Research Article

University Students Behaviour Modelling Using the K-Prototype Clustering Algorithm

Delali Kwasi Dake, Esther Gyimah, and Charles Buabeng-Andoh

Department of ICT Education, University of Education, P.O. Box 25, Winneba, Ghana

Correspondence should be addressed to Delali Kwasi Dake; dkdake@uew.edu.gh

Received 26 January 2023; Revised 28 June 2023; Accepted 14 July 2023; Published 7 August 2023

Academic Editor: Venkatesan Rajinikanth

Copyright © 2023 Delali Kwasi Dake et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Counselling students remains a pre-eminence for most tertiary institutions in Ghana to the extent that institutions now have counselling units that extend to the departmental level. This study used the K-prototype machine learning algorithm to cluster students’ behaviour based on 28 relevant attributes and further proposed a classification model. The analysis of the experimental outcomes using the elbow method reveals the formation of three distinct clusters with decreasing intra-cluster similarities and increasing inter-cluster distances. The first cluster uniquely consists of active learners with three or more roommates, primarily in the first year. The second cluster with the highest membership consists mainly of second-year students who exhibit passive classroom conduct and reside in a two-occupancy hostel. The third cluster contains a mixture of third and final-year students who are highly passive in class and live in a tenancy occupancy of two. After clustering, the K-nearest neighbours, logistic regression, naive Bayes (NB), and AdaBoost ensemble algorithms were implemented to create a model for future learner cluster prediction. Simulation results using the tenfold cross-validation technique show that AdaBoost (NB) has the highest accuracy of 99.88% with an F-measure score of 0.999 and receiver operating characteristic–area under the curve value of 1.00.

1. Introduction

Education has become the priority of many continents, including Africa. Over the years, the development of many countries has been centred on educating their citizenry, which is critical for human capital development [1, 2]. In the last two decades, global access to higher education has increased from 19% to 38%, according to a report by the United Nations Educational, Scientific, and Cultural Organization [3, 4]. The rising enrolment across institutions has enormous benefits, but it also presents a challenge in students’ behaviour modelling and counselling. In order to create a school environment devoid of crime and unsavoury characters, it is essential to have a comprehensive understanding of students’ behaviour and patterns [5]. Positive behaviour fosters a productive learning environment with successful learning outcomes, whereas negative behaviour disrupts the learning process and ultimately results in lower grades [6]. Bicard et al. [7] define behaviour as something that can be observed, measured, and repeated by a person. Bergner [8] defines the eight parameters of behaviour. These parameters include the want, know, know-how, performance, achievement, personal characteristics, identity, and significance. Kwasnicka et al. [9] alluded to Bergner [8] behaviour parameter definitions but recommended innovative approaches to understanding human behaviour and changes.

Enrollment at tertiary institutions in Ghana continues to rise in line with the global trend [4]. According to a report by EducationWeb [10], Ghana’s implementation of a free senior high school policy in 2017 increased enrolment by 20% across tertiary institutions. As learner enrollments continue to rise, monitoring students’ conduct has become necessary. The counselling department of every tertiary institution in Ghana has a primary goal to encourage students to behave well and ultimately improve their lives [11]. In addition, guidance and counselling equip learners to discover their interests, goals, habits, skills, and values [12]. Without counselling, students become frustrated, confused, suicidal, and disappointed, especially regarding low grades, relationships, and family [13]. The behaviourism study approach in the counselling unit across tertiary institutions in Ghana is outdated, with far-reaching effects on learners [14, 15].
Intelligent student behaviour pattern discoveries based on data with predictive outcomes will provide proactive counselling, as opposed to the reactive approaches still prevalent in an era defined by machine learning (ML).

Growth and innovation in education are critical even as learner enrolment continues to rise globally. Education 4.0, which aligns with the fourth industrial revolution (Industry 4.0), has integrated technology and paradigms to transform the future of education. Successful learning outcomes of university students are relevant to Industry 4.0 with autonomous and cyber-physical systems [16]. Behavioural modelling in Education 4.0 has bases in active and collaborative learning, and cyber-physical systems [16]. Behavioural modelling in early life can influence later success [25]. The agricultural domain with ML has resulted in harvest intelligent engineering with post-harvest prediction [26,27]. The smart grid sector has seen ML deployment mainly in load forecasting, fault detection, energy optimisation, and security identification [28,29]. Other related application domains of ML, such as driverless cars, sports, computer vision, image recognition, and social media [30], have contributed to improving life and technological advancements.

Recently, ML application in the educational sector has taken a positive trend. Educational data mining (EDM) relates to the application of ML algorithms to educational data and covers aspects, including classification, clustering, regression, and reinforcement learning (RL) [31]. Literature review [32–34] shows that the most prevailing research in EDM involves predicting students’ academic performance, intelligent tutoring systems, feedback modelling, assessment modelling, and retention modelling. Students’ behaviour modelling and counselling remain a topical EDM research area, but cluster analytics and predictive modelling in EDM are limited.

The primary objective of this study is to cluster learners’ behavioural traits in a traditional campus and expose hidden patterns in each cluster for proactive counselling. Clustering or unsupervised learning is an ML technique that involves grouping data points. Each cluster or grouping should have distinctive features that are highly dissimilar to those of other clusters [35]. Clustering techniques minimise intra-cluster similarity while maximising inter-cluster similarity for high cluster cohesiveness [35]. In addition, a classification model for students cluster groupings was built by comparing conventional and ensemble algorithms. In line with the objectives, the study was guided by four research questions:

1. What is the appropriate number of clusters from the numeric and categorical data set for student categorisation?
2. To what extent do cluster groupings vary regarding members and the likelihood of a student belonging to a cluster?
3. What are the similarities and distinct features of each cluster, and what factors indeed influenced the cluster groupings?
4. Which classification algorithm has the highest performance metric for the futuristic prediction of learner clusters?

The main contributions of the study are as follows:

1. A novel study on the conventional campus behaviour of university students based on 28 relevant attributes using ML algorithms.
2. Comparison between the fivefold and the tenfold cross-validation techniques in building a classification model for students behaviour prediction.
3. A summarised review of the studies involving student behaviour, the algorithms utilised, and the major limitations of each study.

The rest of the paper is organised as follows: Section 2 discusses relevant literature. Section 3 examines the methodological procedure. Section 4 deals with the cluster simulation and results. Section 5 discusses the findings and compares the results to existing literature. In Section 6, we conclude by outlining the study’s relevance to practitioners, its limitations, and future research.

2. Review of Literature

One of the problems of higher education institutions (HEIs) is estimating students’ academic success rates. This is critical because academic success is often used as a criterion to assess the quality of HEIs, which ultimately can improve enrolment rates [36]. With the recent COVID-19 pandemic, many HEIs were challenged to migrate to e-learning systems that required remote and online teaching and assessments.
Students’ academic success was put to the test due to students’ behaviour, such as academic integrity violations, accessibility, and motivation towards academic work on e-learning platforms [37–39]. To this end, predicting students’ academic success has remained a challenge for HEIs. According to Alyahyan and Düşetgör [36], psychological attributes such as students’ interests and personal behaviour are among the factors that affect students’ academic success. Analysing students’ behaviour using intelligent methods and predicting students’ academic success is of great significance in identifying potentially deviant students [40] who need academic counselling. Even though students’ behaviour modelling during online studies is crucial, especially in a pandemic era, traditional campus behaviour modelling of students is equally vital since learning in a physical setting has significant advantages.

Related research on behaviour modelling in predicting students’ academic success shows extensive use of ML techniques in online studies [41–44].

2.1. Students Groupings Using Clustering. Clustering has been widely used in analysing students’ behaviour. Moubayed et al. [18] used K-means clustering on 486 numeric data instances to observe online student learning patterns and behaviour. The engagement metrics of their questionnaire were divided into interaction-related and effort-related attributes. Using the WEKA tool to run the K-means, three cluster models, including 2-level, 3-level, and 5-level, were tested using the silhouette coefficient. Even though the 2-level cluster model resulted in the highest performance in the silhouette coefficient value, the 3-level model was more effective in identifying students with low engagement levels. The experimental results of their study indicate that the number of logins and the average duration to submit assignments have a more significant impact on students’ engagement than the number of contents read and accessed. In behavioural modelling of the mental health state of learners on campus, Srividya et al. [45] examined 656 data instances using K-means clustering to determine the optional cluster groupings for each category. The study used the K-mode clustering algorithm on learners’ sequences with categorical values. The learners’ sequences were grouped into 16 clusters based on the Felder and Silverman Learning Style Model. The clustering produced a yes or no sequence of categorisations for students’ learning styles. In analysing students’ behaviour online during Covid-19, Ge et al. [19] implemented the K-prototype algorithm on 5,015,344 instances of numeric and categorical data. The silhouette coefficient optimally selected two clusters as the inflexion point of the data set. The study also utilised non-parametric statistics and reported the Kruskal–Wallace $H$ tests and $p$-values for the aspects of the data set that are non-normal after clustering. The experimental findings demonstrate that students with superior learning media exhibit better online learning behaviour and results. Second, students with inferior learning media outnumber those with better learning media in economically developed and underdeveloped regions. In a similar study, Liu and d’Aquin [20] used the K-prototype clustering algorithm to classify students based on demographic characteristics and online engagement and how these can influence their learning achievement. The data set included categorical—students’ demographic data and numeric data—students’ behaviour data. The elbow method was applied to determine seven clusters. The outcomes demonstrated that the K-prototype algorithm was effective at grouping students with similar characteristics and levels of learning achievement so that solutions can be tailored to their different needs. Another study by Palani et al. [21] used cluster analysis to group students with similar online behaviour in virtual learning environments (VLEs) to better identify students with low engagement at the early stages of a course. The study used three clustering algorithms; Gaussian mixture, hierarchical, and K-prototype clustering algorithms. For the clustering models, both categorical and numeric data were used, and the models were implemented in Python 3.7 using Jupyter Notebook and Scikit-learn libraries. The data set was executed with three clusters to group the students with similar online behaviour in the VLE. The results showed that the K-prototype algorithm clustered the low-engagement students better with highly partitioned clusters than the other algorithms. Mingyu et al. [46] proposed a prediction method that used K-prototype clustering on 13,613 students’ data from a University in China. The data consisted of categorical and numerical attributes collected from four perspectives: basic information, study behaviour, internet behaviour, and living behaviour of the students. The K-prototype clustering algorithm and the CatBoost-SHAP-based model for prediction were applied on the data set. The experimental results showed that the K-prototype algorithm is very effective for clustering datasets that are both categorical and numerical when compared to other clustering algorithms. According to the findings, the dormitory environment, breakfast time, awards, and good reading habits contributed to students’ high grades. Asif et al. [47] applied the X-means clustering with Euclidean distance to determine high-performing and low-performing students based on the courses that are high and low indicators. They realised that, throughout the 4 years of undergraduate studies, they mostly stayed in the same cluster. Krizanić [48] applied K-means clustering to understand students learning behaviour from an e-learning environment log data. This resulted in the formation of three clusters with varying members. The first group is learners with minimum access to course content, while the second group has medium access. In contrast, the third group has high access to the e-learning course content.

2.2. Predictive Modelling of Cluster Groups Using Classification. After clustering, predictive modelling of students’ behaviour is relevant to determine the right cluster groupings of new students for early academic advice and counselling. Even though limited studies performed classification after clustering, implementing classification algorithms creates a predictive model for futuristic learner groupings and personalised counselling [43]. Hussain et al. [49] concluded that the K-nearest neighbour (KNN) is more efficient than other supervised algorithms for predicting students’ behaviour groupings.
after clustering. Al-Shehri et al. [50] compared KNN and support vector machine (SVM) for students’ grades and behaviour modelling. The SVM algorithm was slightly more efficient as a predictive model with a higher correlation coefficient of 0.96 when compared to the 0.95 coefficient value of the KNN algorithm. Srividya et al. [45], after clustering, implemented a predictive model using decision tree (DT), SVM, logistic regression (LR), KNN, naive Bayes (NB), and the ensemble. The ensemble algorithm performed with the highest accuracy of 90. After implementing the X-means clustering, Asif et al. [47] proposed a predictive model by comparing nine classification algorithms. The NB outperformed other classifiers with an accuracy of 83.65. Krizanic [48], after clustering, implemented the DT algorithm for futuristic categorisation.

Table 1 provides a summary of the literature review, indicating the algorithms used by each study, the best classification algorithm, and the major limitation for each study.

### 3. Methodology

As depicted in Figure 1, this study modified the knowledge discovery in databases (KDD) [51] methodology by adding the clusters, the K-prototype, and classification modules. The KDD process sequentially consists of the stages: data selection, preprocessing, transformation, ML algorithms/K-prototype, knowledge discovery, and prediction.

#### 3.1. Data Selection.

Approval was sort from the University of Education, Winneba Research Ethics Committee to collect data from the students. The authors can confirm that the
The study complied with all ethical regulations and students’ consent was duly obtained. In addition, students were guaranteed that their data would be kept private and anonymous. Using the convenience sampling method, data was collected from students in the Department of Information and Communication Technology at the University of Education, Winneba. The convenience sampling method is a non-probability technique that selects sample members based on availability, accessibility, and economic conditions [52]. Students in years 1–4 completed the Google Forms. The study’s objectives were clearly stated on the questionnaire given to participants, and respondents were asked to agree to an ethics consent form. The students’ privacy and non-disclosure policies were strictly enforced throughout the KDD process. A total of 913 responses categorised under personal biodata, family life, senior high school (SHS) tracker, and university tracker were retrieved.

### 3.1.1. Attribute Description

Each attribute for clustering has relevance in students’ behaviour modelling. As shown in Table 2, the attributes were logically segmented under sections peculiar to learners’ development in family and academic environments.

### 3.2. Data Preprocessing

In the data preprocessing phase, noisy, inconsistent, missing values, and unrelated data were removed from the dataset. After cleaning the data, 905 out of 913 responses were ready for cluster model construction. In percentage terms, 99.12% of valid data were consistent with the study’s objectives.

### 3.3. Data Transformation

The preprocessed data is converted to comma-separated values (CSV) in Excel. The CSV data file now contains characters separated by commas. The final stage of data transformation is the uploading of the CSV file into the Jupyter Notebook platform. The Jupyter Notebook is a server–client application that allows the configuration and running of a data science project.


K-means is a centroid-based partitioning algorithm that clusters unlabelled data into discrete clusters [53]. K-means form the basis for K-modes and K-prototype algorithms. In K-means, n instances of data are partitioned into k clusters where $k < n$, and each data instance is assigned to the cluster with the nearest mean. The K-means algorithm strives to reduce the squared error objective function to improve the intra-cluster and decrease the inter-cluster similarities [53]. Similar to the Gaussian expectation-maximisation algorithm, the K-means algorithm locates cluster centroids in data using Euclidean distance between the two data points [54].

As depicted in the process flow diagram in Figure 2, the K-means algorithm uses a series of algorithmic steps to assign data points to a cluster. The process flow is defined in Algorithm 1.

In measuring the similarity between objects, the K-means uses the Euclidean distance [55] to measure the distance between two data points, as shown in Equation (1).

$$\text{Dist}_{ij} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$  

In Equation (1), $q$ and $p$ are sample data points in an $n$-dimensional feature space, resulting in a multivariate mean in the Euclidean space. The K-means algorithm is limited to numeric data sets and performs poorly with categorical values. Clustering involving categorical values utilises the K-modes algorithm [56]. The K-modes algorithm in modifying K-means for categorical data set uses a simple matching dissimilarity measure [57] and replaces the cluster means by modes.

In the questionnaire administered to participants, both numeric and categorical attributes were considered. The study, therefore, implemented the K-prototype algorithm on the data set. The K-prototype algorithm integrates the K-means and the K-mode algorithms for mixed-type objects.
As shown in Figure 3, the K-prototype uses the Euclidean distance of K-means and the dissimilarity measurement of K-modes to cluster the data points. The process flow for the K-prototype is defined in Algorithm 2.

As shown in Equation (2) [56], the K-prototype algorithm, in addition to the Euclidean distance measure in Equation (1), uses the dissimilarity difference measure for categorical data.

\[
D(x_{ij}, y_{ij}) = \left(\frac{m_{x_{ij}} + m_{y_{ij}}}{m_{x_{ij}} \times m_{y_{ij}}}\right) \times \delta(x_{ij}, y_{ij})
\]

and

\[
\delta(x_{ij}, y_{ij}) = \begin{cases} 
0 & \text{if } x_{ij} = y_{ij} \\
1 & \text{if } x_{ij} \neq y_{ij}
\end{cases}
\]

where \(x_{ij} = \text{value of mode } x_i \text{ on attribute } a_j\), \(y_{ij} = \text{value of object } y_i \text{ on attribute } a_j\), \(m_{x_{ij}} = \text{number of times } x_{ij} \text{ appears in the set of modes on attributes } a_j\), \(m_{y_{ij}} = \text{number of times } y_{ij} \text{ appears in the set of modes on attributes } a_j\), and \(\delta(x_{ij}, y_{ij})\) is the dissimilarity difference measure.

4. Cluster Simulation and Results

The results simulation utilised the Anaconda Python distribution software from Continuum Analytics (Conda) with NumPy, SciPy, pandas, matplotlib, and sklearn. For the presentation layer of the project, the Jupyter Notebook, which combines markdown text with Python source code to create a canvas, was utilised.

4.1. Data Type. The data type of the attributes in the study consists mainly of the object string and integer types. Two attributes have integer data types, while the remaining 26 have string data types. The data types are derived from the questionnaire administered to the respondents and are relevant for modelling the behaviour patterns after clustering.

4.2. Detecting the Number of Clusters

Research Question 1: What is the appropriate number of clusters from the numeric and categorical data set for student categorisation?

To respond to the research question and automate the number of clusters, the elbow method for optimal cluster selection is implemented. The K-prototype uses the cost function that combines numeric and categorical variables to determine the number of clusters. The point of inflexion on the curve, referenced as the elbow, is a good indicator of the cluster divisions. According to the scree plot in Figure 4, the optimal number of clusters at the point the elbow occurs with minimal change in the cost function is at \(K = 2\). Since
the cluster formation starts from zero, the elbow function produces three clusters.

Research Question 2: To what extent do cluster groupings vary regarding members and the likelihood of a student belonging to a cluster?

In response to research Question 2, the K-prototype predict cluster function, as shown in Figure 5, was used to assign each student to one of three clusters. Cluster 0 corresponds to the first cluster, Cluster 1 to the second, and Cluster 2 to the third. The groups are formed based on data resemblance patterns between the 905 numerical and categorical data instances.

4.3. Detecting the Number of Students in Each Cluster. Each cluster contains a varying number of members. As shown in Figure 6, Cluster 1 has the highest number of members, 366. Cluster 0 is the second largest, with 280 members, while Cluster 2 has the list, with 259 members. The number of clusters members for each group is unique based on similarity features. In percentage terms, Cluster 1 has 40.44%, greater than 30.94% of Cluster 0 and 28.69% of Cluster 2.

4.4. Cluster Member Distribution, Patterns, and Visualisation

Research Question 3: What are the similarities and distinct features of each cluster, and what factors indeed influenced the cluster groupings?

In response to research Question 3, the variation of each cluster represents the distribution of each member concerning the cluster’s characteristics. As shown in Figure 7, Clusters 0
and 2 members converge with greater feature similarities than Cluster 1. The similar behaviour of members of Cluster 1 has levels of deviation that decreases any generalisation recommendation for the cluster members. Cluster 1 members diverge significantly in features. In contrast, the generality of Clusters 0 and 2 exemplify good cluster behaviour, with the majority of its members displaying comparable characteristics.

The numerical mean of the clusters, as shown in Table 3, indicates that the majority of Cluster 0 members are level 100 students with an average of four siblings. Cluster 1 consists primarily of level 300 and 400 students with an average of five siblings, while Cluster 2 consists of level 200 students with an average of four and five siblings.

The respective mean and standard deviation distribution across the three clusters are shown in Table 4. From the results, Cluster 1 has the worst standard deviation, as seen
in the cluster distribution in Figure 7. This also reveals that, relative to the mean, Cluster 1 data instances are more prevalent and less reliable than those from the other two clusters. Cluster 1 generalisation is problematic for counselling its members as a unit. Clusters 0 and 2 have a lower standard deviation relative to the mean, indicating that their members are more similar. The numeric data type and results have a significant effect in analysing object-based categorical clustering outcomes.

The categorical cluster representation is depicted in Table 5, with each cluster feature representative of the cluster members. Cluster 0 members are mostly level 100 students between the ages of 21–23. The cluster 0 members primarily reside in a hostel with three or more roommates and are active in class. These categorical features distinguish Cluster 0 members from Clusters 1 and 2. Age, number of roommates, and class participation are essential components of Cluster 0. The findings demonstrate that the number of roommates in Cluster 0 significantly affects members during class participation. The members of Cluster 0 are generally active. Clusters 1 and 2 for categorical data have similar features but differ in numeric data representation. The members of Clusters 1 and 2 are generally 24 years old and above and stay in two in-a-room hostel facilities but are quiet during class engagements. Cluster 2 members are predominantly level 200 students, whereas Cluster 1 members are, on average, level 300 and level 400 students.

4.5. Predictive Modelling of Cluster Groups

Research Question 4: Which classification algorithm has the highest performance metric for the futuristic prediction of learner clusters?

In response to research Question 4, we examined the prediction abilities of the KNN, NB, LR, and AdaBoost ensemble algorithms in a tenfold and fivefold cross-validation study.

Three classification metrics, including accuracy, F-measure, and receiver operating characteristic–area under the curve (ROC–AUC), were utilised to measure the performance of the algorithms. As depicted in Tables 6 and 7, the AdaBoost ensemble algorithm using NB has the highest accuracy of 99.88 from the tenfold cross-validation technique. The AdaBoost multi-classifier outperformed the other conventional supervised algorithms for accuracy.

5. Discussion and Findings

This study implemented the K-prototype algorithm to group students with similar behavioural features into distinct groups. Since the data types included numeric and categorical information, applying K-means and K-modes algorithms would have generated dissimilar clusters. The research utilised the elbow function to form three unique clusters with behavioural trends and patterns. Experimental results show that level 100 students who stay in a hostel with three or more roommates are generally active during lessons. The inferred cluster indicates that the number of roommates influences level 100 students’ ability to engage a lecturer in a class. In contrast, students in levels 200, 300, and 400 are typically passive during classroom engagements and usually reside in a hostel with two roommates. Since level 100 students are new in the academic environment, parents prefer a hostel facility with a higher number of roommates because of the added security. As a result of their enthusiasm and engagement at the hostel, level 100 students are most likely to participate during lesson delivery. The accommodation status of mostly two in a room at upper levels suggests a more private lifestyle, reflected in their passive behaviour during lesson delivery. These hidden trends among various cluster groupings have implications for academic performance and associated
### Table 5: The three clusters.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
</tr>
<tr>
<td>Age range</td>
<td>21–23</td>
<td>24-above</td>
<td>24-above</td>
</tr>
<tr>
<td>Marital status</td>
<td>Single</td>
<td>Single</td>
<td>Single</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Akan</td>
<td>Akan</td>
<td>Akan</td>
</tr>
<tr>
<td>Religion</td>
<td>Christianity</td>
<td>Christianity</td>
<td>Christianity</td>
</tr>
<tr>
<td>My guardians/parents are</td>
<td>Friendly</td>
<td>Friendly</td>
<td>Friendly</td>
</tr>
<tr>
<td>Do you have siblings</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>What is the gender of your siblings</td>
<td>Both male and Female</td>
<td>Both male and female</td>
<td>Both male and female</td>
</tr>
<tr>
<td>What is the social class of your family (financial)</td>
<td>Middle</td>
<td>Middle</td>
<td>Middle</td>
</tr>
<tr>
<td>Educational background of your parents</td>
<td>Averagely educated</td>
<td>Averagely educated</td>
<td>Averagely educated</td>
</tr>
<tr>
<td>SHS school category</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
</tr>
<tr>
<td>Residential status in SHS</td>
<td>Boarding</td>
<td>Boarding</td>
<td>Boarding</td>
</tr>
<tr>
<td>My SHS is located in the</td>
<td>Semi-urban region</td>
<td>Semi-urban region</td>
<td>Semi-urban region</td>
</tr>
<tr>
<td>Did you take part in extra-curricular activities in SHS?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Subject studied in SHS</td>
<td>General Aarts</td>
<td>General arts</td>
<td>General arts</td>
</tr>
<tr>
<td>After lectures, I normally</td>
<td>Stay behind and learn</td>
<td>Stay behind and learn</td>
<td>Stay behind and learn</td>
</tr>
<tr>
<td>After the day is over, I prefer to</td>
<td>Walk alone to where I stay</td>
<td>Walk alone to where I stay</td>
<td>Walk alone to where I stay</td>
</tr>
<tr>
<td>During lectures, I prefer to</td>
<td>Be active, answer questions</td>
<td>Be quiet</td>
<td>Be quiet</td>
</tr>
<tr>
<td>I prefer to</td>
<td>Study alone</td>
<td>Study alone</td>
<td>Study alone</td>
</tr>
<tr>
<td>In a group discussion, I prefer to be</td>
<td>Active</td>
<td>Active</td>
<td>Active</td>
</tr>
<tr>
<td>Residential status in the university</td>
<td>Hostel</td>
<td>Hostel</td>
<td>Hostel</td>
</tr>
<tr>
<td>Accommodation status at where you live</td>
<td>Three or more in a room</td>
<td>Two in a room</td>
<td>Two in a room</td>
</tr>
<tr>
<td>Residential condition at where you stay</td>
<td>Quiet</td>
<td>Quiet</td>
<td>Quiet</td>
</tr>
<tr>
<td>Do you take part in extra-curricular activities in the university?</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Are you working whiles schooling?</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Are you dating on campus (in any relationship?)</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Table 6: Classification performance using tenfold cross-validation.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F-measure</th>
<th>ROC–AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>99.77</td>
<td>0.998</td>
<td>1.00</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>83.20</td>
<td>0.830</td>
<td>0.921</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>99.33</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>AdaBoost (NB)</td>
<td><strong>99.88</strong></td>
<td>0.999</td>
<td>1.00</td>
</tr>
<tr>
<td>AdaBoost (KNN)</td>
<td>83.20</td>
<td>0.830</td>
<td>0.921</td>
</tr>
<tr>
<td>AdaBoost (LR)</td>
<td>99.33</td>
<td>0.993</td>
<td>0.999</td>
</tr>
</tbody>
</table>

KNN, K-nearest neighbours; LR, logistic regression. Bold values signify the highest accuracy.

### Table 7: Classification performance using fivefold cross-validation.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F-measure</th>
<th>ROC–AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>99.66</td>
<td>0.997</td>
<td>1.00</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>81.76</td>
<td>0.816</td>
<td>0.909</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>99.55</td>
<td>0.996</td>
<td>0.999</td>
</tr>
<tr>
<td>AdaBoost (NB)</td>
<td>99.66</td>
<td>0.997</td>
<td>1.00</td>
</tr>
<tr>
<td>AdaBoost (KNN)</td>
<td>81.76</td>
<td>0.816</td>
<td>0.909</td>
</tr>
<tr>
<td>AdaBoost (LR)</td>
<td>99.55</td>
<td>0.996</td>
<td>0.999</td>
</tr>
</tbody>
</table>

KNN, K-nearest neighbours; LR, logistic regression.
behaviours. In order to improve learners’ negative behavioural choices and habits, which have the potential to affect their academic performance, it is necessary to provide tailored counselling to distinct clusters.

Comparative cluster analysis between this study and similar literature reveals distinct results because of differences in research objective, target audience, and data set. Unlike most studies that delve into online students’ behaviour, we looked at the conventional campus behaviour of students and its relevance to class engagement. However, the use of the K-prototype algorithm for mixed data types is represented in the findings in literature. Moubayed et al. [18] used the K-means algorithm on numeric data, and the results of their study indicate that the number of logins and the average duration to submit assignments online significantly impact students’ engagement more than the number of contents read and accessed. Ge et al. [19] implemented the K-prototype algorithm on mixed data type to observe students’ behaviour online. The results show that students with superior learning media exhibit better online learning behaviour. Liu and d’Aquin [20] used the K-prototype algorithm to cluster students based on demographic features and online engagement. The result indicates that mature active students with higher educational qualifications are more successful in interacting with the virtual environment. Palani et al. [21] compared Gaussian mixture, hierarchical, and K-prototype clustering algorithms to identify low-engagement students in a virtual learning environment. The K-prototype algorithm clustered the low-engagement students better. Asif et al. [47] applied the X-means to determine high-performing and low-performing students based on the courses that are high and low indicators in a traditional classroom. They realised that learners stay in the same cluster throughout the 4 years of their studies. Asif et al. [47] further implemented a classification model for future cluster prediction of learners.

Predictive modelling of students for early counselling using classification mechanisms is ideal after clustering. As a result of the formation of distinct class labels after clustering, the study compared conventional algorithms with the AdaBoost ensemble to determine the best-performing classifier for the data set. The classification result shows high accuracy of 99.88 for AdaBoost (NB) using the tenfold cross-validation technique. The strong performance of the ensemble algorithm is supported by the findings of Srividya et al. [45] when they compared the ensemble bagging to SVM, LR, KNN, and (NB). Second, the study by Asif et al. [47], with NB having the highest accuracy, supports our findings when NB has the highest accuracy of 99.77 among the conventional algorithms.

6. Conclusion and Future Work

This study investigated four research questions with the final aim of advising instructors and management with intelligent cluster information on students for proactive counselling and policy. The first question involves cluster groupings for distinct student categorisation. The results show that automating the number of clusters is based on the data type, responses, and the number of data instances. The second question strives to derive differences between the cluster groupings. The variation of cluster groupings is based on cluster membership. Each data instance using the K-prototype algorithm must belong to one cluster group with similar features. The third question involves cluster membership distribution, whether similar or distinct. The result detailed features that distinguish clusters. Mean and standard deviation are the primary statistical measure for numeric data, while categorical data are in the cluster membership. The final question relates to a classification model for future prediction. With the predictive model, new students will automatically join similar clusters using the best-supervised learning algorithm.

6.1. Implications to Theory and Practice. In the peculiar case of the University of Education, Winneba, class participation concerns instructors and management with complaints of passive learners at higher levels. According to the findings from this study, instructors should pay attention to students that reside in one or two-room occupancy and engage them further during class hours. Instructors should also monitor the academic performance of these students to provide proactive guidance, particularly about the roommates and motivation. The predictive model will also help instructors to determine the cluster groupings early for personalised counselling of students.

6.2. Future Work. In a future study, we will compare deep learning algorithms to the ensemble multi-classifiers and use digital assessment variables to model student behaviour. Deep learning, which uses neural networks with multiple levels of representation, has gained usage with higher classification performance metrics from diverse application domains.

Data Availability

Data used in this study are available upon request from the corresponding author.

Additional Points

Key Lessons Learnt. Student clustering is vital for personalised and collaborative learning, especially in Education 4.0. Instead of the generic method of learner groupings without any analytics, the drive for academic institutions to meet the demands of Industry 4.0 requires the integration of disruptive technologies in the classroom. The implementation of a clustering algorithm should be based on the data type and the number of data instances. Even though clustering alone reveals interesting learner groupings, classification is essential for the future prediction of newer students.

Limitations of the Study. Even though the research clustered learner groupings and revealed interesting patterns, we faced data size limitations due to the availability of respondents. Second, the lack of previous literature, especially on conventional campus behaviour modelling, hindered discussions on similar data attributes for cluster performance analysis and predictions. Finally, it was impossible to compare the findings from the research to machine learning-based
behaviour modelling in Ghana. Similar literature was unavailable in Ghana to the best of our knowledge.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**References**


