Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 11, No. 1, March 2023, pp. 152~163
ISSN: 2089-3272, DOI: 10.52549/ijeei.v11i1.4376

Mixed-type Variables Clustering for Learners' Behavior in Flipped Classroom Implementation

Daniel Febrian Sengkey[1], Angelina Stevany Regina Masengi[2]
[1]Department of Electrical Engineering, Universitas Sam Ratulangi, Manado, Indonesia
[2] Department of Pharmacology and Therapy, Universitas Sam Ratulangi, Manado, Indonesia

Article Info

Article history:

Received Dec 14, 2022
Revised Jan 14, 2023
Accepted Feb 20, 2023

Keyword:

Learning behavior analysis
Clustering
Mixed-type variables
Blended learning
Flipped classroom

ABSTRACT

Numerous approaches have been developed to group learners' behavior in an online/blended learning environment. However, most clustering analyses in this particular field only consider numeric features despite the existence of categoric features that are found important in other studies. In this study, we compare K-Means and K-Prototypes algorithms to cluster learners' behavior in a flipped classroom implementation. From the model selection, we found that the model produced by the K-Prototypes algorithm — which included categoric features — is a better one. The statistical analysis of the clustering results of the selected K-Prototypes model shows significant differences in most of the inter-cluster comparisons, implying a good separation of the data. More importantly, we can identify the behavior in each cluster which then can be used to help learners in achieving better results in learning.

Corresponding Author:

Daniel Febrian Sengkey,
Department of Electrical Engineering,
Universitas Sam Ratulangi,
Jl. Kampus UNSRAT, Bahu-Manado, 95115, Indonesia
Email: danielsengkey@unsrat.ac.id

1. INTRODUCTION

Even though the adoption of Information Technology (IT)-assisted learning has surged in recent years due to the COVID-19 pandemic, it has existed since the 1960s [1]–[4]. An instance of the implementation of IT-assisted learning is Electronic Learning (e-Learning), where electronic devices such as Compact Disc is used to deliver learning materials across geographically separated learning participants, and later, online learning, where learners can access the learning materials through a server on the Internet [5]. Now, the utilization of technology is found as brings positive impacts to learners with digital backgrounds, such as the ones in the present time [6]. It is also reported that blended learning — where the learners have pre-classroom learning activities, provided in a similar way as online learning — allows a more collaborative learning experience, and even comes up as a preferred method by the learners [6]–[8]. The method holds potential, not only for undergraduates but also for graduate students [9], [10]. It also positively contributes to learning achievements [8], [11].

Some studies in this field specifically addressed the use of learning media [12]–[15]. It is explainable since the use of the right media would give the utmost benefit to the learners, since the different types of media might have different impacts on learning achievements [15], [16]. The studies [12]–[15] were dedicated to comparing three types of learning media: ones with only text and images; recorded slideshow with audio narration from the lecturer/instructor; and recorded slideshow with the lecturer appearing in-frame, explaining the particular topic. Several aspects have been examined, such as the preferences of the learners, their behavior in a flipped classroom (which one got the earliest access? Which one was accessed more frequently?), and how these factors contribute to the learning achievements. These studies made use of questionnaires to gather

feedback from students, and later, the studies in [13]–[15] augmented the data with records from the Learning Management System (LMS).

LMS log has been known to provide beneficial information regarding learning activities. The approach has been used to develop a system that will alert educators regarding the students that plausibly failed the course [17]. LMS log has also been used to estimate learners' performance [18], model learners' approach to studying [19], [20], and how learners interact with the LMS [21]. It is also used as the basis for learning analytics, to predict learners' performance [18]. In [22], eye strain detection is developed with the possibility of embedding it onto an LMS, makes possible for educators to design learning materials. Unfortunately, the students' activities that are recorded in an LMS logging system might produce a lot of data. For example, the log file used in [13]–[15] consists of more than 63 thousand records. This leads to an issue with the analysis of data with such volume [23].

Recently, with the emergence of Machine Learning (ML) applications in various fields, we have seen some works in IT-assisted learning that adopt the technique, as we summarized in Table 1. Li and Yoo used Bayesian Markov Chain to model learning styles in online learning based on the learners' activities [24]. As a result, they can identify behavioral changes when learners move from one lab to another one. Similarly, Köck and Paramythis used Discrete Markov Models to discover information regarding learners [25].

Table 1. Former studies of the implementation of cluster analysis in modeling learners' behavior, ordered by publication year.

Reference	Data Source	Algorithm	Publication Year
Li and Yoo [24]	LMS (tutoring system) and questionnaire	Markov Chain (Bayesian)	2006
Chen et al. [26]	LMS	Fuzzy	2009
Hogo [27]	Web server access log	Fuzzy (FCM and KFCM)	2010
Köck and Paramythis [25]	Published dataset [28], [29]	Discrete Markov Model	2011
Jovanovic et al. [30]	Questionnaire on cognitive styles	K-Means[1]	2012
Bovo et al. [31]	LMS (Moodle log data)	Expectation Maximization, Hierarchical Clustering, Simple K-Means, X-Means	2013
Akçapınar [19]	LMS (Moodle log data)	K-Means	2015
Liu and D'Aquin [32]	Published dataset [33]	K-Prototypes	2017
Charitopoulos et al. [21]	LMS (Moodle database)	K-Means	2017
Triayudi and Fitri [34]	LMS	Single Linkage Dissimilarity Increment Distribution-Global Cumulative Score Standard (SLG)	2019
Moubayed et al. [35]	LMS	K-Means	2020
Palani [36] and Palani et al. [37]	Published dataset [33]	FCM, AHC, K-Prototypes, Gaussian Mixture Model	2020 [36], 2021 [37]
Nalli et al. [38]	LMS (Moodle database)	K-Means, Mean-Shift Clustering, Agglomerative Clustering, Density-based spatial clustering of applications with noise (DBSCAN), Gaussian Mixture Model, SOM	2021
Ge et al. [39]	Questionnaire	K-Prototypes	2021
Dhaiouir et al. [40]	LMS	SOM	2022

Fuzzy Clustering was used by Chen et al. to cluster learners' behaviors based on the number of clicks and the number of stays on the resources [26]. The work of Hogo also used fuzzy clustering, specifically the

[1] K-Means was adapted for use over the categorical cognitive styles data

Fuzzy C-means (FCM) and the Kernelized Fuzzy C-means (KFCM), with the latter coming up as a better algorithm [27].

K-Means is another clustering algorithm that gains popularity in recent years. It works by calculating and then minimizing the Euclidean distance between each data point and a cluster centroid. Jovanovic et al. used K-Means to cluster learners' cognitive styles collected by using a questionnaire, which was later utilized to develop learning modules for learners with specific needs [30]. Akçapınar [19] also used the same algorithm to cluster learners in an online class based on Moodle log data. Similarly, Charitopoulos et al [21]. extracted the records in Moodle database to get the Time Between Action (TBA) of the learners while using the LMS, then fed the data into the K-Means algorithm to cluster the learning contents based on the TBA of the learners. K-Means was also used to model the engagement levels of the learners, with the number of logins and average duration of assignment submission as the indicative ones [35]. It is found as the best algorithm for the automated creation of a heterogeneous group of learners and is implemented as a Moodle plugin [38]. In another study, K-Means is used to cluster learners' personalities based on the Myers-Briggs Type Indicator (MBTI) [41].

Other clustering algorithms that have been adopted for behavioral analysis of the learners in an IT-assisted learning environment are Agglomerative Hierarchical Clustering (AHC) and Self-Organizing Map (SOM). Triayudi and Fitri [34] developed a new AHC algorithm to cluster the interpersonal of the students. SOM was studied by Nalli et al. [38], and compared to the other clustering algorithms. It is also used by Dhaiouir et al. [40] to group learners to help them find the Massive Open Online Course (MOOC) that suits their profile.

Despite their wide application, in this particular field of study, the numeric-based clustering algorithms have a drawback, they could only work with quantitative data whereas in learning behavioral analysis, the data could be in numeric and categoric data. Despite the possibility of adapting K-Means for categoric data as done by Jovanovic et al. [30], using an algorithm that is specified for the mixed type of data such as K-Prototypes [42] is a preferable method. For instance, Liu and D'Aquin used K-Prototypes to build clusters of learners since they also incorporated categorical variables such as gender, region, and highest education [32]. In another study, Palani [36] and Palani et al. [37] found that K-Prototypes yield better separation compared to other clustering algorithms when applied for the identification of at-risk learners. Ge et al. [39] studied how learning media influenced learners' familiarity with online learning, hence they used K-Prototypes as their clustering algorithm.

Even though we have seen various implementations of cluster analysis in IT-assisted learning, there is one drawback: most studies are constrained to numeric variables despite there being non-numeric variables that might have an impact on this particular matter, such as the learning media. It is proven to be an important factor in learning, as shown by previous studies[12]–[15], [39], [43], [44]. In our literature search, the study by Ge et al. [39] is the only one that includes learning media in their cluster analysis, however, they did not include the learners' achievement data. Therefore, in this study, we seek to deepen the findings from the previous works that discuss learners' behaviors in a flipped classroom implementation by employing clustering of mixed data types. The remainder of this article is organized as follows: in Section 2, we present the methods we used then followed by the Result and Discussion of the findings in Section 3. Last, in Section 4, we conclude our work as well as present some possibilities for future work.

2. RESEARCH METHOD

In this study, we used the same data that were used in Refs. [13]–[15]. The data came from the processed LMS log, results of the in-class quiz, and course design, as shown in Table 2. Clustering is achieved by using two well-known algorithms: K-Means and K-Prototypes. Even though it is clear in Section 1 that K-Prototypes are suitable for these data, we incorporated K-Means as a comparison. As in this work, we intend to study the effect of clustering with the mixed-type data, hence no modification has been made to both algorithms.

When clustering the data, K-Prototypes included all the features while K-Means only included the numerical features. The optimum number of clusters was decided by using the Elbow Method, which is a comparison of the within-cluster sum of squares (WCSS). For evaluation purposes, we used LightGBM (LGBM) classification algorithm [45] with the cluster number as the label. This approach is selected due to the unusual nature of the study, where we compared the results of clustering, not only the performance of the algorithms. In using a classifier for cluster evaluation, if the clustering result is good enough, there should be clear distinctions between the data belonging to a cluster compared with data that belong to another cluster. With such distinctions between clusters, then a good classifier model can be built. LGBM was chosen since it could accommodate categoric variables.

The model selection is based on the cross-validation (CV) score of LGBM since it reflects the quality of the model, which indirectly demonstrates the separation between clusters. The CV score is one of the evaluation metrics for the classification models. These metrics are commonly represented between 0.00 to 1.00 or in percentage. Despite the differences between models being small, they are still considered important in deciding the best model, as demonstrated in [46], [47]. Besides using the CV score, the LGBM model is also evaluated by using Shapley Additive Explanation (SHAP) value to see the influence of each feature [48]. Then, the feature with a low SHAP value is dropped and the process started again from clustering, without the dropped feature. These steps were repeated until the cross-validation score reached a considerably sufficient number.

Table 2. Clustering features

Feature	Type	Source	Description
Name	Categoric	LMS log and In-class quiz results	Learner's name
Session	Categoric	LMS log, Course Design	The session code of the lecture
Media Type	Categoric	LMS log, Course Design	The code of the media type as defined in [12]–[15], [43], [44]
Time Delta (First access time related to lecture schedule)	Numeric	Processed LMS log	The number of hours when a learner for the first time made access to a particular learning media before the related lecture started. If the first access was made after the schedule, the value was recorded as a negative one.
Access frequency	Numeric	Processed LMS log	The number of accesses made by a particular learner towards a particular learning media.
Score	Numeric	In-class quiz results	The number of correct answers made by a learner for each topic, where each topic was delivered by a certain type of learning media.

3. RESULTS AND DISCUSSION

3.1. Model Selection

In the first phase, a model (KM1) was built using K-Means with only Time Delta, Access Frequency, and Score features. Based on the Elbow Plot (Figure 1), the ideal number of clusters is five. Then another model (KP1) was built using K-Prototypes, now with the categorical features included. As can be seen from Figure 1, the ideal number of clusters is also five for this model. When evaluated with LGBM (see Table 3), the KM1 is the better one despite the existence of features that were not used when building the model. The SHAP value of KM1 (Figure 2) suggests that Session has very little impact. On the other hand, the SHAP value of KP1 (Figure 3), Session is also the feature with the least impact, however, here it has a higher impact than KM1 since KP1 incorporated this feature when building the clustering model.

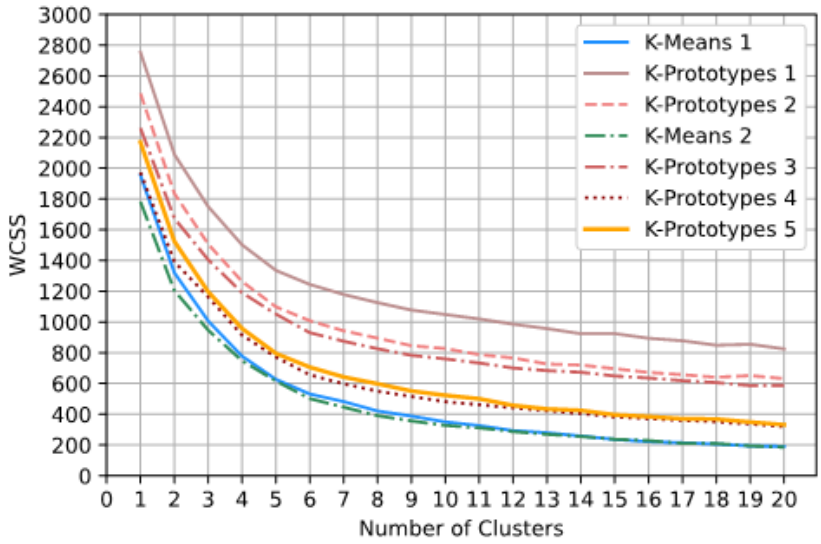

Figure 1. Elbow plot for all models

Table 3. Comparisons of LGBM cross-validation score

Model	Algorithm	LGBM score	CV
K-Means 1 (KM1)	K-Means	97.70%	
K-Prototypes 1 (KP1)	K-Prototypes	95.21%	
K-Prototypes 2 (KP2)	K-Prototypes	97.24%	
K-Means 2 (KM2)	K-Means	97.89%	
K-Prototypes 3 (KP3)	K-Prototypes	96.73%	
K-Prototypes 4 (KP4)	K-Prototypes	96.65%	
K-Prototypes 5 (KP5)	K-Prototypes	98.00%	

Table 4 shows the features used to build each clustering model. Since Session has the least impact on the LGBM model, another clustering model is built without this feature. As the KM1 since the beginning did not use Session, then the new clustering model is built with K-Prototypes (KP2). The sequence of building the clustering model, evaluating with LGBM and SHAP, dropping features with the least impact, or modifying features were repeated several times until we get the KP5 model that is considered adequate (Figure 4). The models KM2, KP3, and KP4 were built by dropping the first access made by learners more than 72 hours after the related lecture since they are considered outliers or abnormal access [13]–[15]. However, the evaluation with LGBM to KP3 and KP4 did not give a better cross-validation score compared to KP2, therefore, in KP5 the same feature sets with KP4 were used, but the time delta included all data. The result is better than any of the previous models. Table 3 shows the cross-validation scores of each model.

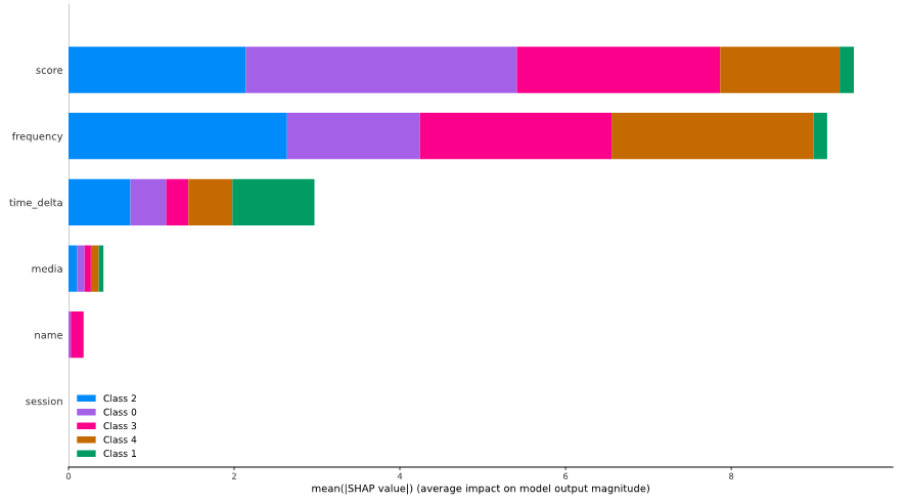

Figure 2. SHAP summary plot of KM1 model

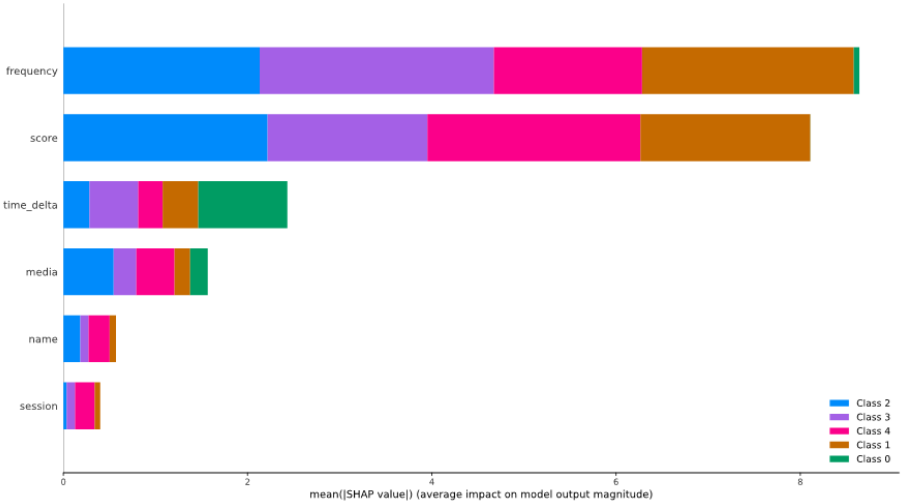

Figure 3. SHAP summary plot of KP1

Table 4. Features used to build each clustering model.

Feature	KM1	KP1	KP2	KM2	KP3	KP4	KP5
Name		✓	✓		✓		
Session		✓					
Media Type		✓	✓		✓	✓	✓
Time Delta	✓	✓	✓	✓[2]	✓[2]	✓[2]	✓
Access frequency	✓	✓	✓	✓	✓	✓	✓
Score	✓	✓	✓	✓	✓	✓	✓

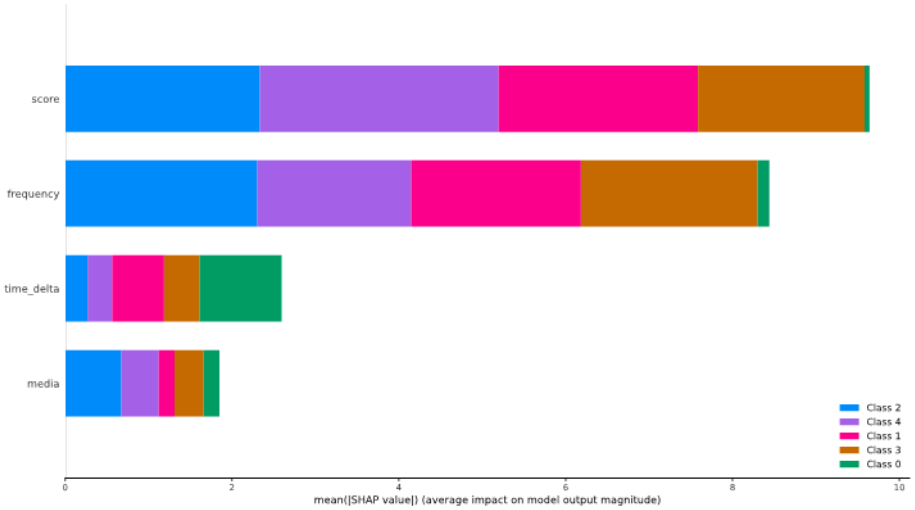

Figure 4. SHAP summary plot of KP5

3.2. Cluster Analysis

Based on the clustering results of the model KP5, we made a cluster analysis to gain knowledge about the learners' behavior in each cluster. Figure 5 shows the two-dimensional representation of the clustering results. For reading convenience, the Time Delta in Figure 5 is limited to no later than 72 hours after the lecture schedule, despite the KP5 including all the data in each feature selected. The first access made after 72 hours was not plotted since these outliers are obscuring the trend in the graph. As can be visually observed in Figure 5, each cluster has its trend whether in terms of the media, access frequency, first access time, and the achievements of the learners. Figure 6 shows the count of access to each media type. This chart shows that SAD is the only media type accessed by learners in Cluster 0, while in Cluster 1, it is the dominant one. Learners in Cluster 2 preferred video-based media. On the other hand, the learners in Clusters 3 and 4 mainly chose the text-based media with a surge in the latter.

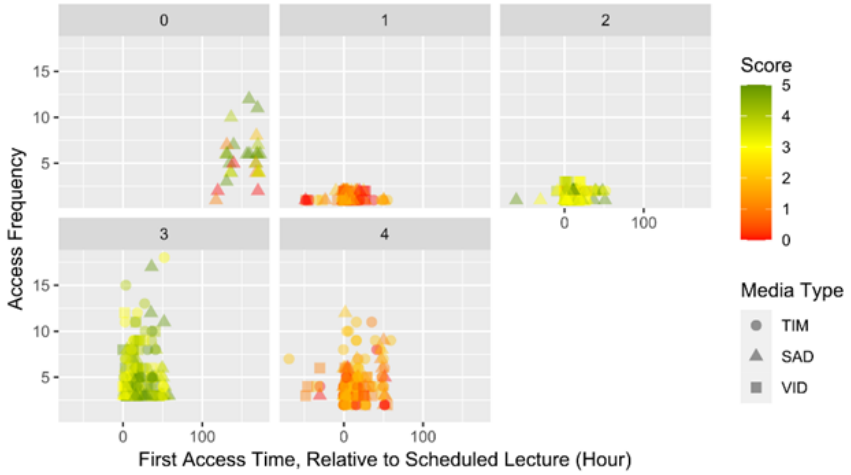

Figure 5. A scatter plot with all data and features of KP5. The time delta only includes access no later than 72 hours after each lecture.

[2] First access made before the lecture or no more than 72 hours after the lecture

Even though from visual observation to Figure 5 the trends are already visible, this method is not reliable and must be supported with statistical tests.

Table 5 shows the results from the Shapiro-Wilk test, applied to each numeric variable in each cluster, grouped by the media type. With a 95% confidence level ($\alpha = 0.05$), this distribution normality test indicates that most of the values are not normally distributed, hence further tests must be carried out with non-parametric tests. Therefore, the Kruskal-Wallis test was applied to each numeric variable, grouped by cluster. Table 6 presents the results of the test. It can be seen that with $\alpha = 0.05$, the p-values indicate that for all variables, there are differences between every cluster. This result was then followed by a pairwise comparison between clusters, by using the Wilcoxon test with Bonferroni adjustment, to see where the significant differences take place. The results are presented in Table 7.

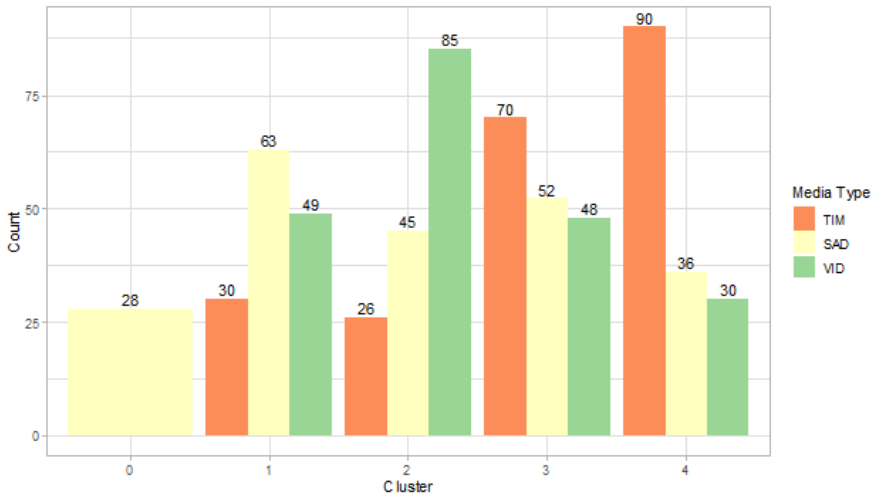

Figure 6. Media and count of learners in each cluster.

Table 5. Results of the Shapiro-Wilk test for each numeric variable grouped by media in each cluster.

		Time Delta		Access Frequency		Score	
Cluster	Media	W	p-value	W	p-value	W	p-value
0	SAD	0.832	< 0.001	0.934	0.078	0.806	< 0.001
1	TIM	0.546	< 0.001	0.404	< 0.001	0.770	< 0.001
1	SAD	0.390	< 0.001	0.595	< 0.001	0.782	< 0.001
1	VID	0.661	< 0.001	0.472	< 0.001	0.805	< 0.001
2	TIM	0.652	< 0.001	0.637	< 0.001	0.583	< 0.001
2	SAD	0.618	< 0.001	0.632	< 0.001	0.776	< 0.001
2	VID	0.664	< 0.001	0.786	< 0.001	0.873	< 0.001
3	TIM	0.925	< 0.001	0.853	< 0.001	0.789	< 0.001
3	SAD	0.894	< 0.001	0.706	< 0.001	0.778	< 0.001
3	VID	0.919	0.003	0.845	< 0.001	0.773	< 0.001
4	TIM	0.791	< 0.001	0.839	< 0.001	0.752	< 0.001
4	SAD	0.863	< 0.001	0.830	< 0.001	0.760	< 0.001
4	VID	0.898	0.007	0.901	0.009	0.717	< 0.001

Table 6. Results of the Kruskal-Wallis H test for each numeric variable grouped by cluster.

Variable	n	H	df	p-value
Score	652	468.088	4	< 0.001
Time Delta	652	197.904	4	< 0.001
Access Frequency	652	471.044	4	< 0.001

It can be seen from the results that presented in Table 7, the difference between clusters for each is significant in almost all pairs of clusters. The test only yielded insignificant results in the Score comparison between Cluster 0 and Cluster 2, between Cluster 0 and Cluster 3, between Cluster 1 and Cluster 4; and in the Access Frequency between Cluster 0 and Cluster 3. Despite the insignificant differences between Cluster 0 and Cluster 3 in the two numeric variables, these two clusters still have differences in the preferred media and in the Time Delta. It is interesting since almost in all cases, Time Delta and Access Frequency are always closely related, but between these two clusters, the difference in Access Frequency is insignificant, while the difference

in Time Delta is significant. This implies that although the learners in both clusters access the media at a similar frequency, the relative access time before the scheduled lecture is not similar. This phenomenon can also be visually inferred from the scatter plot in Figure 5.

Table 7. Results of Pairwise Wilcoxon Test with Bonferroni adjustment.

Variable	group1	group2	n1	n2	Z	p	p.adj	Sig.
Score	0	1	28	142	3289	< 0.001	< 0.001	***
Score	0	2	28	156	2353.5	0.493	1.000	ns
Score	0	3	28	170	2252	0.633	1.000	ns
Score	0	4	28	156	3495	< 0.001	< 0.001	***
Score	1	2	142	156	247.5	< 0.001	< 0.001	***
Score	1	3	142	170	0	< 0.001	< 0.001	***
Score	1	4	142	156	9212.5	0.007	0.068	ns
Score	2	3	156	170	10555.5	0.001	0.007	**
Score	2	4	156	156	23980.5	< 0.001	< 0.001	***
Score	3	4	170	156	26520	< 0.001	< 0.001	***
Time Delta	0	1	28	142	3976	< 0.001	< 0.001	***
Time Delta	0	2	28	156	4368	< 0.001	< 0.001	***
Time Delta	0	3	28	170	4760	< 0.001	< 0.001	***
Time Delta	0	4	28	156	4368	< 0.001	< 0.001	***
Time Delta	1	2	142	156	7929	< 0.001	< 0.001	***
Time Delta	1	3	142	170	3724	< 0.001	< 0.001	***
Time Delta	1	4	142	156	5798.5	< 0.001	< 0.001	***
Time Delta	2	3	156	170	7125	< 0.001	< 0.001	***
Time Delta	2	4	156	156	9779	0.003	0.027	*
Time Delta	3	4	170	156	16725	< 0.001	< 0.001	***
Access Frequency	0	1	28	142	3854	< 0.001	< 0.001	***
Access Frequency	0	2	28	156	4137	< 0.001	< 0.001	***
Access Frequency	0	3	28	170	2615	0.397	1.000	ns
Access Frequency	0	4	28	156	2978	0.002	0.019	*
Access Frequency	1	2	142	156	7188	< 0.001	< 0.001	***
Access Frequency	1	3	142	170	0	< 0.001	< 0.001	***
Access Frequency	1	4	142	156	425	< 0.001	< 0.001	***
Access Frequency	2	3	156	170	325	< 0.001	< 0.001	***
Access Frequency	2	4	156	156	1567.5	< 0.001	< 0.001	***
Access Frequency	3	4	170	156	17045.5	< 0.001	< 0.001	***

ns: not significant
*: $p \leq 0.05$
**: $p \leq 0.01$
***: $p \leq 0.001$

Measuring the correlations between variables in a cluster can give more insight into the learners' behavior in each cluster. Here we adopt the method used in [15] by treating the media as an ordinal variable, starting with the simplest form of media, TIM, to the most complex one, VID. Additionally, we also measured the correlation between Cluster and Media Type, treating both as categoric variables. Table 8 shows the result of the Chi-square test between clusters and media types. The categoric correlation implies that there is a significant difference in the media distribution in each cluster, as proof that the selected clustering model has succeeded to cluster the data also based on the categoric feature. It is interesting that for each pair of variables, the correlation strengths and directions could be different from one cluster to another.

Table 8. Result of the Test of Independence to the Clusters and the Media types.

$\chi 2$	df	p-value
148.93	8	< 0.001

Table 9 shows the correlations for the numeric variables in the same cluster. As described earlier, between Clusters 0 and 3, even though the difference in Time Delta is significant while not in Access Frequency, it can be seen that the correlation between stronger in Cluster 0.

Table 9. Intra-cluster correlations of the numeric variables

Correlation	0	1	2	3	4
Time Delta – Media	-	-0.023***	-0.167***	-0.028***	0.017***
Time Delta – Access Frequency	0.231**	0.126***	0.099***	0.014***	0.145***
Time Delta – Score	0.136***	0.013***	-0.109***	0.173***	-0.006***
Media – Access Frequency	-	0.008***	0.182***	-0.183**	-0.036***
Media – Score	-	-0.083***	0.018***	0.208***	0.064***
Access Frequency – Score	0.461*	-0.078***	-0.023***	0.000***	0.151***

* $p \leq 0.05$
** $p \leq 0.01$
*** $p \leq 0.001$

Based on the inter-cluster analysis, Table 10 shows the summaries of the observed behavior of the learners in each cluster.

Table 10. Observed behavior in each cluster, based on the KP5 model.

Cluster	Observed Behavior
0	Learners in this cluster only accessed the SAD media type with the access frequency almost distributed evenly. The first accesses were generally made long before the lecture (more than 100 hours earlier). The achievements are also distributed almost evenly. Based on the correlations in Table 7, learners in this cluster will get a much higher score if they make more frequent access to the learning media.
1	Learners in this cluster mainly accessed the SAD media type with low access frequency. The first accesses were generally made around the lecture schedule. The achievements are quite low in this cluster. Based on the correlations in Table 7, all variables almost have no correlation to the other variables except for the access frequency and time delta which has a relation.
2	This cluster shares some similarities with cluster 1, except for the earlier access and slightly better achievements. However, based on the correlations in Table 7, the time delta and access frequency have a weaker correlation here, but the media type and access frequency are found to have a relation.
3	Learners in this cluster achieved higher scores compared to those in the other clusters. There are various media types used by the learners with medium to high access frequency nearing the scheduled lecture. There is an anomaly with the learners in this cluster where there were learners who accessed a learning media for the first time not long before the lecture yet still got high scores, as reflected by the correlations in Table 7, where the access frequency does not contribute to the learners' achievement. It can be seen also that more complex media types, such as video-based ones have a positive impact on achievements.
4	Learners in this cluster tend to access the TIM media type. Despite the first accesses being made before the lecture, there is a noticeable number of them that were made after the lecture. Even though the trends of access frequency and time delta are almost similar to the ones in cluster 3, unfortunately, learners in this cluster achieved lower scores.

Compared to the previous works in [13]–[15], where the learners were only treated as a single population without giving any concern to the behavioral aspects that influenced either learning styles and or achievements, the clustering with mixed-type variables was found to give more meaningful insights where each cohort might be treated differently to achieve a better overall learning performance. Therefore, the results could be beneficial either for the instructors where ones might set up a learning plan where the methods and media accommodate the specific needs of the learners or for the higher education management parties where they can put students with similar interests and or behavior at the same class where the particular instructor can apply a specific treatment.

4. CONCLUSION

In this era of digital disruption, many aspects of human life have adopted technology, including education. The flipped classroom is an implementation of IT-assisted learning, where learners might access a particular topic before class, and then discuss it in class with peers and/or instructor, as well as explore more advanced issues within a such topic, with the assistance of the instructor. In this situation, learning media is an important factor to deliver the learning materials. Unfortunately, regardless of the abundance of studies to group learners according to their behavior in online and blended learning implementation, including flipped classroom implementation, learning media is rarely discussed since most works in this field only consider numeric variables.

In this work, we implemented K-Means and K-Prototypes clustering algorithms to the behavior data of learners in a flipped classroom. The implementation of K-Means only included numeric features while the implementation of K-Prototypes included the categoric feature, such as learning media. The model evaluation shows that K-Prototypes yield a model that can be identified better by the classification algorithm. The statistical analysis of the clustering result shows significant differences in most inter-cluster comparisons. These findings may bring positive implications where learners can be grouped, then those who might fail or achieve a low score in the evaluation can be identified, and actions be taken earlier to overcome such unwanted conditions.

On the other hand, K-Prototypes and especially K-Means have been known for decades. As machine learning research and adoptions are growing massively in the present, more recent methods could be adapted and applied to the learners' behaviors clustering cases as a plausible future work.

ACKNOWLEDGMENTS

This article is part of research funded by Budget Implementation Registration Form, Universitas Sam Ratulangi, Ministry of Education, Culture, Research, and Technology. The authors wish to express their gratitude to the Office of Research and Community Service Universitas Sam Ratulangi for funding this research.

REFERENCES

[1] "The rise of online learning during the COVID-19 pandemic | World Economic Forum. "https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/ (accessed Mar. 21, 2022).

[2] M. Aparicio, F. Bacao, and T. Oliveira, "An e-Learning Theoretical Framework," *J. Educ. Technol. Syst.*, vol. 19, no. 1, pp. 292–307, 2016, [Online]. Available: http://www.jstor.org/stable/jeductechsoci.19.1.292.

[3] P. Qiao, X. Zhu, Y. Guo, Y. Sun, and C. Qin, "The Development and Adoption of Online Learning in Pre- and Post-COVID-19: Combination of Technological System Evolution Theory and Unified Theory of Acceptance and Use of Technology," *J. Risk Financ. Manag. 2021, Vol. 14, Page 162*, vol. 14, no. 4, p. 162, Apr. 2021, doi: 10.3390/JRFM14040162.

[4] N. M. Almusharraf and S. H. Khahro, "Students Satisfaction with Online Learning Experiences during the COVID-19 Pandemic," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 21, pp. 246–267, Nov. 2020, doi: 10.3991/IJET.V15I21.15647.

[5] J. L. Moore, C. Dickson-Deane, and K. Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," *Internet High. Educ.*, vol. 14, no. 2, pp. 129–135, Mar. 2011, doi: 10.1016/j.iheduc.2010.10.001.

[6] A. A. Okaz, "Integrating Blended Learning in Higher Education," *Procedia - Soc. Behav. Sci.*, vol. 186, pp. 600–603, May 2015, doi: 10.1016/j.sbspro.2015.04.086.

[7] S. Hubackova and I. Semradova, "Evaluation of Blended Learning," *Procedia - Soc. Behav. Sci.*, vol. 217, pp. 551–557, Feb. 2016, doi: 10.1016/j.sbspro.2016.02.044.

[8] S. D. E. Paturusi, T. Usagawa, and A. S. M. Lumenta, "A study of students' satisfaction toward blended learning implementation in higher education institution in Indonesia," in *2016 International Conference on Information & Communication Technology and Systems (ICTS)*, 2016, pp. 220–225, doi: 10.1109/ICTS.2016.7910302.

[9] T. Usagawa and K. Ogata, "Potential of e-Learning for Enhancing Graduate and Undergraduate Education," *IPTEK J. Proceeding Ser.*, no. 1, pp. KS2-3-KS2-6, 2015, [Online]. Available: http://iptek.its.ac.id/index.php/jps/article/view/1115.

[10] S. D. E. Paturusi, Y. Chisaki, and T. Usagawa, "Development and Evaluation of the Blended Learning Courses at Sam Ratulangi University in Indonesia," *Int. J. e-Education, e-Business, e-Management e-Learning*, vol. 2, no. 3, pp. 242–246, 2012, doi: 10.7763/IJEEEE.2012.V2.118.

[11] H. B. Seta, T. Wati, A. Muliawati, and A. N. Hidayanto, "E-Learning Success Model: An Extention of DeLone & McLean IS' Success Model," *Indones. J. Electr. Eng. Informatics*, vol. 6, no. 3, pp. 281–291, Sep. 2018, doi: 10.52549/IJEEI.V6I3.505.

[12] C. T. Gozali, S. D. E. Paturusi, and A. M. Sambul, "Studi Preferensi Mahasiswa terhadap Jenis Media Pembelajaran Daring," *J. Tek. Inform.*, vol. 13, no. 4, pp. 39–46, 2018, Accessed: May 18, 2019. [Online]. Available: https://ejournal.unsrat.ac.id/index.php/informatika/article/view/24115.

[13] D. F. Sengkey, S. D. E. Paturusi, and A. M. Sambul, "Identifying Students' Pre-Classroom Behaviors in a Flipped Learning Environment," *J. Sustain. Eng. Proc. Ser.*, vol. 1, no. 2, pp. 143–149, Sep. 2019, doi: 10.35793/joseps.v1i2.19.

[14] D. F. Sengkey, S. D. E. Paturusi, and A. M. Sambul, "Perbandingan Akses Mahasiswa terhadap Media Pembelajaran Daring dalam Penerapan Flipped Classroom," *J. Tek. Elektro dan Komput.*, vol. 9, no. 1, pp. 31–38, Jun. 2020, doi: 10.35793/JTEK.9.1.2020.28634.

[15] D. F. Sengkey, S. D. E. Paturusi, and A. M. Sambul, "Correlations between Online Learning Media Types, First Access Time, Access Frequency, and Students' Achievement in a Flipped Classroom Implementation," *J. Sist. Inf.*, vol. 17, no. 1, pp. 44–57, Apr. 2021, doi: 10.21609/jsi.v17i1.1008.

[16] L. E. Sherman, M. Michikyan, and P. M. Greenfield, "The Effects of Text, Audio, Video, and In-person Communication on Bonding between Friends," *Cyberpsychology J. Psychosoc. Res. Cybersp.*, vol. 7, no. 2, Jul. 2013, doi: 10.5817/CP2013-2-3.

[17] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: A proof of concept," *Comput. Educ.*, vol. 54, no. 2, pp. 588–599, Feb. 2010, doi: 10.1016/j.compedu.2009.09.008.

[18] Y. Zhang, A. Ghandour, and V. Shestak, "Using Learning Analytics to Predict Students Performance in Moodle LMS," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 20, pp. 102–115, Oct. 2020, doi: 10.3991/IJET.V15I20.15915.

[19] G. Akçapınar, "Profiling Students' Approaches to Learning through Moodle Logs," in *Proceedings of*

Multidisciplinary Academic Conference on Education, Teaching and Learning (MAC-ETL 2015), 2015, pp. 9–15, Accessed: Jan. 17, 2022. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.719.9791&rep=rep1&type=pdf.

[20] S. S. Kusumawardani and S. A. I. Alfarozi, "Kajian Penggunaan Data Log Mahasiswa untuk Berbagai Permasalahan Analisis Pembelajaran," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 9, no. 4, pp. 365–374, Dec. 2020, doi: 10.22146/jnteti.v9i4.779.

[21] A. Charitopoulos, M. Rangoussi, and D. Koulouriotis, "Educational data mining and data analysis for optimal learning content management: Applied in moodle for undergraduate engineering studies," in *2017 IEEE Global Engineering Education Conference (EDUCON)*, Apr. 2017, pp. 990–998, doi: 10.1109/EDUCON.2017.7942969.

[22] Riandini, S. A. Aditya, R. N. Wardhani, and S. Setiowati, "Prediction of Digital Eye Strain Due to Online Learning Based on the Number of Blinks," *Indones. J. Electr. Eng. Informatics*, vol. 10, no. 2, pp. 452–462, Jun. 2022, doi: 10.52549/IJEEI.V10I2.3500.

[23] A. Aldholay, O. Isaac, A. N. Jalal, F. A. Anor, and A. M. Mutahar, "Towards a better understanding of the Organizational Characteristics that affect Acceptance of Big Data Platforms for Academic Teaching," *Indones. J. Electr. Eng. Informatics*, vol. 9, no. 3, pp. 766–773, Sep. 2021, doi: 10.52549/IJEEI.V9I3.2902.

[24] C. Li and J. Yoo, "Modeling student online learning using clustering," in *Proceedings of the 44th annual southeast regional conference on - ACM-SE 44*, 2006, vol. 2006, p. 186, doi: 10.1145/1185448.1185490.

[25] M. Köck and A. Paramythis, "Activity sequence modelling and dynamic clustering for personalized e-learning," *User Model. User-adapt. Interact.*, vol. 21, no. 1–2, pp. 51–97, Apr. 2011, doi: 10.1007/s11257-010-9087-z.

[26] J. Chen, K. Huang, F. Wang, and H. Wang, "E-learning behavior analysis based on fuzzy clustering," *3rd Int. Conf. Genet. Evol. Comput. WGEC 2009*, pp. 863–866, 2009, doi: 10.1109/WGEC.2009.214.

[27] M. A. Hogo, "Evaluation of e-learning systems based on fuzzy clustering models and statistical tools," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6891–6903, Oct. 2010, doi: 10.1016/J.ESWA.2010.03.032.

[28] K. R. Koedinger, K. Cunningham, A. Skogsholm, and B. Leber, "An open repository and analysis tools for fine-grained, longitudinal learner data," in *Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings*, 2008, no. May 2014, pp. 157–166.

[29] K. R. Koedinger, R. S. J. d. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper, "A Data Repository for the EDM Community: The PSLC DataShop," in *Handbook of Educational Data Mining*, CRC Press, 2010, pp. 65–78.

[30] M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study," *Int. J. Comput. Intell. Syst.*, vol. 5, no. 3, pp. 597–610, Jun. 2012, doi: 10.1080/18756891.2012.696923.

[31] A. Bovo, S. Sanchez, O. Heguy, and Y. Duthen, "Clustering moodle data as a tool for profiling students," in *2013 2nd International Conference on E-Learning and E-Technologies in Education, ICEEE 2013*, 2013, pp. 121–126, doi: 10.1109/ICELETE.2013.6644359.

[32] S. Liu and M. D'Aquin, "Unsupervised learning for understanding student achievement in a distance learning setting," in *2017 IEEE Global Engineering Education Conference (EDUCON)*, Apr. 2017, pp. 1373–1377, doi: 10.1109/EDUCON.2017.7943026.

[33] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, J. Vaclavek, and A. Wolff, "OU Analyse: analysing at-risk students at The Open University," Mar. 2015, Accessed: Oct. 25, 2022. [Online]. Available: http://www.laceproject.eu/learning-analyticsreview/analysing-at-risk-students-at-open-university/.

[34] A. Triayudi and I. Fitri, "A new agglomerative hierarchical clustering to model student activity in online learning," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 17, no. 3, pp. 1226–1235, Jun. 2019, doi: 10.12928/TELKOMNIKA.V17I3.9425.

[35] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means," *Am. J. Distance Educ.*, vol. 34, no. 2, pp. 137–156, Apr. 2020, doi: 10.1080/08923647.2020.1696140.

[36] K. Palani, "Identifying At-Risk Students in Virtual Learning Environment using Clustering Techniques," National College of Ireland, Dublin, 2020.

[37] K. Palani, P. Stynes, and P. Pathak, "Clustering Techniques to Identify Low-engagement Student Levels," in *Proceedings of the 13th International Conference on Computer Supported Education*, Apr. 2021, pp. 248–257, doi: 10.5220/0010456802480257.

[38] G. Nalli, D. Amendola, A. Perali, and L. Mostarda, "Comparative Analysis of Clustering Algorithms and Moodle Plugin for Creation of Student Heterogeneous Groups in Online University Courses," *Appl. Sci.*, vol. 11, no. 13, p. 5800, Jun. 2021, doi: 10.3390/app11135800.

[39] G. Ge *et al.*, "Analyzing Differences between Online Learner Groups during the COVID-19 Pandemic through K-Prototype Clustering," *J. Data Anal. Inf. Process.*, vol. 10, no. 1, pp. 22–42, Dec. 2021, doi:

10.4236/JDAIP.2022.101002.
[40] I. Dhaiouir, M. Ezziyyani, and M. Khaldi, "Smart Model for Classification and Orientation of Learners in a MOOC," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 05, pp. 19–35, Mar. 2022, doi: 10.3991/IJET.V17I05.28153.
[41] A. Talasbek, A. Serek, M. Zhaparov, S. Moo-Yoo, Y. K. Kim, and G. H. Jeong, "Personality Classification Experiment by Applying k-Means Clustering," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 16, pp. 162–177, Aug. 2020, doi: 10.3991/IJET.V15I16.15049.
[42] Z. Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values," in *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*, 1997, pp. 21–34.
[43] D. F. Sengkey, S. D. E. Paturusi, A. M. Sambul, and C. T. Gozali, "A Survey on Students' Interests toward On-line Learning Media Choices (A Case Study from the Operations Research Course in the Department of Electrical Engineering, UNSRAT)," *Int. J. Educ. Vocat. Stud.*, vol. 1, no. 2, pp. 146–152, Jun. 2019, doi: 10.29103/ijevs.v1i2.1527.
[44] D. F. Sengkey, A. M. Sambul, and S. D. E. Paturusi, "Penilaian Mahasiswa terhadap Jenis Media Pembelajaran dalam Penerapan Flipped Classroom," *J. Tek. Elektro dan Komput.*, vol. 8, no. 2, pp. 103–110, Aug. 2019, doi: 10.35793/JTEK.8.2.2019.25029.
[45] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Oct. 26, 2022. [Online]. Available: https://github.com/Microsoft/LightGBM.
[46] K. Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, B. Kerim, and R. Budiarto, "Important Features of CICIDS-2017 Dataset For Anomaly Detection in High Dimension and Imbalanced Class Dataset," *Indones. J. Electr. Eng. Informatics*, vol. 9, no. 2, pp. 498–511, May 2021, doi: 10.52549/IJEEI.V9I2.3028.
[47] M. J. Pendekal and S. Gupta, "An Ensemble Classifier Based on Individual Features for Detecting Microaneurysms in Diabetic Retinopathy," *Indones. J. Electr. Eng. Informatics*, vol. 10, no. 1, pp. 60–71, Mar. 2022, doi: 10.52549/IJEEI.V10I1.3522.
[48] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of Machine Learning Models Using Improved Shapley Additive Explanation," pp. 546–546, Sep. 2019, doi: 10.1145/3307339.3343255.

BIOGRAPHY OF AUTHORS

Daniel Febrian Sengkey is an Assistant Professor at the Undergraduate Program in Informatics, Department of Electrical Engineering, Faculty of Engineering, Universitas Sam Ratulangi, Manado-Indonesia. He graduated from the Undergraduate Program in Electrical Engineering of the same department in 2012. Later in 2015, he achieved his Master of Engineering degree from the Master Program in Electrical Engineering, under the Information Technology concentration, in the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta-Indonesia. His current research interest is in the implementation of Machine Learning in various fields.

Angelina Stevany Regina Masengi achieved her Bachelor of Medicine and Medical Doctor profession from the Faculty of Medicine in 2008 and 2010, respectively. She holds a master's degree in Biomedics, achieved in 2016 from the Master's Program in Biomedical Science. Since 2018, she is a tenured lecturer at Universitas Sam Ratulangi. Despite her assignment in the Department of Clinical Pharmacology and Therapy at the Faculty of Medicine, she is participating actively in teaching activities at several undergraduate programs, namely: Medicine, Nursing, Dentistry, and Pharmacy. She was also a member of the teaching team of the Bioinformatics course, in the Undergraduate Program in Informatics. She has an interest in behavioral studies.