one misclassified sample and reducing the error count from one to zero. In contrast, for the K-Prototypes target, adding K-Medoids features slightly decreases accuracy to 98.63 percent, with the number of misclassifications increasing from three to four samples. McNemar test results indicate that these differences are not statistically significant (p-value = 0.317), reflecting the very small number of discordant prediction pairs.

Despite the lack of statistical significance, the observed pattern remains practically relevant. The improvement on the K-Medoids target demonstrates that K-Prototypes captures broader health-related information that can support more precise partitioning, while K-Medoids, which focuses primarily on morphometric characteristics, does not provide incremental information for predicting K-Prototypes clusters.

The main contributions of this study are as follows: (1) empirical evidence of complementary behavior between K-Prototypes and K-Medoids in health-profile clustering, (2) demonstration that cross-algorithm validation can reveal an information hierarchy between clustering methods, (3) validation that a dual-algorithm strategy can support efficient health screening in pesantren environments with limited resources, and (4) achievement of high and stable classification performance suitable for practical deployment.

This study is limited by the use of data from a single pesantren and a single academic year, as well as by the restricted set of health variables used. Future research is encouraged to incorporate multi-site data, richer clinical indicators, alternative clustering approaches, and longitudinal analysis to further validate and extend the proposed framework.

REFERENCES

[1] I. Amalia *et al.*, "Combating Infectious Diseases Threat among Students in Islamic Boarding School (Pondok Pesantren): A Pilot Assessment," *J. Community Empower. Heal.*, vol. 6, no. 1, p. 7, Apr. 2023, doi: 10.22146/jcoemph.77426.

[2] E. Rianti, A. Triwinarto, A. Rodoni, and Elina, "Enhancing Health Quality of Islamic Boarding School Students through Hygiene Practices in Depok and Banten, Indonesia," *Indian J. Forensic Med. Toxicol.*, vol. 13, no. 4, p. 1661, 2019, doi: 10.5958/0973-9130.2019.00545.0.

[3] F. H. Ruslana and S. Mulyono, "The Relationship of Cultural Values with Clean and Healthy Life Behaviour among Islamic Boarding School Students in Indonesia," *J. Public health Res.*, vol. 11, no. 2, Apr. 2022, doi: 10.4081/jphr.2021.2739.

[4] J. Olufemi Ogunleye, "The Concept of Data Mining," 2022, pp. 1–34. doi: 10.5772/intechopen.99417.

[5] Venkata Mahesh Babu Batta, "Machine Learning," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 583–591, Apr. 2024, doi: 10.48175/IJARSCT-17677.

[6] D. Zelterman, "Clustering Methods," 2022, pp. 305–351. doi: 10.1007/978-3-031-13005-2_11.

[7] A. Majumder, "Classification Models in Machine Learning Techniques," 2023, pp. 1–16. doi: 10.4018/978-1-6684-8531-6.ch001.

[8] S. Yadav, "Heart Disease Prediction Using Machine Learning," *INTERANTIONAL J. Sci. Res. Eng. Manag.*, vol. 08, no. 07, pp. 1–14, Jul. 2024, doi: 10.55041/IJSREM36858.

[9] A. Pathak *et al.*, "Application of Machine Learning K-Means Clustering

and Linear Regression in Determining the Risk Level of Pulmonary Tuberculosis," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, IEEE, Sep. 2024, pp. 1–6. doi: 10.1109/COMPAS60761.2024.10796963.

[10] H. Hafid and S. Annisa, "IMPLEMENTATION OF K-MEDOIDS AND K-PROTOTYPES CLUSTERING FOR EARLY DETECTION OF HYPERTENSION DISEASE," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 19, no. 1, pp. 465–476, Jan. 2025, doi: 10.30598/barekengvol19iss1pp465-476.

[11] A Priyanka and D. C. Chandrasekar, "Efficient Slice Creation in Network Slicing using K-Prototype Clustering and Context-Aware Slice Selection for Service Provisioning," *Int. J. Recent Technol. Eng.*, vol. 12, no. 5, pp. 12–20, Jan. 2024, doi: 10.35940/ijrte.E7973.12050124.

[12] H. Jridi, M. A. Ben HajKacem, and N. Essoussi, "Parallel K-Prototypes Clustering with High Efficiency and Accuracy," 2020, pp. 380–395. doi: 10.1007/978-3-030-59065-9_29.

[13] R. Septian and D. Darnah, "Penerapan Algoritma K-Medoids pada Pengelompokan Wilayah Provinsi di Indonesia Berdasarkan Indikator Pendidikan," *EKSPONENSIAL*, vol. 14, no. 2, p. 85, Nov. 2023, doi: 10.30872/eksponensial.v14i2.1150.

[14] L. Lenssen, E. Schubert, A. Krivošija, E. Schubert, A. Lang, and S. Hess, "Cluster Analysis," in *Fundamentals*, De Gruyter, 2022, pp. 179–248. doi: 10.1515/9783110785944-005.

[15] A. AbdElSamea and S. M. Saif, "K-medoid clustering containerized allocation algorithm for cloud computing environment," *J. Electr. Syst. Inf. Technol.*, vol. 11, no. 1, p. 35, Sep. 2024, doi: 10.1186/s43067-024-00161-1.

[16] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[17] Y. H. Chang *et al.*, "Machine learning–based triage to identify low-severity patients with a short discharge length of stay in emergency department," *BMC Emerg. Med.*, vol. 22, no. 1, pp. 1–10, 2022, doi: 10.1186/s12873-022-00632-6.

[18] T. Phung, K. Reese, I. Shpitser, and R. Bhattacharya, "Recursive Equations For Imputation Of Missing Not At Random Data With Sparse Pattern Support," Jul. 2025, [Online]. Available: http://arxiv.org/abs/2507.16107

[19] H. Al Azies, F. A. Rohmatullah, H. B. Rochmanto, and D. Putri, "TOWARDS OPTIMIZATION : DATA-DRIVEN APPROACH K-MEDOIDS CLUSTERING ALGORITHM FOR REGIONAL EDUCATION QUALITY," vol. 12, no. 3, 2022.

[20] A. F. Purba, Mustafid, and K. Puspita, "PENERAPAN ALGORITMA k-PROTOTYPE UNTUK PENGELOMPOKAN DESA DI KABUPATEN BEKASI BERDASARKAN INFRASTRUKTUR DIGITAL," vol. 13, pp. 479–489, 2025, doi: 10.14710/j.gauss.13.2.479-489.

[21] A. B. Mawardi, R. S. Pradini, M. S. Haris, and G. Boosting, "Komparasi Algoritma Boosting untuk Prediksi Gangguan Tidur," vol. 13, no. 3, doi: https://doi.org/10.23960/jitet.v13i3.7281.

[22] E. Arazo, D. Ortego, P. Albert, N. E. O. Connor, and K. Mcguinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning".

[23] J. Xu, T. Li, D. Zhang, and J. Wu, "Ensemble clustering via fusing global and local structure information," *Expert Syst. Appl.*, vol. 237, p. 121557, Mar. 2024, doi: 10.1016/j.eswa.2023.121557.

[24] H. Surbakti and T. A. Munandar, "K-Means-Based Pseudo-Labeling Technique in Supervised Learning Models for Regional Classification Based on Types of Non-Communicable Diseases," *J. Online Inform.*, vol. 10, no. 2, pp. 465–473, Nov. 2025, doi: 10.15575/join.v10i2.1609.

[25] A. Garrocho-Rangel, S. Aranda-Romo,

R. Martínez-Martínez, V. Zavala-Alonso, J. C. Flores-Arriaga, and A. Pozos-Guillén, "Fundamentals of Nonparametric Statistical Tests for Dental Clinical Research," *Dent. J.*, vol. 12, no. 10, p. 314, Sep. 2024, doi: 10.3390/dj12100314.