Journal of
Information Systems Engineering
and Business Intelligence

Vol.11, No.3, October 2025
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Optimizing Tuition Fee Determination with K-Means Cluster Relabeling Based on Centroid Mapping of Principal Component Pattern

Wiyli Yustanti [1)*], Andi Iwan Nurhidayat [2)], Muhammad Iskandar Java [3)]

[1)]*Department of Information System, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia*
[1)]wiyliyustanti@unesa.ac.id

[2)]*Department of Informatic Management, Faculty of Vocaion, Universitas Negeri Surabaya, Surabaya, Indonesia*
[2)]andyl34k5@unesa.ac.id

[3)]*Information Technology Development Center, Universitas Negeri Surabaya, Surabaya, Indonesia*
[3)]muhammadjava@unesa.ac.id

Abstract

Background: Tuition fee in Indonesian public universities is determined based on the socioeconomic status of prospective students. In this context, students are assigned to tuition fee groups after passing the selection process through achievement-based or computer-based exams. However, the current grouping system shows overlapping distributions, indicating the need for a more precise classification method.
Objective: This research aims to improve the accuracy of tuition fee group assignments by refining the clustering structure and relabeling the classification dataset.
Methods: A total of 13 socioeconomic variables were used to predict tuition fee groups. This research used K-Means clustering algorithm and a relabeling process using centroid mapping of principal components to balance original and newly generated labels. To assess the effectiveness of the relabeling process, six classification algorithms, namely Decision Tree (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), were used. Statistical tests at a 5% significance level were conducted to evaluate improvements in classification accuracy.
Results: The relabeling process significantly enhanced prediction accuracy compared to the original dataset. The refined clustering structure reported better classification performance across all six algorithms, showing the effectiveness of the proposed method.
Conclusion: The results showed that robust clustering and a relabeling method improved the precision of tuition fee classification systems. The proposed framework provided a data-driven solution for refining classification models, ensuring a fairer distribution of tuition fee based on socioeconomic indicators. The novelty lies in the centroid-based relabeling, which uses principal component patterns to enhance interpretability and classification accuracy. The method was adaptable for global use in any educational system using socioeconomic-based fee classification. Future research should explore alternative clustering methods and additional socioeconomic factors to enhance classification accuracy.

Keywords: *K-Means Clustering, Machine Learning, Relabeling Process, Socioeconomic Indicators, Tuition Fee Classification*

Article history: Received 25 February 2025, first decision 22 July 2025, accepted 6 October 2025, available online 28 October 2025

I. INTRODUCTION

The issue of financial burden in accessing higher education is a global challenge since many countries ensure equitable access and maintain financial sustainability of institutions. Different methods, such as income-based tuition, fee-tiering, and needs-based subsidies, have been implemented to address these challenges. In Indonesia, financial burden associated with higher education is a central concern, prompting the implementation of Single Tuition Fee (STF) policy across Public Universities (PUs). Therefore, tuition fee remains affordable, considering the varying financial capacities of students and families. STF policy was formally introduced in 2013 by the Ministry of Education and Culture. This policy was consistent with Article 88 of Law No. 12 of 2012 on Higher Education, where operational costs for students were standardized and adjusted based on economic capacity. The policy was partly driven by the

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print) © 2025 The Authors. Published by Universitas Airlangga.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
doi: http://dx.doi.org/10.20473/jisebi.11.3.445-458

increasing burden of non-tuition-related fees. Furthermore, STF rates are implemented in phases and categorized into several tiers. In this context, students from economically disadvantaged backgrounds pay lower tuition rates, while those with greater financial means are assigned higher rates. According to the Ministry of Education Regulation No. 2 of 2024, STF applies to all academic levels, from Diploma to Doctoral programs, in Public Universities. However, specific provisions require that STF rates must be divided into two groups, with the lowest tiers designated for students from economically disadvantaged backgrounds. Additional higher-level groups may be established by universities, with fees adjusted up to the maximum standard operational cost of each program. In contrast to other countries adopting income-based or needs-based tuition models, Indonesia implements STF through a centralized national policy. The actual application varies across public universities and often lacks a standardized framework for socioeconomic assessment. This inconsistency creates challenges in categorizing students fairly and accurately to explore data-driven methods in evaluating and improving existing groupings. These classification tasks belong to unsupervised learning from a scientific perspective. Theoretically, a good cluster has high homogeneity (similarity) and heterogeneity (difference) [1].

In a good structure, STF group predictions for new students are expected to be more accurate [2]. All state universities have adopted a variety of socioeconomic indicators to identify STF category. However, previous research relied solely on university-derived STF groupings as ground truth without validating internal consistency or cluster quality. The quality of STF clusters was assessed by using validity indices and relabeling before classification. This research is among the first to incorporate a cluster validation phase before relabeling and classification, ensuring more reliable predictions based on structurally valid STF groupings. Previous research [3] identified variables impacting students' socioeconomic circumstances, and 16 of the 43 initial variables substantially affected the calculation of students' STF. The group will use a feature selection method based on Support Vector Machine (SVM) algorithm [4], and has an accuracy rate of 81% with F1-Score metric. Several research developed a decision-making system based on the fuzzy-c-means algorithm [5], [6], [7]. A decision tree-based algorithm for determining STF group is also applied through comparison experiments with the J48, ID3, and Naive Bayes (NB) algorithms. This research [8] concluded that J48 had better accuracy than ID3 and NB algorithms, with a value using F1-Score of 91.1%. In line with the efforts, recent research [9], [10], [11], [12] has emphasized the importance of relabeling and clustering methods in improving classification accuracy. The integration of clustering validity measures is supported before classification.

Based on the description above, this research aims to address the problem of unreliable STF group labels by introducing a systematic validation and relabeling framework before classification. Before conducting supervised learning-based modelling, the results neglect the conditions of variety inside and across clusters. Previous research reported that the outcomes of grouping with the university formulations were sufficient for use as training and testing materials to forecast STF level of new students. Therefore, this research aims to (1) assess the quality of STF grouping results produced by the university formula using cluster validity metrics, (2) implement STF grouping labelling process based on the cluster goodness index, and (3) conduct experiments for classification algorithm-based cluster relabeling to compare the accuracy before and after the process. The results are expected to contribute to (1) improving the structure of the dataset used for predicting new student STF and (2) providing recommendations for a classification algorithm with the best performance.

The main contributions and novelty of this research are threefold. First, a validation process is introduced using clustering validity indices to assess quality. Second, a relabeling method is proposed to improve label reliability before classification, which addresses potential inconsistencies in the original data. Third, the impact of the relabeling method is evaluated through classification experiments using multiple algorithms. In Indonesian higher education context, this is among the first research to incorporate a systematic validation and relabeling framework in STF classification process.

II. METHODS

The research framework adopted Knowledge Data Discovery (KDD) process in data mining, which was proposed by the knowledge discovery community [13]. KDD process consisted of five phases, namely (1) Selection, (2) Pre-processing, (3) Transformation, (4) Data Mining, and (5) Evaluation. The general phases in KDD procedure started from selection to obtain relevant data from the source, and continued with the pre-processing phase. The next phase was transformation, which included converting the data into a format suitable for analysis. This was followed by the data mining phase, where analytical methods were applied to identify patterns or models. Finally, the interpretation or evaluation phase focused on assessing and interpreting the results. Fig. 1 shows the overview of this process, including the modifications applied. In this research, procedure modifications were carried out in the transformation and modelling phases. During the transformation process, the raw data used as input in the clustering method were standardized and dimensionally reduced with principal component analysis. The value of the main component was

used as input in the clustering method. The results produced a random label from index 0 to k. For the cluster to be meaningful, pattern matching of the origin and random labels was carried out using a component analysis with a loading factor of more than 0.5 [14]. The following explanation provided a more detailed understanding of the phases in KDD process.

1) Data Selection

Table 1 shows the distribution of STF dataset from 15875 students at Universitas Negeri Surabaya with categories 1 to 8 over 5 years (2017–2021). The lowest and highest categories (1st and 8th) have the smallest and largest tuition fees, respectively. Tuition fee is the amount of student payment per semester applied consistently in the research period. This STF scheme is applied to undergraduate programs as regulated by the Ministry of Education. Tuition fee ranges from IDR 500,000 to IDR 7,500,000 in Categories 1 to 8, respectively.

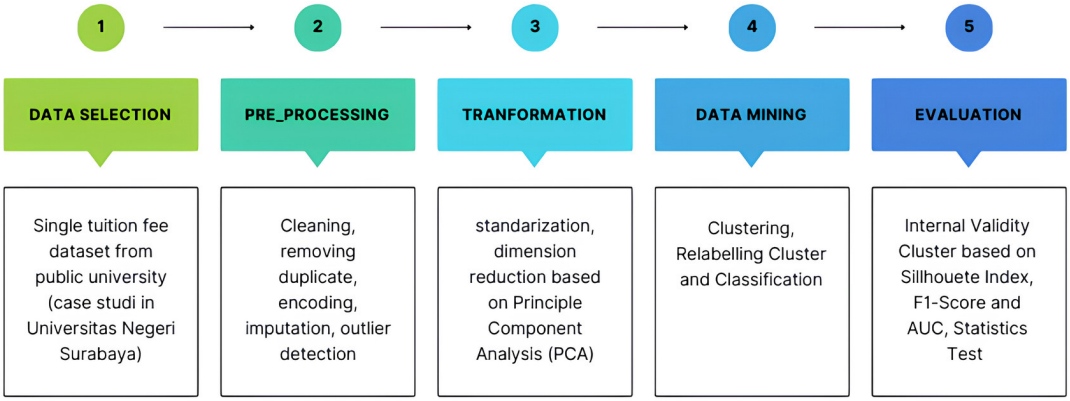

Fig. 1 Research Framework based on KDD Process.

The categorization is based on socio-economic criteria, including household income, electricity usage, parental occupation, and other verifiable indicators submitted during the registration process. The final STF category is determined by the university after evaluating the documents and data submitted. Furthermore, the class value for the tuition category group will be the label or target variable in the classification modelling. Table 1 shows the unequal distribution of the student population across STF categories, with high and low concentration in Categories 4 to 6 and 1 and 8, respectively. This class imbalance may affect the performance of classification models, which tend to be biased toward the majority classes. Therefore, addressing the imbalance through optimized relabelling and modelling strategies is a key focus of this research.

TABLE 1
THE NUMBER OF STUDENTS DISTRIBUTED IN STF CATEGORY

Year	Single Tuition Fee (STF) Category							
	1	2	3	4	5	6	7	8
2017	20	103	294	1266	1367	36	0	0
2018	21	126	298	1481	1557	44	0	0
2019	54	240	306	830	999	501	48	0
2020	30	211	86	1112	964	355	35	1
2021	36	72	311	1122	1201	709	39	0
Total	161	752	1295	5811	6088	1645	122	1
							Total	15875

The variables used reflect the socio-economic characteristics of students' parents. The target variable was tuition fee group (Y), and the predictors X_1 to X_{13} are socio-economic indicators seen in Table 2.

2) Pre-Processing

Pre-processing process is an important phase before implementing Principal Component Analysis (PCA) to ensure that the data provides optimal results. This phase consists of three main phases, namely (1) data cleaning to address missing or inconsistent values, (2) normalization to ensure all variables are on the same scale, and (3) outlier detection to minimize distortion in PCA results. The encoding process is required to convert categorical data into a numerical format [15]. After the data is clean and standardized, the next phase is to evaluate the correlation between the variables. This ensures that PCA can effectively reduce the data dimensions and retain relevant information. A meticulous pre-processing process improves the quality of PCA analysis and interpretation of results.

TABLE 2
THE RESEARCH VARIABLES

Variable	Value	Type	Description
Y	1,2,3,4,5,6,7,8	Ordinal	Single Tuition Fee Level
X1	1,2,3,4,5,6	Ordinal	Father's Occupation
X2	1,2,3,4,5,6	Ordinal	Mother's Occupation
X3	1,2,3,4,5	Ordinal	Father's Salary
X4	1,2,3,4,5	Ordinal	Mother's Salary
X5	1,2,3,4,5	Ordinal	Number of Dependent
X6	1,2,3,4,5	Ordinal	Building Land Tax Billing
X7	1,5,6	Ordinal	Number of Cars
X8	1,4,5,6	Ordinal	Number of Motorbikes
X9	1,2	Nominal	Electricity Source
X10	1,2,3	Nominal	Residential Status
X11	1,2,3,4,5	Ordinal	Electricity Billing
X12	1,2,3,4	Ordinal	Mother's other income
X13	1,2,3,4	Ordinal	Father's other income

3) Transformation

Principal component analysis is a widely used strategy for dimensionality reduction. The central concept generates new features and projects the original data onto these features to maximize the total variation in the data. The processes for calculating the principal components are as follows: (a) Standardize variables. (b) Calculating the Covariance Matrix to determine Correlation. (c) Determine the Principal Components by calculating the Eigenvalues and Eigenvectors of the Covariance Matrix. (d) Determine which Principal Components to retain for further research based on differences in Components utilizing a Scree Plot. (e) Transform data along the Principal Components axis.

PCA is defined as an orthogonal linear transformation that converts the data to a new coordinate system. The highest and second largest variance by some scalar projection of the data are on the first and second coordinates, respectively [16].

4) Data Mining

In this phase, three main activities are carried out in data mining modelling. These are (1) clustering using K-Means algorithm, (2) relabeling the results using principal component loading factors, and (3) classification modelling to evaluate prediction performance. The first phase is clustering through K-Means algorithm due to the simplicity and efficiency in grouping data with ordinal and nominal characteristics. K-Means can effectively identify socioeconomic patterns based on employment, income, and family assets, most of which represent ordinal scales. The algorithm is computationally efficient, suitable for medium-sized datasets, and capable of providing meaningful insights to support data-driven decision-making, such as STF level determination. The relatively simple data structure and controlled number of dimensions allow K-Means to provide informative and relevant cluster results [17]. The second phase is the relabeling process on the result label. Computationally, cluster labels are randomly assigned to each group formed from index 0 to k. K-Means algorithm is given a value of k=8, based on the number of STF groups desired by the university. Furthermore, this research uses the loading factor value of the variable on the main components to interpret the group related to the socioeconomic level of students. The proposed algorithm for the relabeling method is shown in Algorithm 1. The third phase includes applying six classification algorithms and three validation scenarios to evaluate the prediction of STF groupings. The proposed algorithm for the relabeling method is shown in Table 3. A new dataset is formed after relabeling and used to predict STF groups through the classification method. In this context, a total of 6 classification algorithms and 3 validation scenarios are compared. The dataset has better prediction accuracy performance than before relabeling.

5) K-Means Algorithm

K-Means is selected due to wide application and effectiveness in unsupervised partitioning of data with underlying continuous attributes, including socioeconomic variables related to financial profiles, as reported in previous research [9], [18]. K-Means is a vector quantization method known as clustering to divide a set of n observations into k clusters. The procedure of K-Means algorithm is as follows [19]: (a) Determine a random cluster centre and the number of clusters (k). (b) Determine the distance from each data point to the cluster's center. (c) Assign the data to clusters with the least distance. (d) Determine the cluster center. (e) Keep going back and forth between phases 2 and 4 until no more data is moving to the other clusters. (f) Cluster Relabeling.

In text grouping analysis, cluster labelling issues are frequently experienced. The clarity labelling requires the selection of a human-readable descriptive label for the cluster produced by the document method. In this context, ordinary clustering algorithms do not produce the labels. The cluster labelling algorithm analyzes the document's contents to identify labels. Research related to cluster labelling can be found in [20] and [21], where the method

examines interconnection paths between crucial locations separating various cluster contours and analyzes the topology of the function representing Support Vector Clustering (SVC) cluster outlines. Distinct clusters are identified and connected to the appropriate cluster. Therefore, this research is connected to the results of [22] concerning the labelling of SVC. In another research [23], a mechanism known as Maximal Resemblance Data Labeling (MARDL) is used. Each unlabeled data point is assigned to the correct cluster based on a new categorical grouping called N-Nodeset Importance Representative (NNIR). NNIR represents the cluster with the importance of attribute value combinations. The cluster labelling method aims to improve the structure that can be differentiated between clusters. Therefore, the labelling must be connected to the original data label because the data comes from poor results and is geared towards better clustering. In the algorithm, the centroids of the actual labels are mapped to K-Means. Selection is based on the closest distance between the original centroid and the cluster to prevent double mapping. The label is updated by replacing the original to match the mapping results of the updated prediction.

Algorithm 1

Relabelling Cluster

```
function relabelling_cluster()
    # Data Reading and Preprocessing
    original_data ← read_data_from_database()
    selected_features ← perform_feature_selection(original_data)
    pca_features ← apply_pca(selected_features)
    reduced_features ← select_first_two_principal_components(pca_features)

    # Calculate Real Centroids
    real_centroids ← calculate_centroids(reduced_features)

    # Clustering with K-Means
    cluster_centroids ← apply_kmeans_clustering(reduced_features)

    # Centroid Mapping
    cluster_order ← initialize_empty_list()
    for each real_centroid in real_centroids
        closest_index ← find_nearest(real_centroid, cluster_centroids, cluster_order)
        append_to_cluster_order(cluster_order, closest_index)
    end for

    # Update Predicted Labels
    label_mapping ← create_label_mapping(real_centroids, cluster_order)
    updated_labels ← update_labels_with_mapping(label_mapping)

    # Handle Imbalanced Data
    x_train, y_train ← oversample_data(reduced_features, updated_labels)

    # Train the SVM Model
    clf ← train_svm_model(x_train, y_train)

    # Save Model and Results
    save_trained_model(clf)
    return updated_labels, clf
end function
```

6) Classification-Based Cluster Relabeling

The labeled input data is required for modeling to predict a group from a set of items. Research [24], [25], [26], [27], [28], and [29] contains previous investigations reporting the labeling outcomes used in the classification modeling procedure. Cluster-based data relabeling (CBDR) is a recent strategy introduced in the latest research [30] that enables linear classifiers to operate successfully on nonlinear data. The concept divides the data set into multiple class-specific clusters without overlapping and relabelling. A linear classifier can be used on the relabeled data to

obtain cluster-based linear decision boundaries. Extensive trials have shown that CBDR significantly improves the performance of linear classifiers and outperforms the nonlinear counterparts. According to additional experiments, CBDR has also increased the classification performance of nonlinear classifiers. The most substantial outperformance was observed with skewed data. Another research used the class decomposition method [31] and [32] investigated the hierarchical and K-Means clustering methods. The results showed that class breakdown using K-Means and hierarchical clustering enhanced F1-score of NB classifier. An experimental design is also carried out using a relabeling dataset based on PCA component to accomplish the third goal. There is an element m_{pqr}, where p is the type of dataset used, p=1,2, and q is the fold level in the cross-validation process. In this context, q=3,5,10, while r is the type of classification algorithm (r = 1,2,3,4,5,6). Furthermore, m is a metric used to measure the goodness of classification predictions in the form of F1-Score or AUC (Area Under the Curve) values. The classification algorithm used is Decision Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), NB, Random Forest (RF), and SVM. A substantial difference in the accuracy of the prediction results can be produced between the data with the original and new labels based on the proposed method. The experiment used Jupiter-lab on a computer with an Intel Core i7 6th-generation processor and 64 GB of RAM. In hyperparameter optimization, the Grid Search method is used with the following parameter ranges in Table 3.

TABLE 3
RANGE VALUE OF HYPERPARAMETER TUNING

Algorithm	Parameter		Range Value
DT	a.	criterion	['gini', 'entropy']
	b.	min_samples_leaf	[1, 10, 100, 500]
	c.	min_samples_split	[1, 10, 100, 500]
KNN	a.	n_neighbors	[1, 10, 100, 500]
	b.	weights	['uniform', 'distance']
	c.	metric	['euclidean', 'manhattan']
LR	a.	penalty	[‘l1’,’l2’]
	b.	C	[0.1, 1, 10,100]
	c.	F1	[1.0, 0.5, 0.1]
	d.	solver	['liblinear']
NB	a.	prior	None
	b.	smoothing	np.logspace(0,-9, num=100)
RF	a.	min_samples_leaf	[1, 10, 100, 500]
	b.	max_depth	[1, 10, 100, 500]
	c.	min_samples_split	[1, 10, 100, 500]
SVM	a.	kernel	[‘linear’,’rbf’]
	b.	C	[1, 0.1, 0.01]
	c.	Gamma	[0.1, 1, 10,100]
	d.	Decision function shape	['ovo', 'ovr']

7) Evaluation and Validation

This research uses the Silhouette Index, widely applied in validating the structure of clusters in unsupervised learning tasks to assess the quality of clustering before relabeling and classification [33]. The Silhouette value measures the fitting level of a data point in the assigned cluster. The Silhouette score *s(i)* ranges from -1 to +1, with higher values since the object is well and poorly matched to the main and neighboring clusters, respectively. A clustering configuration is considered good when most data points have high silhouette values. Conversely, increased low or negative values indicate poor clustering or an inappropriate number of clusters. The Silhouette Index is computed using various distance measures, such as Euclidean or Manhattan distances. In Equation (1), point i belongs to cluster C_I, and the value *a(i)* is defined as the average distance from *i* in the same cluster.

$$a(i) = \frac{1}{|C_I|-1}\sum_{j\in C_I, i\neq j} d(i,j) \quad (1)$$

Where *d(i,j)* is the distance between the *i-th* and *j-th* point in the same cluster (C_I), and $|C_I|$ is the number of points in the C_I. The smaller value of *a(i)* can be interpreted as a better cluster assignment. In contrast, *b(i)* is the average distance between a point I and every other point outside the C_I cluster determined by Equation (2):

$$b(i) = \min_{J\neq I} \frac{1}{|C_I|}\sum_{j\in C_I, i\neq j} d(i,j) \quad (2)$$

With equations (1) and (2), form the *s(i)* equation (3),

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}, if|C_I| > 1 \quad (3)$$

Cross-validation is an effective evaluation method, specifically on multiclass data with an imbalanced distribution. The advantage is the ability to provide more stable and accurate model performance estimates than single evaluation methods, such as train-test splits. Cross-validation ensures that each point is used to make the evaluation results more representative of the overall dataset by dividing the data into multiple subsets [34]. In imbalanced data, the use of metrics such as F1-score becomes relevant in combining precision and recall. Cross-validation ensures that the resulting F1-score does not rely on a data share but consistently reflects the performance across multiple scenarios to provide a more reliable assessment [35], [36]

III. RESULTS

Based on the method, the results of initial data exploration are significant in visualizing data structures based on STF groups. Dimensional reduction uses PCA to obtain visualization results in a two-dimensional graph. During the pre-processing phase, the original data were transformed using ordinal encoding and validated for feasibility through the Kaiser-Meyer-Olkin (KMO) and Bartlett statistical test to determine the viability of PCA. Bartlett's Test of Sphericity aims to analyze the hypothesis that the variables are uncorrelated in the population (H0). The null hypothesis is rejected when the p-value is small or less than the significance level (5%). The calculation results of the p-value for the Bartlett Test are 0.000 since H_0 is rejected. The KMO is a measure of Sampling Adequacy (MSA), an index used to assess the accuracy of PCA. In this context, PCA is not advised when MSA is less than 0.5. MSA value for the dataset is 0.5818, greater than 0.5. The application of PCA is expected to reduce dimension and extract relevant components. Based on the eigenvalues in Fig.2, the predictor variable from the dataset can be reduced to 5 components (eigenvalue >1). The data is standardized before PCA process is carried out. Furthermore, the five new variables are used as a new dataset for the clustering process. The total variation represented by the extracted components is 63.17%.

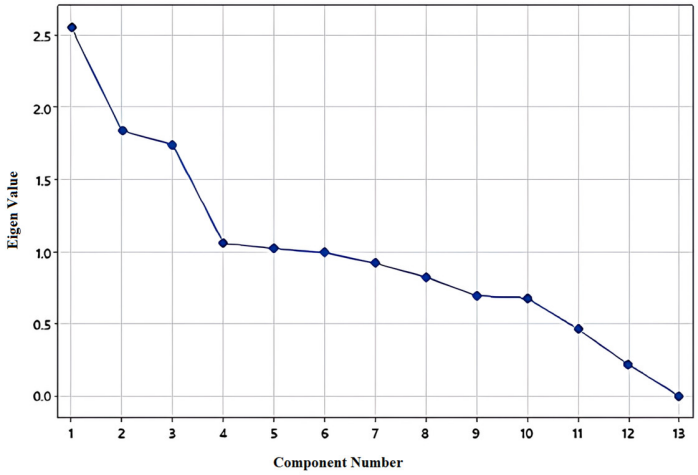

Fig. 2. Scree Plot for Dimension Reduction

The result of dimension reduction using PCA can be used to create visualizations between the first and second components based on STF-level labels in the original dataset. Fig. 3a shows that the clusters between STF are not separated, and overlap occurs. A value of -0.054 is obtained when measured using the Silhouette Index. Since the value of the cluster goodness is close to -1, re-clustering should be conducted to increase the value of the Silhouette Index. A new data cluster form is generated using K-Means algorithm, as shown in Fig. 3b. The location of objects between clusters in Fig. 3b is clearly separated. The label given by K-Means is a random number with no meaning as an STF level. Therefore, a strategy must be developed to map the random label given by K-Means with the original label. A value of 0.591(close to 1) is obtained when measured using the Silhouette Index in the right cluster. The number of clusters (k) used in K-Means algorithm is eight. This is because there are eight tuition fee categories applied in Universitas Negeri Surabaya. In Fig.3b, the dots on the plot represent students grouped by STF class (1 to 8). This visualization enables clearer identification of student distributions based on socioeconomic levels, after dimensionality reduction and re-clustering. X-axis and Y-axis represent the 1st and 2nd Components, which are a linear combination of native features designed to capture variation in data. The first component captures the most dominant variance in the dataset, assumed to reflect underlying socioeconomic status. Students with lower STF are located on the left side and may have certain centralized and relatively uniform socioeconomic traits. In contrast, students with high STF are

located at the top or right side, showing more significant differences in socioeconomic profiles. Clusters with low STF appear to be tighter, suggesting similar features between the groups. Furthermore, students in the category of high STF may have more diverse socioeconomic backgrounds. This scatter plot indicates overlap or mixing between certain groups in areas with overlapping colors and points. In the area around the center, several dots of similar or spatially adjacent colors appeared. This shows that the distinguishing features of the groups may be less obvious. Even though there is a tendency for more separate distributions, some points are not entirely in the dominant cluster. Therefore, there are students in the category with a socioeconomic profile close to a lower STF level. In addition, the unbroken color gradient from dark to light indicates that the transition between STF levels is not fully explicitly classified. Some STF categories at the intermediate level have similar characteristics in the two-dimensional space of PCA results. The scatter plot reports the potential for overlap between several STF levels, specifically at the middle and high levels. Therefore, the clustering method should be analyzed using K-Means to make the boundaries between STF groups more precise and separable. Fig. 3b is the scatter plot of K-Means clustering result with the number of 8 clusters.

Fig. 3 (a) Scatter Plot of Original Dataset based on PCA (b) Scatter Plot of K-Means

Fig. 3b shows the scatter results of K-Means clustering, which is a visual improvement from the initial plot. This clustering shows that K-Means algorithm successfully separates the data into several groups based on similarities in PCA space. Areas that previously showed overlap between groups, such as clusters at intermediate STF levels, are more clearly separated. K-Means manages to group similar points into a more centralized region without significant overlaps. Some clusters appear relatively isolated from others, showing the unique characteristics of the algorithm. The clusters in the middle are denser and tighter, indicating a substantial uniformity in the data. In contrast, the clusters in the top right and bottom right are scattered, suggesting that the data has higher variability. Since the cluster labels do not have meaning as STF levels, the next phase is to use the proposed relabeling algorithm. In principle, the center point of the clustering labeling is compared with the original STF label. This visualization provides a stronger foundation for evaluating the accuracy of K-Means grouping. The clustering shows that the data can be separated more systematically than the initial scatter. However, the random labels need validation or interpretation to have meaning at STF level.

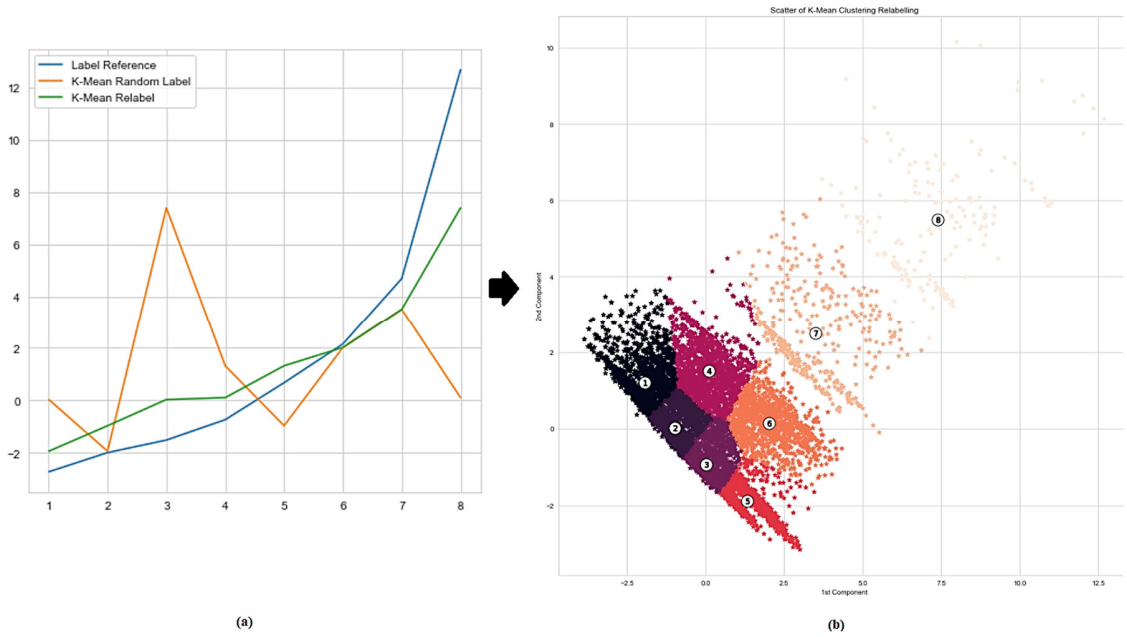

Fig. 4 Scatter Plot K-Means Cluster Relabel-based PCA Centroid Pattern Matching

Fig. 4a is a visualization of K-Means cluster relabeling process to rearrange the clustered labels. In this graph, the label is reported with an orange line, which tends to be random because the initial determination of the centroid is carried out randomly. The original reference label has a blue line indicating the pattern. To improve the interpretation of the results, a relabeling process is carried out by rearranging the cluster labels. This relabeling is shown with a green line. The process includes replacing the orange label with green based on the similarity of the pattern to the reference centroid. Therefore, the green label indicates a cluster in line with the original pattern. Fig. 4b shows the scatter plot results of clustering after the relabeling process. The scatter graph shows the data distribution based on two main components with cluster labels. The clustering results report the cluster's clarity level and the boundaries based on the data distribution. A well-defined cluster can be used for more accurate analysis or decision-making to predict STF cohorts.

TABLE 4
F1-SCORE FROM CLASSIFICATION MODEL

Dataset (p)	fold (q)	Classification Algorithm (r)					
		DT	KNN	LR	NB	RF	SVM
Original Labeled Data	3	82.98	83.11	58.48	59.65	84.29	81.03
	5	83.41	83.41	58.35	59.64	84.69	81.22
	10	83.63	83.49	58.47	59.69	84.85	81.58
Relabeled Data	3	99.50	98.98	92.82	95.54	99.65	99.85
	5	99.51	99.11	92.85	95.50	99.66	99.85
	10	99.56	99.15	92.93	95.56	99.73	99.60

A comparison is conducted in response to the accuracy results of the classification algorithm based on the dataset. The calculation results for the average of F1-Score and AUC can be shown in Tables 4 and 5, respectively. In general, there are differences in results between F1-score measurements for data with the original label and the new label. Statistical tests for two paired samples are necessary to determine the significance. The selection of the appropriate test statistic must be checked against the normal distribution. The normality test was carried out through the Kolmogorov-Smirnov test. The results showed that F1-score and AUC data were not normally distributed.

Fig. 5 presents the graphical representation of the data shown in Table 4. The average F1-score accuracy of the dataset with the new labels is consistently original labels. X-axis represents the experiment index, ranging from 1 to 18, which corresponds to the combination of 3-fold cross-validation and 6 classification algorithms (3 × 6 = 18 data points).

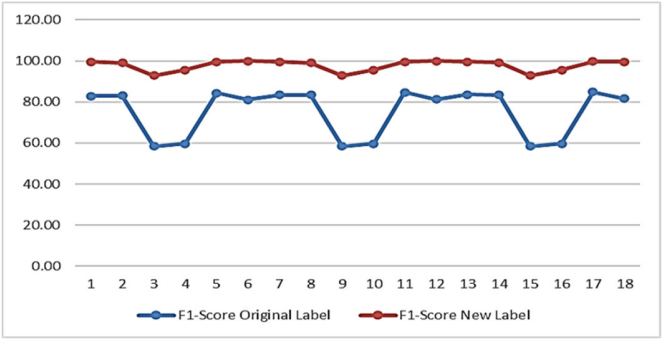

Fig. 5 F1-Score Between Original Label and New Label dataset

The graph compares the accuracy performance based on F1-Score between the data with the original and the relabeled labels. The red line shows better stability than the blue. This is important to ensure that model performance remains optimal across the various data conditions and classification methods. The average F1-score for the new label is consistently higher than the original, indicating an improvement in the quality of predictions. Fig. 5 visualizes F1-Score values, and the results presented in Table 5 are reported in a comparative chart in Fig. 6.

TABLE 5
AUC FROM CLASSIFICATION MODEL

Dataset (p)	fold (q)	Classification Algorithm (r)					
		DT	KNN	LR	NB	RF	SVM
Original Labeled Data	3	96.33	97.73	91.54	92.49	98.48	97.38
	5	96.51	97.84	91.54	92.48	98.55	97.44
	10	96.62	97.92	91.55	92.48	98.59	97.49
Relabeled Data	3	99.81	99.96	99.18	99.66	100.0	100.0
	5	99.81	99.63	99.17	99.66	100.0	100.0
	10	99.56	99.15	92.93	95.56	99.73	99.60

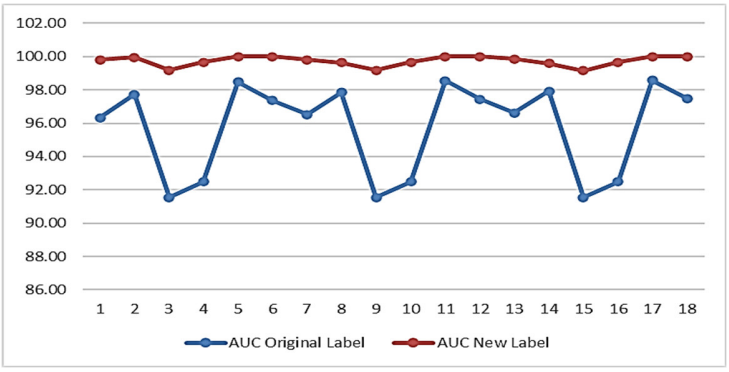

Fig. 6 AUC between the Original Label and the New Label dataset

This graph shows a comparison of AUC values between the original and the relabeled dataset. The red line tends to be more stable and has a higher AUC value than the blue line. Therefore, relabeling improves the performance of the model in differentiating between STF classes. Fluctuations in the blue line show that the original dataset is less consistent. The relabeling results positively impact data quality and model stability. Hypothesis testing is carried out since there is a difference in the accuracy results of the dataset structure with the original label and relabeling. The hypotheses to be tested are as follows:

H_0: There is no difference between F1-score or AUC measurements before and after K-Means cluster relabeling.
H_1: There is a difference between F1-score and AUC measurements before and after K-Means cluster relabeling.

TABLE 6.

THE WILCOXON SIGNED-RANK TEST FOR F1-SCORE AND AUC

Null Hypothesis	Z-Statistics	P-Value (Asymp. Sig. 2-tailed)	Decision
F1-Score New Label = F1-Score Original Label	-3.724b	0.000	H0 Rejected
AUC New Label = AUC Original Label	-3.724b	0.000	H0 Rejected

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

The Wilcoxon statistical test for paired samples shows that the results of F1 Score or AUC average on the dataset with the new label are significantly improved since the p-value is less than 5%. The data statistically support the results of F1-Score or AUC being significantly different between the dataset with the new and original labels.

IV. DISCUSSION

The application of clustering and relabeling methods to student socio-economic data significantly improves the accuracy of tuition fee grouping predictions in public universities. Therefore, this research addresses the issue of ambiguous label distributions in existing institutional datasets using centroid-based relabeling. Several research using fuzzy clustering [5], [6], [7] successfully applied unsupervised methods to categorize students. However, systematic validation of the resulting clusters was unavailable. This method introduces a validation phase before classification, ensuring that the clusters are generated and evaluated for structural quality using validity indices. The validation phase ensures that classification models are trained on reliable and meaningful labels. Table 7 shows the summary of relevant previous research.

TABLE 7
COMPARISON OF RELATED RESEARCH AND THE PRESENT WORK

Research	Method Used	Number of Clusters	Accuracy	Advantages	Limitations	Gap Addressed
[5], [6], [7]	Fuzzy Clustering	Varies (e.g., 3–5)	Not systematically validated	Able to categorize students without supervision	Lack of cluster validity assessment; ambiguous labels remain	This research adds a validation phase before classification
[9], [10]	Support Vector Machine (SVM)	–	~81% accuracy	Strong performance on structured data	Dependent on raw institutional labels; no relabeling	The relabeling process increases accuracy across all algorithms
[11], [12]	Label refinement and Classification	–	Not explicitly reported	Highlighted the importance of refining data labels	Did not integrate cluster validation	This research combines both cluster validation and relabeling
This research	Relabeling with Validation and Multi-Algorithm Classification (DT, KNN, NB, RF, SVM, ANN)	6–8 (depending on dataset)	>81%, statistically significant improvements (F1-Score & AUC)	Systematic, consistent improvement, practical for tuition fee classification	Relies on high-quality socio-economic data and computational resources	Provides both methodological and practical contributions in the Indonesian context

Table 7 compares applied methods, number of clusters, reported performance scores, as well as the strengths and limitations. Even though SVM [9], [10] reported accuracy levels of 81%, this research with six classification algorithms, namely DT, KNN, NB, RF, SVM, and Artificial Neural Networks (ANN), showed that the relabeling process obtained higher score across all models. Statistical tests at a 5% significance level confirm that the improvements are statistically significant, as reflected in F1-Score and AUC values. The results of previous investigations [11], [12] emphasized the importance of refining data labels before classification. However, this research integrates validation and relabeling into a unified and systematic framework directed to the context of STF classification. From a practical perspective, accurate tuition fee classification has direct implications for the equitable allocation of subsidies and scholarships since students from lower-income families receive the required support.

This research should be placed in a broader international context. In the United States, tuition fee is primarily market-driven, creating affordability challenges and dependence on financial aid [37]. In Europe, Germany and Finland maintain tuition free higher education, while others adopt targeted aid mechanisms to provide more limited support [38]. In Asia and Asia-Pacific, financing models are highly diverse, ranging from strong public funding to extensive cost-sharing and private provision shaped by challenges of access, equity, and governance [39], [40], [41].

These contrasts suggest the uniqueness of STF policy, which embeds equity directly into tuition-setting through cross-subsidization. The proposed data-driven method offers insights adapted to other contexts seeking fairer and more transparent tuition allocation. However, this research has several limitations that should be acknowledged. First, the method depends on the availability of high-quality socio-economic data, which may vary across institutions. Second, scalability may be challenged by the computational resources required for large-scale datasets. Third, the results are limited to Indonesian higher education context, requiring validation in different educational systems.

V. CONCLUSIONS

In conclusion, the quality of the grouping of education costs (STF) is successfully evaluated based on the university formula. The results of the preliminary analysis show that the distribution of STF labels is unclear and does not fully reflect the socio-economic data structure of students. This reports the need for better methods to ensure a more accurate clustering of STF. Furthermore, the grouping is rearranged using a centroid mapping-based relabeling process by implementing K-Means algorithm. The method produces new labels in line with the actual data patterns, providing a more substantial basis for predictive analysis. The relabeling process is validated by testing six classification algorithms, namely DT, KNN, NB, LR, RF, and SVM. The experimental results show that the relabeled data significantly improves the accuracy of the predictions compared to the original data, as evidenced by statistical tests at a significance level of 5%. Therefore, clustering based on the cluster goodness index can improve the quality of the dataset structure for future STF prediction. RF algorithm and SVM are the top recommendations in similar applications. This research significantly improves the quality of datasets for STF grouping and provides practical recommendations for higher education institutions. Universities can also develop more accurate and fairer prediction systems with better data structures and optimal classification algorithms. These results support data-driven decision-making in education cost allocation and provide a framework applied to various other social policies. The institutional context serves as a data source, even though a case from Universitas Negeri Surabaya is used to show the method. The primary contribution of this research lies in the methodological framework for cluster relabeling and classification optimization, generalized and applied to other institutions with similar tuition grouping challenges. The results explicitly answer the questions by confirming that the proposed relabeling method addresses the imbalance between original labels and socio-economic patterns.

Author Contributions: *Wiyli Yustanti*: Conceptualization, funding acquisition, investigation, methodology, project administration, resources, writing-original draft. *Andi Iwan Nurhidayat*: Data curation, visualization, and writing - review & editing. *Muhammad Iskandar Java* : Software, data curation, resources, validation, visualization..

All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Grant Scheme of the Research and Community Service Unit (LPPM) of Universitas Negeri Surabaya in 2024, under contract number B/116044/UN38.III.1/LK.04.00/2024, and the Director of Research, Technology, Research and Community Service (DRTPM) under contract number B/61572/UN38.III.1/LK.04.00/2024.

Acknowledgments: The authors express their gratitude for the financial support and facilitation provided by the institution.

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

Data Availability: The data that has been used is confidential.

Informed Consent: There were no human subjects.

Institutional Review Board Statement: Not applicable.

Animal Subjects: There were no animal subjects.

ORCID:
Wiyli Yustanti: https://orcid.org/0000-0002-9574-7072
Andi Iwan Nurhidayat: https://orcid.org/0009-0004-5100-4543
Muhammad Iskandar Java: https://orcid.org/0009-0009-3311-987X

REFERENCES

[1] D. Hooshyar, Y. Yang, M. Pedaste, and Y. M. Huang, "Clustering Algorithms in an Educational Context: An Automatic Comparative Approach," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3014948.
[2] M. Aamir and S. M. Ali Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, pp. 436–446, May 2021, doi: 10.1016/j.jksuci.2019.02.003.
[3] W. Yustanti and Y. Anistyasari, "A Polychoric Correlation to Identify the Principle Component in Classifying Single Tuition Fee Capabilities on the Students Socio-Economic Database," in *IOP Conference Series: Materials Science and Engineering*, 2018, p. 012150. doi: 10.1088/1757-899X/288/1/012150.
[4] W. Yustanti, Y. Anistyasari, and E. M. Imah, "Determining student's single tuition fee category using correlation based feature selection and support vector machine," *Proceeding International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, vol. January, pp. 172–176, 2018, doi: 10.1109/ICACSIS.2017.8355029.
[5] Indrawati, Anwar, and N. Amalia, "Determination System of Single Tuition Group Using a Combination of Fuzzy C-Means Clustering and Simple Additive Weighting Methods," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, 2019. doi: 10.1088/1757-899X/536/1/012148.
[6] A. W. Sugiyarto, R. Pamungkas, A. R. Rasjava, and A. M. Abadi, "Fuzzy Multi Attribute Decision Making (FMADM) Implementation for Classifying Student's Single Tuition Fee (UKT) Based on Android Applications," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Dec. 2019. doi: 10.1088/1742-6596/1397/1/012061.
[7] H. Syahputra, Sutrisno, and S. Gultom, "Decision Support System for Determining the Single Tuition Group (UKT) in State University of Medan Using Fuzzy C-Means," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Mar. 2020. doi: 10.1088/1742-6596/1462/1/012071.
[8] T. F. Abidin, S. Rizal, T. M. Iqbalsyah, and R. Wahyudi, "Decision tree classifier for university single rate tuition fee system," *International Journal of Business Intelligence and Data Mining*, vol. 17, no. 2, pp. 258–271, 2020, doi: 10.1504/IJBIDM.2020.108764.
[9] W. Yustanti, N. Iriawan, and Irhamah, "A Hybrid Evaluation Index Approach in Optimizing Single Tuition Fee Cluster Validity," *Proceeding of 6th International Conference on Information Technology, Information Systems and Electrical Engineering: Applying Data Sciences and Artificial Intelligence Technologies for Environmental Sustainability, ICITISEE 2022*, pp. 154–159, 2022, doi: 10.1109/ICITISEE57756.2022. 10057653.
[10] W. Yustanti, N. Iriawan, and Irhamah, "Categorical encoder based performance comparison in preprocessing imbalanced multiclass classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, pp. 1705–1715, 2023, doi: 10.11591/ijeecs.v31.i3.pp1705-1715.
[11] D. L. S. Reddy, M. Ramchander, B. R. Babu, and M. Geetalatha, "Comparitive study of outlier analysis methods in improving classifier accuracy on categorical data," *International Conference on Microelectronics, Computing and Communication, MicroCom 2016*, vol. 1, pp. 1–6, 2016, doi: 10.1109/MicroCom. 2016. 7522476.
[12] I. Škrjanc, J. Iglesias, A. Sanchis, D. Leite, E. Lughofer, and F. Gomide, "Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A Survey," *Inf Sci (N Y)*, vol. 490, 2019, doi: 10.1016/j.ins.2019.03.060.
[13] K. J. , Cios, W. , Pedrycz, and R. W. Swiniarski, *Data Mining Methods for Knowledge Discovery*, vol. 458. Boston: The Springer International Series in Engineering and Computer Science, 1998. doi: https://doi.org/10.1007/978-1-4615-5589-6_1.
[14] F. L. Gewers *et al.*, "Principal component analysis: A natural approach to data exploration," *ACM Comput Surv*, vol. 54, no. 4, 2021, doi: 10.1145/3447755.
[15] W. Yustanti, N. Iriawan, and Irhamah, "Categorical encoder based performance comparison in preprocessing imbalanced multiclass classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, 2023, doi: 10.11591/ijeecs.v31.i3.pp1705-1715.
[16] D. A. Simovici, "Dimensionality Reduction Techniques," in *Linear Algebra Tools for Data Mining*, 2023. doi: 10.1142/9789811270345_0013.
[17] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi, and I. W. A. S. Darma, "The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 13–24, 2024, doi: 10.20473/jisebi.10.1.13-24.
[18] K. Fahriya and W. Yustanti, "Optimalisasi Jumlah Klaster Uang Kuliah Tunggal pada Data Sosial Ekonomi Mahasiswa," *Journal of Emerging Information Systems and Business Intelligence (JEISBI)*, vol. 02, no. 02, pp. 73–77, 2021, [Online]. Available: https://rb.gy/5dzjg
[19] G. Gan, C. Ma, and J. Wu, *Data Clustering : Theory, Algorithms and Applications*. American Statistical Association and the Society for Industrial and Applied Mathematics. 10, 2007. doi: 10.1017/ CBO978 1107415324.004.
[20] V. D'Orangeville, M. A. Mayers, M. E. Monga, and M. S. Wang, "Efficient cluster labeling for support vector clustering," *IEEE Trans Knowl Data Eng*, vol. 25, no. 11, pp. 2494–2506, 2013, doi: 10.1109/TKDE. 2012. 190.
[21] W. Zhu and Y. Fan, "Relabelling Algorithms for Large Dataset Mixture Models," Mar. 2014.
[22] J. Lee and D. Lee, "An improved cluster labeling method for support vector clustering," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 3, pp. 461–464, Mar. 2005, doi: 10.1109/TPAMI.2005.47.
[23] H. L. Chen, K. T. Chuang, and M. S. Chen, "On data labeling for clustering categorical data," *IEEE Trans Knowl Data Eng*, vol. 20, no. 11, pp. 1458–1471, 2008, doi: 10.1109/TKDE.2008.81.
[24] A. A. Klaib, A. A. Milad, and M. A. Algaet, "A New Approach for Labelling XML Data," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2021*, Institute of Electrical and Electronics Engineers Inc., Sep. 2021, pp. 603–607. doi: 10.1109/3ICT53449.2021.9581352.
[25] W. C. Sleeman IV *et al.*, "A Machine Learning method for relabeling arbitrary DICOM structure sets to TG-263 defined labels," *J Biomed Inform*, vol. 109, Sep. 2020, doi: 10.1016/j.jbi.2020.103527.

[26] M. Sperrin, T. Jaki, and E. Wit, "Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models," *Stat Comput*, vol. 20, no. 3, pp. 357–366, 2010, doi: 10.1007/s11222-009-9129-8.
[27] Z. Li, J. Li, Y. Liao, S. Wen, and J. Tang, "Labeling clusters from both linguistic and statistical perspectives: A hybrid approach," *Knowl Based Syst*, vol. 76, pp. 219–227, Mar. 2015, doi: 10.1016/j.knosys.2014.12.019.
[28] R. Kusumaningrum and Farikhin, "An Automatic Labeling of K-means Clusters based on Chi-Square Value," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Mar. 2017. doi: 10.1088/1742-6596/801/1/012071.
[29] H. Wan, "Cluster-Based Supervised Classification," Ulster University, Northen Ireland, 2020.
[30] H. Wan, H. Wang, B. Scotney, J. Liu, and X. Wei, "Cluster-based Data Relabelling for Classification," *Information Sciences SSRN*, pp. 1–32, Jul. 2022.
[31] S. Banitaan, A. B. Nassif, and M. Azzeh, "Class decomposition using K-means and hierarchical clustering," in *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, Institute of Electrical and Electronics Engineers Inc., Mar. 2016, pp. 1263–1267. doi: 10.1109/ICMLA.2015.169.
[32] B. Sowan, N. Matar, F. Omar, M. Alauthman, and M. Eshtay, "Evaluation of class decomposition based on clustering validity and k-means algorithm," in *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ACIT50332.2020.9300084.
[33] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J Comput Appl Math*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
[34] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, vol. 3, no. 10, pp. 27–38, 2013, [Online]. Available: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633
[35] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," pp. 1–13, 2020.
[36] N. Sano and Y. Hattori, "Utility evaluation measures for categorical data by classification performance," *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2019-Novem, pp. 356–361, 2019, doi: 10.1109/ICDMW.2019.00059.
[37] S. Brint, "Challenges for higher education in the United States: The cost problem and a comparison of remedies," *Eur J Educ*, vol. 57, no. 2, pp. 181–198, Jun. 2022, doi: 10.1111/ejed.12496.
[38] K. Czarnecki, T. Korpi, and K. Nelson, "Student support and tuition fee systems in comparative perspective," *Studies in Higher Education*, vol. 46, no. 11, pp. 2152–2166, 2021, doi: 10.1080/03075079.2020.1716316.
[39] M. Bray, "Financing higher education: Patterns, trends and optionsT," *Prospects (Paris)*, vol. 30, pp. 331–348, Sep. 2000, doi: https://doi.org/10.1007/BF02754057.
[40] A. Welch, "Governance Issues in South East Asian Higher Education: Finance, Devolution and Transparency in the Global Era," *Asia Pacific Journal of Education*, vol. 27, no. 3, 2007, doi: 10.1080/02188790701601805.
[41] W. J. Jacob, D. Neubauer, and H. Ye, "Financing trends in Southeast Asia and Oceania: Meeting the demands of regional higher education growth," *Int J Educ Dev*, vol. 58, pp. 47–63, Jan. 2018, doi: 10.1016/J. IJEDU DEV.2016.11.001.

Publisher's Note: Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.