Rank-Based Variable Minimization Using Clustering Algorithm

Allemar Jhone P. Delima and Virnille C. Francisco

***Abstract*—This paper introduces variable minimization with the advent of data mining techniques, particularly the cluster analysis using the K-means algorithm. The results of the simulation process validated the actual ranking of the variables in each category from the 407 records of student-respondents. A noticeable decrease of variables in determining the instructional performance of teachers in the Caraga Region, Philippines is evident in the simulation results. The variables that were clustered by the algorithm that conforms to have the lowest rank based on the actual survey was removed. Results showed that there were a total of fourteen (14) variables removed from the thirty-item survey questionnaire. In this case, it only proved that like any other data mining algorithms, the cluster analysis, particularly with the K-Means algorithm, is also effective in determining what variables in the data set to omit based on the groupings with the lowest rank it generated.**

***Index Terms*—Clustering, instructional performance, rank-based, K-Means, variable minimization.**

I. Introduction

Data preprocessing is considered to be one of the foremost techniques that are useful in the knowledge discovery process [1]. Data reduction, as an essential preprocessing technique in data mining, is performed by selecting and deleting unnecessary features and or variables. Maximized accuracy through the reduced number of attributes [2] and better understandability and interpretability of results are among the many benefits perceived in data reduction [3].

It is well known that in some instances, reducing the original training set or variables by selecting the most representative information is advisable [4]. Reducing the size of the dataset is beneficial in increasing the ability of generalization properties of the model. It also helped in reducing problems in space and computational time as well as lessening the size of formulas obtained by an induction algorithm on the reduced datasets [5].

In this paper, another technique of data reduction is shown. The proposed approach is the use of cluster analysis mainly with the K-Means algorithm to validate the survey conducted and then perform the data reduction in those clusters with variables who obtained the lowest rank in the survey.

II. Related Literature

In the era of data mining [6]-[8] and machine learning, data reduction procedures play vital importance [9]. It aims to obtain a fast, accurate, and adaptable model that quickly respond to incoming changes due to low computational complexity [3].

Different data mining techniques such as decision tree, attribute subset selections, clustering, data cube aggregation, and more are used for data reduction. [10]. Reducing samples called instance or horizontal reduction and reducing attribute/feature called vertical reduction are some of the many types of data reductions. Data reduction methods are dependent on the data mining methods used.

Feature selection algorithms such as Information Gain Ratio (IGR) attribute evaluation, Correlation-based Feature Selection (CFS), Symmetrical Uncertainty (SU) and Particle Swarm Optimization (PSO) algorithms were used to improve the performance accuracy of MLP, Simple Logistic, Rotation Forest, Random Forest and C4.5 prediction algorithms [11]. A notion that reducing original training set or variables by selecting the most representative information is advisable, yet obtaining nearly the same result or data-driven output [12]. Minimizing the size of the dataset aids in increasing the ability of generalization properties of the model. It also helped in lessening the space and computational time as well as minimizing the size of formulas used by the algorithm on the execution process [13]. Maximized accuracy through the reduced number of attributes [14] and better understandability and interpretability of results are among the many benefits perceived in data reduction [15].

The use of genetic algorithm for data reduction to enhance the accuracy of a hybrid K-means and C4.5 prediction model through the integration of the crossover-improved genetic algorithm in the quest to elevate the prediction rate for students leaving the university was conducted. Results showed that the hybrid prediction model integrating the modified genetic algorithm to k-means segmentation and c4.5 algorithms yielded a higher prediction rate [16].

A new crossover operator of the genetic algorithm called Cross Average Crossover was proposed. With the new crossover operator of the genetic algorithm, the variable reduction process on the dataset used has removed 90% of the variables converging to a global solution after ten generations compared to the traditional genetic algorithm which only removed 25%. The result showed that the reduction process after ten generations using the modified genetic algorithm outperformed the traditional genetic algorithm, but a degradation phenomenon [17] was depicted. The GA with CAX operator and rank-based selection function eliminated those individuals with higher fitness values due to the structure of its mating scheme [18].

Further, the imperialist competitive algorithm (ICA), which is a type of evolutionary algorithm (EA), was used to optimize the artificial neural network (ANN). The result

Manuscript received March 23, 2019; revised August 6, 2019.

A. J. Delima and V. Francisco are with the College of Engineering and Information Technology, Surigao State College of Technology, Surigao City, 8400 Philippines (e-mail: allemarjpd@ssct.edu.ph, vim2004@gmail.com).

doi: 10.18178/ijmlc.2019.9.5.842

showed that prediction using ICA-ANN model outperformed the ANN model when used alone [19]. Further, a model comprising the genetic algorithm and SVM for vaccine design was conducted. An average prediction with 93.50% accuracy on IEDB dataset was achieved while the accuracy of 95.125% was generated using Wang benchmark dataset [20]. Meanwhile, linear and nonlinear models using ARIMA and ANN respectively were used. The adaptation of hybrid methodology combining ARIMA and Deep Neural Network (DNN), which is an ANN model with multiple hidden layers, was considered as the optimal model for predicting roll motion compared to the non-hybrid models. It was found out that DNN-ARIMA hybrid model showed improved forecast accuracy and was identified to be very effective [21].

Furthermore, a study using the classification method in data reduction was conducted. The weighted k-nearest neighbor classifier was utilized to prove the importance of data reduction. After using uniform random sampling selection for data reduction in both direction-instances and attributes, accuracy is preserved, and an increase of execution time was observed [22].

Meanwhile, an algorithm called Edited Nearest Neighbor (ENN) for editing data was developed. The concept starts with $S = T$, where if instance in S does not agree with the majority of its k nearest neighbors, it is removed. Although, it does not reduce the data as much as other reduction algorithms do [23].

Since cluster analysis is considered as one of the most important techniques in data mining, it has attracted more and more attention in this big data era [24]. It recognizes patterns or clusters from a set of objects. In general, data clustering divides set of objects into groups or clusters where the objects in the same cluster or group are identical to each other than to objects from the other clusters [25].

This method is currently used in a wide array of application such that of government, education, market, health, and social science researches, image and pattern recognition and processing, web informatics, business, biology, geology, and other branches of science [26].

The use of K-Means segmentation technique and C4.5 algorithm to build a prediction model for customer loyalty in multimedia service provider was conducted. The integration of K-Means and C4.5 algorithm have yielded an increase of 79.33% accuracy prediction from the identified 69.23% accuracy with the C4.5 algorithm alone [27].

With the use of the K-Means clustering algorithm, the quest to eliminate the cluster with variables that are considered to be the lowest in rank as identified in the result of the actual survey is conducted.

III. Methodology

The experimental result for clustering was implemented using KNIME (Konstanz Information Miner) [28] analytics platform. Fig. 1 shows the node structure of the K-means clustering executed in KNIME. The node for the K-Means is connected and then positioned after the node of the imported CSV file of the dataset. The node color manager comes after as it put distinctions to the results to be generated later. The node scatter plot shows the scatter plot of the clusters while the interactive table is used to view the result in a table manner. In the obtained results throughout the use of the KNIME analytics platform in performing k-means clustering, it is denoted that Cluster_0 is equal to Cluster 1 and Cluster_1 is equal to Cluster 2.

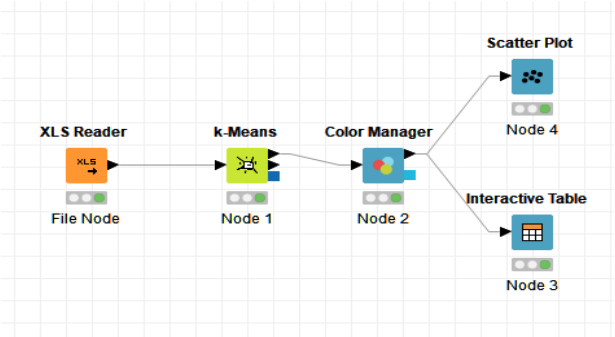

Fig. 1. Node structure of K-Means algorithm in KNIME.

A. Data

In this paper, a total of 407 records of student-respondents in the evaluation of the faculty instructional performance from the four State Universities and Colleges (SUC) in Caraga Region, Philippines were used as the datasets. There were thirty (30) variables that represent the faculty instructional performance having divided into six (6) parts viz., methodology, classroom management, student discipline, assessment of learning, student-teacher relationship, and peer relationship. Each category has five items. Each item are represented with a mean value obtained from the male and female student-respondents.

B. Clustering

Clustering algorithms divides the group of objects into clusters, where objects in each cluster are similar to each other [29]. K-means algorithm appears since 1965 and is by far the most used clustering algorithm due to simplicity in its implementation and effectiveness [30]. It is an unsupervised clustering with $N*k*d$ scheme. K-means algorithm uses Euclidean distance to group a set of n data points to k cluster. The steps of the K-means algorithm are the following:

1. Randomly select k centers to be used as initial centroids ($A_1, A_2, \ldots A_k$) from the datasets.
2. Assign each similar object centroids x_i to cluster C_j.
3. If all input data has been assigned, update the cluster centers to:

$$Z_j^* = \frac{1}{n_j} \sum_{X_j \in C_j} X_i \qquad \text{(a)}$$

where n_j is the number of data that belongs to cluster c_j.

4. If the result of the iteration is similar from the previous iteration, terminate the loop as the process of the k-means clustering has reached its stability.

Fig. 2 shows the graphical representation of the processes of the K-means algorithm from the concept of [31].

TABLE I: Indexed Dataset for Instructional Performance in Methodology

Items	Male	Female
1	3.273	3.132
2	3.407	3.323
3	3.407	3.393
4	3.387	3.315
5	3.28	3.222

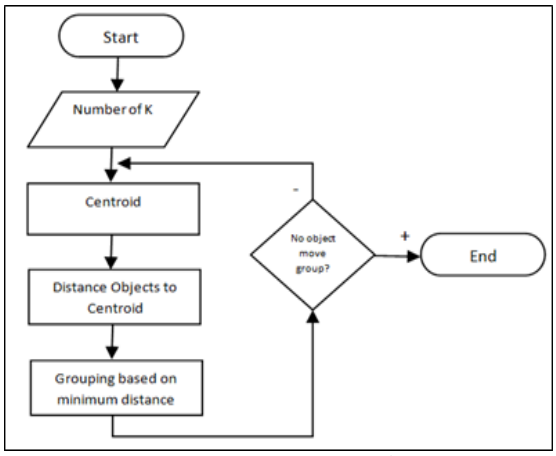

Fig. 2. K-means nearest neighbour flowchart.

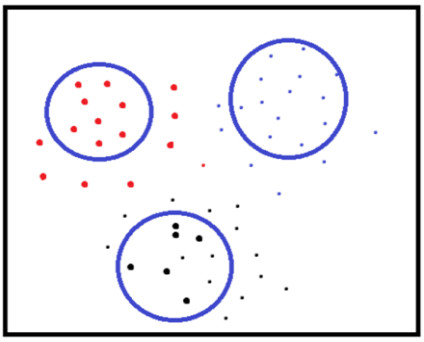
Fig. 3. Clustering scatter plot.

Table I shows the indexed datasets for the category Methodology. The first five items in the thirty-item survey questionnaire are represented by their mean value from the male and female respondents.

Row ID	D Male	D Female	S Cluster
1.0	3.273	3.132	cluster_0
2.0	3.407	3.323	cluster_1
3.0	3.407	3.393	cluster_1
4.0	3.387	3.315	cluster_1
5.0	3.28	3.222	cluster_0

Fig. 4. Clustering result for instructional performance in methodology.

TABLE II: FACULTY INSTRUCTIONAL PERFORMANCE IN METHODOLOGY AS PERCEIVED BY THE STUDENT-RESPONDENTS

Statement	Rank
1. utilizes varied designs/ techniques/ activities suited to the different types of learners.	5
2. explains learning goals and instructional procedures to the students.	2
3. uses real-life examples in the class to sustain students' interest in learning.	1
4. creates a situation that encourages students to use critical thinking.	3
5. delivers accurate/ relevant/ updated content knowledge.	4

Fig. 4 shows the clustering result of each variable under the category methodology performed in KNIME. Variables 1 and 5 belong to Cluster 1, while the rest belongs to Cluster 2. The result of K-Means algorithm proved that there is a similarity in the attributes between variables 1 and 5 and it conforms to the result in the survey questionnaire deployed that both items obtained the least rating from the student respondents. Table II shows that the least rating of the student respondent went to the items 1 and 5, making them the 5th and 4th in rank, respectively. In general, since items under Cluster 1 has the least rating, these variables will be removed. It also means that items under Cluster 2 are the most effective methodologies employed in the delivery of instruction as perceived by the student respondent.

TABLE III: INDEXED DATASET FOR INSTRUCTIONAL PERFORMANCE IN CLASSROOM MANAGEMENT

Items	Male	Female
1	2.987	3.012
2	2.62	2.58
3	3.027	2.922
4	3.413	3.171
5	3.267	3.276

Table III shows the indexed datasets for the category classroom management. The next five items in the thirty-item survey questionnaire are represented by their mean value from the male and female respondents.

Row ID	D Male	D Female	S Cluster
1.0	2.987	3.012	cluster_0
2.0	2.62	2.58	cluster_1
3.0	3.027	2.922	cluster_0
4.0	3.413	3.171	cluster_0
5.0	3.267	3.276	cluster_0

Fig. 5. Clustering result for instructional performance in classroom management.

Fig. 5 shows that variables 1, 3, 4, and 5 under the category classroom management belongs to Cluster 1, while variable 2 belongs to Cluster 2. The result of the K-means algorithm proved that there is a similarity in the attributes between the variables in Cluster 1. It also conforms to the result in the survey questionnaire deployed that variable 2 in Cluster 2 obtained the least rating from the student respondents making it the 5th in rank as shown in Table IV. Therefore, the item under Cluster 2 will be removed. It also means that items under Cluster 1 are the most effective classroom management scheme in the delivery of instruction as perceived by the student respondent.

TABLE IV: FACULTY INSTRUCTIONAL PERFORMANCE IN CLASSROOM MANAGEMENT AS PERCEIVED BY THE STUDENT-RESPONDENTS

Statement	Rank
1. establishes routines to maximize instructional time	3
2. organizes and assign the daily cleaners.	5
3. employs an effective system of the classroom set-up.	4
4. employs strategies to maximize the use of resources in learning activities.	2
5. implements rules/ policies inside the classroom	1

TABLE V: INDEXED DATASET FOR INSTRUCTIONAL PERFORMANCE IN STUDENT DISCIPLINE

Items	Male	Female
1	3.233	3.249
2	3.14	3.128
3	3.407	3.482
4	3.467	3.497
5	3.293	3.327

Table V shows the indexed datasets for the category of student discipline. The third set of five variables in the thirty-item survey questionnaire are represented by their mean value from the male and female respondents.

Fig. 6 shows the clustering result of each variable under the category student discipline performed in KNIME.

Variables 3, 4, and 5 belongs to Cluster 1 while the rest belongs to Cluster 2. The result of the K-means algorithm proved that there is a similarity in the attributes between variables 1 and 2. It also conforms to the result in the survey questionnaire deployed that both items obtained the least rating from the student respondents. Table VI shows that the least rating made by the student respondents went to item 1 and item 2, making them the 4th and 5th in rank, respectively. Hence, variables 1 and 2 under Cluster 2 will be removed. It also means that variables under Cluster 1 are the most effective strategies in student discipline as perceived by the student respondent.

Row ID	D Male	D Female	S Cluster
1.0	3.233	3.249	cluster_1
2.0	3.14	3.128	cluster_1
3.0	3.407	3.482	cluster_0
4.0	3.467	3.479	cluster_0
5.0	3.293	3.327	cluster_0

Fig. 6. Clustering result for instructional performance in student discipline.

TABLE VI: FACULTY INSTRUCTIONAL PERFORMANCE IN STUDENT DISCIPLINE AS PERCEIVED BY THE STUDENT-RESPONDENTS

Statement	Rank
1. handles behavior problems with respect to the student's rights.	4
2. imposes disciplinary sanction (s) to the misbehaving student (s).	5
3. encourages student to submit requirement on time.	2
4. motivates students to respect each other.	1
5. allows students to exercise their own creativity.	3

TABLE VII: INDEXED DATASET FOR INSTRUCTIONAL PERFORMANCE IN ASSESSMENT OF LEARNING

Items	Male	Female
1	3.054	3.163
2	3.133	3.089
3	3.293	3.25
4	3.167	3.058
5	3.087	3.086

Table VII shows the indexed datasets for the category assessment of learning. The fourth set of five variables in the thirty-item survey questionnaire are represented by their mean value from the male and female respondents.

Row ID	D Male	D Female	S Cluster
1.0	3.054	3.163	cluster_0
2.0	3.133	3.089	cluster_0
3.0	3.293	3.25	cluster_1
4.0	3.167	3.058	cluster_0
5.0	3.087	3.086	cluster_0

Fig. 7. Clustering result for instructional performance in assessment of learning.

Fig. 7 shows that variables 1, 2, 4, and 5 under the category assessment of learning belongs to Cluster 1 and is distant from variable 3 which belongs to Cluster 2. It also conforms to the result in the survey questionnaire deployed that variable 3 in Cluster 2 obtained the highest rating from the student respondents. The K-means algorithm grouped the variables 1, 2, 4, and 5 in one cluster and was ranked as 2nd, 3rd, 4th, and 5th, respectively, as shown in Table VIII. Therefore, variables under Cluster 1 will be removed. It also means that variable 3 is the most effective assessment of learning scheme as perceived by the student respondent.

Table IX shows the indexed datasets for the category student-teacher relationship. The fifth set of five variables in the thirty-item survey questionnaire are represented by their mean value from the male and female respondents.

TABLE VIII: FACULTY INSTRUCTIONAL PERFORMANCE IN ASSESSMENT OF LEARNING AS PERCEIVED BY THE STUDENT-RESPONDENTS

Statement	Rank
1. constructs a valid and reliable formative and summative tests.	2
2. uses appropriate non-traditional assessment techniques and tools (i.e. journals, rubrics, etc)	3
3. interprets and use test results to improve teaching and learning.	1
4. uses tools for assessing authentic learning.	4
5. provides timely and accurate feedback to students.	5

TABLE IX: INDEXED DATASET FOR INSTRUCTIONAL PERFORMANCE IN STUDENT-TEACHER RELATIONSHIP

Items	Male	Female
1	3.433	3.362
2	3.26	3.245
3	3.4	3.233
4	3.28	3.261
5	3.207	3.16

Row ID	D Male	D Female	S Cluster
1.0	3.433	3.362	cluster_0
2.0	3.26	3.245	cluster_1
3.0	3.4	3.233	cluster_0
4.0	3.28	3.261	cluster_1
5.0	3.207	3.16	cluster_1

Fig. 8. Clustering result for instructional performance in the student-teacher relationship.

Fig. 8 shows the clustering result of each variable under the category student-teacher relationship performed in KNIME. Variables 1 and 3 belong to Cluster 1. The K-means algorithm grouped the variables 2, 4, and 5 and is assigned in Cluster 2. In the survey questionnaire deployed, they were ranked as 4th, 3rd, and 5th, respectively as shown in Table X. Meanwhile, variables 1 and 3 under Cluster 1 topped and seconded the rank. Hence, variables 2, 4, and 5 under Cluster 2 will be removed. It also means that variables under Cluster 1 are the most effective strategies in establishing a student-teacher relationship as perceived by the student respondent.

TABLE X: FACULTY INSTRUCTIONAL PERFORMANCE IN STUDENT-TEACHER RELATIONSHIP AS PERCEIVED BY THE STUDENT-RESPONDENTS

Statement	Rank
1. encourages students to actively participate in the class/school activities.	1
2. allows students to communicate directly to him/her.	4
3. provides equal opportunities for all students.	2
4. promote teamwork among students.	3
5. makes him/herself available to students.	5

TABLE XI: INDEXED DATASET FOR INSTRUCTIONAL PERFORMANCE IN PEER RELATIONSHIP

Items	Male	Female
1	3.433	3.362
2	3.26	3.245
3	3.4	3.233
4	3.28	3.261
5	3.207	3.16

Table XI shows the indexed datasets for the category peer relationship. The sixth and the last set of five variables to complete the thirty-item survey questionnaire are represented by their mean value from the male and female respondents.

Fig. 9 shows that variables 1, 4, and 5 under the category peer relationship belongs to Cluster 1, and variables 2 and 3 belongs to Cluster 2. It also conforms to the result in the survey questionnaire deployed that variable 2 and 3 in Cluster 2 obtained the least rating from the student respondents making them the 5th and 4th in rank respectively as shown in Table XII. Therefore, variables 2 and 3 will be removed. It also means that items under Cluster 1 are the most effective scheme in establishing peer relationship as perceived by the student respondent.

Row ID	D Male	D Female	S Cluster
1.0	3.14	3.218	cluster_0
2.0	3.127	3.078	cluster_1
3.0	3.1	3.124	cluster_1
4.0	3.207	3.319	cluster_0
5.0	3.22	3.226	cluster_0

Fig. 9. Clustering result for instructional performance in peer relationship.

TABLE XII: FACULTY INSTRUCTIONAL PERFORMANCE IN PEER RELATIONSHIP AS PERCEIVED BY THE STUDENT-RESPONDENTS

Statement	Rank
1. demonstrate appropriate behavior in dealing with students/peers/superiors.	3
2. manifest flexibility when deemed necessary.	5
3. exhibit collegiality with colleagues.	4
4. observe professionalism at all times.	1
5. empathize other needs and concern.	2

TABLE XIII: SUMMARY OF SIMULATION RESULTS USING K-MEANS ALGORITHM

Category	Number of Existing Variables	Number of Variables Removed	Variables/ Items Removed
1	5	2	1,5
2	5	1	2
3	5	2	1,2
4	5	4	1,2,4,5
5	5	3	2,4,5
6	5	2	2,3

Table XIII presents the simulation result having listed the number of variables removed in each category.

IV. CONCLUSIONS

The K-means clustering algorithm proved to be an effective mechanism in optimizing variables in the datasets. The variables that have the highest and lowest mean value were clustered correctly. The results of the simulation were compared and are matched to the actual ranking of every item in each category on the survey questionnaire deployed. Those clusters with the lowest rank were removed. It is showed that category four (4) has the most number of variables removed as evident in the centroid distance determined by the algorithm.

REFERENCES

[1] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016.

[2] A. Baldominos, P. Isasi, and U. C. I. I. I. De Madrid, "Feature set optimization for physical activity recognition using genetic algorithms," in *Proc. Companion Publ. 2015 Genet. Evol. Comput. Conf. - GECCO Companion '15*, 2015, pp. 1311–1318.

[3] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017.

[4] S. García, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 417–435, 2012.

[5] J. R. Cano, F. Herrera, and M. Lozano, "On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining," *Appl. Soft Comput. J.*, vol. 6, no. 3, pp. 323–332, 2006.

[6] A. J. P. Delima, "Applying data mining techniques in predicting index and non-index crimes," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 4, pp. 533–538, 2019.

[7] A. J. P. Delima and M. T. Q. Lumintac, "Application of time series analysis for philippines ' inflation prediction," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 1761–1765, 2019.

[8] A. J. P. Delima, "Predicting scholarship grants using data mining techniques," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 4, pp. 513–519, 2019.

[9] I. Czarnowski and J. Piotr, *Data Reduction Algorithm for Machine Learning and Data Mining*, pp. 276–285, 2008.

[10] P. Vora and B. Oza, "Improved data reduction technique in data mining," *Int. J. Comput. Sci. Mob. Comput.*, vol. 2, pp. 169–174, 2013.

[11] H. Almayan and W. Al Mayyan, "Improving accuracy of students' final grade prediction model using PSO," in *Proc. 6th International Conference on Information Communication and Management*, pp. 35–39, 2016.

[12] A. J. P. Delima, A. M. Sison, and R. P. Medina, "A modified genetic algorithm with a new crossover mating scheme," *Indones. J. Electr. Eng. Informatics*, vol. 7, no. 2, pp. 165–181, 2019.

[13] A. J. P. Delima, "An experimental comparison of hybrid modified genetic algorithm-based prediction models," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 1756–1760, 2019.

[14] A. J. P. Delima, A. M. Sison, and R. P. Medina, "Variable reduction-based prediction through modified genetic algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 356–363, 2019.

[15] U. O. Cagas, A. J. P. Delima, and T. L. Toledo, "PreFIC : Predictability of faculty instructional performance through hybrid prediction model," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 7, pp. 22–25, 2019.

[16] M. Y. Orong, A. M. Sison, and R. P. Medina, "A hybrid prediction model integrating a modified genetic algorithm to k-means segmentation and C4.5," in *Proc. TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, pp. 1853–1858.

[17] Y. Fan and D. Chen, "Application of improved adaptive genetic algorithm in train energy saving," in *Proc. 4th International Forum on Decision Sciences, Uncertainty and Operations Research*, 2017, pp. 723–736.

[18] M. Y. Orong, A. M. Sison, and R. P. Medina, "A new crossover mechanism for genetic algorithm with rank-based selection method," in *Proc. 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science*, 2018, pp. 83–88.

[19] D. J. Armaghan, M. Hasanipanah, and E. T. Mohamad, "A combination of the ICA - ANN model to predict air - overpressure resulting from blasting," *Eng. Comput.*, 2015.

[20] B. A. Moghram, E. Nabil, and A. Badr, "Ab-initio conformational epitope structure prediction using genetic algorithm and SVM for vaccine design," *Comput. Methods Programs Biomed.*, 2017.

[21] N. Suhermi, Suhartono, D. D. Prastyo, and B. Ali, "Roll motion prediction using a hybrid deep learning and ARIMA model," *Procedia Comput. Sci.*, vol. 144, pp. 251–258, 2018.

[22] F. Patel, "Large high dimensional data handling using data reduction," in *Proc. International Conference on Electrical, Electronics, and Optimization Techniques*, 2016, vol. 7, pp. 1531–1536.

[23] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst. Man Cybern.*, vol. 2, no. 3, pp. 408–421, 1972.

[24] C. Jinyin, L. Xiang, Z. Haibing, and B. Xintong, "A novel cluster center fast determination clustering algorithm," *Appl. Soft Comput. J.*, vol. 57, pp. 539–555, 2017.

[25] G. Gan and M. K. P. Ng, "K-Means clustering with outlier removal," *Pattern Recognit. Lett.*, vol. 90, pp. 8–14, 2017.

[26] [26] A. Bansal, M. Sharma, and S. Goel, "Improved K-Mean clustering algorithm for prediction analysis using classification technique in data mining," *Int. J. Comput. Appl.*, vol. 157, no. 6, pp. 35–40, 2017.

[27] S. Moedjiono, Y. R. Isak, and A. Kusdaryono, "Customer loyalty prediction in multimedia service provider company with K-Means segmentation And C4.5 algorithm," in *Proc. 2016 International Conference on Informatics and Computing*, pp. 1–6.

[28] L. Feltrin, "KNIME an open source solution for predictive analytics in the geosciences [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 4, 2015.
[29] D. Hand, D. Hand, H. Mannila, H. Mannila, P. Smyth, and P. Smyth, *Principles of Data Mining*, vol. 30, 2001.
[30] L. Morissette and S. Chartier, "The K-Means clustering technique: General considerations and implementation in Mathematica," *Tutor. Quant. Methods Psychol.*, vol. 9, no. 1, pp. 15–24, 2013.
[31] M. Y. Orong, A. M. Sison, and A. A. Hernandez, "Mitigating vulnerabilities through forecasting and crime trend analysis," in *Proc. 2018 5th Int. Conf. Bus. Ind. Res.*, 2018, pp. 57–62.

Allemar Jhone P. Delima was born in Surigao City, province of Surigao del Norte, Philippines. He is a faculty under the College of Engineering and Information Technology at Surigao State College of Technology. He received his B.S. degree from Surigao State College of Technology, Surigao City, in 2014 and his master's degree from the same institution in 2016, both in information technology. He is currently pursuing his doctor in information technology degree in the Technological Institute of the Philippines, Quezon City. His research interest is in data mining, data analytics, and machine learning.

Virnille C. Francisco was born in Surigao City, province of Surigao del Norte, Philippines. She is a graduate with the bachelor of science in computer science, then pursued master degree in information technology and eventually obtained a doctorate degree in education major in technology management (phEdD). She is a licensed professional teacher and is currently connected at the College of Engineering and Information Technology of Surigao State College of Technology.