Predicting Learners' Academic Progression Using Subspace Clique Model in Multidimensional Data

Mr. Oyugi Odhiambo James, Prof. Waweru Mwangi, Dr. Kennedy Ogada
Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

***Abstract*—Subspace clustering examines the traditional clustering techniques that have previously been considered the best approaches to clustering data. This study uses a subspace clustering approach to predict learners' academic progress over time. Using the subspace clustering method, a model was developed that improves the classic Clique by optimizing clustering performance and addresses the clustering challenges posed by inaccuracies due to additional data size and increased dimensionality. The study used an experimental design that included data validation and training to predict students' academic progress. Clustering evaluation metrics including accuracy, precision, and recall measures were identified. The optimized model recorded a better performance index with 98.90% accuracy, 98.50% precision, and 98.50% recall which directly shows the efficiency of the optimized model in predicting learning academic progress through clustering. In this regard, conclusions are drawn for an alternative approach to predictive modeling through cluster analysis, so that educational institutions have a better opportunity to manage learners by ensuring adequate preparation in terms of resources, policies and knowledge. It highlights career guidance for learners based on their academic progress. The result validates the suitability of the model for clustering multidimensional data.**

Keywords*—Subspace clustering; clique model; academic progression; multidimensional data; feature engineering; cross validation and principal component analysis

I. Introduction

A. Introduction

This component describes the background of the study, the objectives and the problems of the study. The literature of the study within the local, regional and international perspectives on the use and application of clique models, the importance and the gaps addressed in this study.

B. Background of Study

Academic progress is important in a learning environment where each learner must be assessed to determine their progress to the next level of learning. Various considerations are taken into account and typically indicators of progress in a learning environment are well established and clearly stated, although this may vary from institution to institution. There are standard learning levels such as certificate, diploma, bachelor's, master's and doctoral. It is also notable that educational institutions need to understand learners' progress to ensure efficient and effective management of learners.

The main objective of this study was to develop a model that predict leaners' academic progression. The improved Subspace Clique model was expected to address the challenges of classical Clique in clustering of data. For instance, the process of finding clusters in multidimensional data space is a complex procedure due to the large number of attributes and tuples involved. In the case of multidimensional data, the density points are at their lowest level. The approach taken by traditional Clique cannot stand due to the inaccuracies and inconsistencies, thus misleading in finding the objective clusters. Because of various scores of attributes involved, clusters are not always found in their actual multidimensional data space [1]. This means that it could be possible to find these clusters in specific subspace of the entire dataset space. The problem of dimensionality is common in data mining, for instance the dimension increases with the increase in the number of attributes in a particular dataset leading to the curse of dimensionality which an optimized Clique can address. In [2], improved Clique algorithm from a hybrid of Clique and K-means, the experiment was conducted on artificially simulated dataset, which revealed that the hybrid was not sensitive to the input parameters used in classical Clique but was silent on the dimensionality challenge and different dataset effect. Ultimately a significant number of attributes are normally dropped before the actual experiments are conducted which may affect the results if not well accounted for [3]. On the other hand, traditional clustering algorithms break when employed in multidimensional data spaces, and that they present many irrelevant attributes that could limit the possibility of clustering. The current subspace clustering methods are mainly used in either numerical or categorical data but not all [4]. The method recommended in this study takes into account the different data types that are subject to proper data preprocessing. In this study, an improved Clique algorithm was proposed, which takes into account students' academic performance in predicting learning progress. This is an area that many researchers have overlooked when conducting behavioral analysis using Clique algorithm.

C. Literature Review

This study is based on the subspace clique clustering technique. The focus of this study was on predicting learners' academic progress using multidimensional data. To achieve this, future forecasts were made based on the available data. Whereas a Cluster is an ordered list of data which have the familiar characteristics [5], Clustering refers to an ill-posed problem which aims to reveal interesting structures in the data or to derive a useful grouping of the observations. However, specifying what is interesting or useful in a formal way is challenging. This complicates the specification of suitable criteria for selecting a clustering method, or a final clustering solution [6] also emphasized this point. He argued that the definition of the true clusters depends on the context and on the

aim of clustering. Therefore, given the data, there is no clear clustering solution, but different aims of clustering imply different solutions, and analysts should in general be aware of the ambiguity inherent in cluster analysis and thus be transparent about their clustering aims when presenting the solutions obtained [7].

CLIQUE is a subspace clustering algorithm that operates on a grid and density basis. By combining the advantages of density-based clustering with the advantages of grid-based clustering, it can find clusters of any shape while still managing large amounts of data [8]. The clustering process starts with a single dimension scaling upwards to higher dimensions. Clique divides the N-dimensional data space into non-overlapping rectangular units from which dense units are identified. A unit is considered highly dense if the sum of the total data points in the unit runs over an input parameter, Clusters are then created from the original data space using the Apriori principle [9]. CLIQUE finds high-quality clusters only in subspaces with the highest dimensionality, making it an efficient method. The threshold density and grid size must be properly adjusted in order to produce meaningful clustering results.

Research on Clique-based Model in Predictive Analysis has been conducted by several scholars, such as [10], One such study focused on forecasting future weights based on a partial order set using the Clique Algorithm for pattern analysis to cluster high-dimensional data [11]. In text mining [12], Customer Segmentation and trend analysis [13], Further, intrusion detection in IoT networks [14]. The study in [15] proposed dimensionality reduction via feature reduction. as well as imaging processing [16]. Student behavioral data [17] among other fields. This study proposed Clique model in predicting leaners academic progression in multidimensional dataset.

Numerous studies have been carried out to forecast the development, effectiveness, and behavior of leaners. For instance, [12] used the vector space model and the clique subspace model to study critical text mining of learners' behavioral patterns in Kenya. The findings demonstrated that even for large data sets, the processing time in Clique is significantly less; however, the researcher noted that grid cell consideration, density threshold, and parameter input requirements are limitations.

Similarly, [18] studied in the USA to find course clusters and course cliques based on the degree of grade correlation between student grades in pairs of courses in one of the universities using subspace clique and network diagrams, where courses are represented as nodes and are connected to courses if they have a high degree of grade correlation. The ability of Clique to cluster academic data is demonstrated by an analysis of the results for course pairs and courses grouped by academic department.

For example, [17] research on student behavior patterns was done in China. In an effort to address the issue, the study suggested an unsupervised clustering framework that combined behavioral data from students with grade point averages to find behavioral patterns, a similar study reflected in study [19]. The suggested framework incorporated the views of statistics and entropy to extract behavior features, which are combined with k-means and density-based spatial clustering of applications with noise (DBSCAN) algorithms to find behavioral patterns. An analysis was conducted to compare the enhanced model with the conventional Clique subspace clustering model. In comparison to the improved model, the results showed that Clique performed a little bit worse. The findings demonstrate that the framework is able to identify both mainstream and anomalous behavioral patterns. Although this was the case, future selection only considered variance and correlation rather than cross validation and principal component analysis, meaning that the data management of clustering was not given much thought.

A similar study by [20] was conducted in Korea on various clustering algorithms with a focus on pattern identification. Clique became one of the algorithms used and tried to study the convergence speed and accuracy of clustering. The study used small data sets that achieved high accuracy with sensitivity to density and hyperparameter tuning. It was not clear how the optimization that resulted in high accuracy was achieved, as the study did not specify the limitations of small datasets and dimensionality management, which has been addressed in this study.

In the United Kingdom, [21] conducted research to design a system for assessing and guiding the mental health of college students. The study examined the applications of clustering data mining algorithms relevant to the researcher's area of interest. The research involved college students and used performance testing to determine the system's accuracy and effectiveness in assessing and managing students' mental health. However, the study acknowledged the pattern recognition ability of the Clique algorithm, although it focused on developing an artificial intelligence-based system for analyzing the mental health of college students. Further suggestions from [22] affirm that the CLIQUE algorithm can perform effective cluster analysis and automatically adapt to different subspaces.

A study by [23] to find nonlinear feature relationship pattern recognition in India. Even in cases where feature relationships are nonlinear, the method enables the discovery of bi-clusters based on feature relationships. The suggested approach used datasets from various domains and did not require the user to provide any parameters. Clustering with the Clique algorithm was used to assess performance. The fact that the new approach outperformed the original Clique, suggests that it needs to be improved. The research in [24] are among the other benchmarked studies with similar findings.

In summary, the studies conducted in Kenya, the United States, the United Kingdom, China, India and Korea using the Clique algorithm clustering technique have not only shown the scope bias but also found serious gaps, which accompany the implementation and use of the Clique algorithm, which require attention. For example, most studies have been conducted to analyze student behavior patterns, but not to direct academic progress and/or performance, which is critical to excelling in an academic learning environment and in managing learners. Other notable challenges would be data preparation before use, data size, methods used for dimensionality reduction, future selection, and consideration of hyperparameters, which in turn affect the accuracy of the clustering method. Other studies that

conducted a comparative analysis between ordinary Clique and other existing models, trying to recommend a model that provides a better performance index for some specific data, which is not necessarily academic in nature did not pay attention to improving the models.

In countering the gaps with studies conducted in Kenya and reflected elsewhere, this study recommends predicting students' academic progress using an improved subspace clique in multidimensional data, which considers model optimization through feature engineering, data standardization, dimensionality reduction and cross-validation.

In other sections, the paper further extends to provide information on study-related work, data acquisition and preprocessing, the design, methods, tools used, and the implementation of the predictive subspace Clique model. The paper concludes by presenting and discussing the results, conducting a comparative analysis, and exploring the significance of the research. Additionally, it draws conclusions and offers recommendations for future work.

II. Related Work

A. Introduction

This section includes a detailed review of the relevant literature and recent studies on implementing the subspace clique model for predicting learners' academic progress in multidimensional data, as well as a summary of results from such applications.

B. Subspace Clique Model

Research on the topic of using Subspace Clique to predict student academic progress and related learning areas, or such behavioral environments is extensive. For example, in a study by [25], blended aerobics learning is analyzed and guided using data mining. Action learning is combined with blended aerobics instruction to promote learning progress. The study proposed a Clique clustering algorithm that meets the following criteria: (1) identify embedded clusters in high-dimensional data subspaces; (2) scale; (3) understand end-user results; and (4) predict cluster descriptions into minimized density expressions to promote understanding [25]. In the modern educational environment, blended learning has enormous growth potential. The study examined the practical value of meta-learning theory in its application to aerobics instruction. The results showed that students have potential meta-learning, can deliberately improve students' meta-learning ability, and it is important to increase students' interest in aerobic skills through appropriate learning strategies [26]. Subjects were tested on basic questions, aerobic technical skills and physical fitness in the experimental data set. A comparative analysis of the semester's academic performance was reviewed following a teaching exam that lasted 15 weeks, equivalent to a full semester at the university. After applying the classic clique model to the dataset and analyzing the results using accuracy, precision, and recall metrics, it was found that the model had an accuracy of 86.5%, a precision of 85.99%, and a recall of 85.9%. Although the accuracy was higher on average, the study's margin of error of 14.5 percent on accuracy was noted and could be associated with certain observations such as the unclear explanation of data management, feature selection, and dimensionality reduction with respect to cross-validation.

In a separate study conducted by [24], investigating clustering algorithms based on grids, the focus was on the appropriate selection of grid cells that contributed to the field, and a novel grid-based algorithm that employs an automated method for calculating the number of grid cells was proposed. The study covered the idea of grid cells in the Clique algorithm and used the Clique model to contrast the outcomes of the upgraded model. A simulated educational dataset was run through a Python script in order to identify learning patterns, which produced results with accuracy of 95.23%, precision 95.0% and a recall of 95.20%. The well-known Clique algorithm was used in the experiment as a benchmark, allowing a quick pinpoint on a few observations; even though the Clique model was efficient at establishing clusters when clustering data. Notably, it required two input parameters, i.e., the threshold for density and the number of intervals. The parameters in this case were difficult to calculate. Based on this study, we can conclude that the model did not function as well as the researcher had hoped.

In a related study, [17] used the clustering approach to analyze student behavior patterns. The study's goal was to assist the institution in setting targeted rules, particularly for unexpected patterns, and in determining more effective ways to provide specialized services and management. To address the issue, the study took into account a clustering technique. The study analyzed the relationships between various behavior patterns and students' grade point averages by conducting experiments on six different types of behavioral data generated by university students (eating behavior, shopping behavior, library entry behavior, and gateway login behavior, respectively). The six attributes were gathered from 9024 university students. Grade point averages served as the foundation for the computation of extrinsic metrics, which quantify the relationship between various behaviors and academic achievement. The characteristics corresponding to each type of behavior were derived from a central tendency perspective. With a comparative performance evaluation against the Clique algorithm, the study proposed a hybrid clustering algorithm that combines the best features of DBSCAN and K-means. Using the Clique algorithm, the results showed an accuracy of 92.0%, precision of 91.83% and recall of 90.12%. This explained the model's ability to group educational data. The researcher added that the improved model performed better than Clique by achieving an accuracy performance of 92.0%, which was only a small improvement. The experimental results show that the proposed method can not only detect abnormal behavior patterns but also identify different behavior patterns more accurately. Based on the clustering results, student departments can take more targeted interventions and specialized services. The study recommended that future work focus on creating meaningful features, creating new distance measures, reducing the dimensionality of feature spaces, and, lastly, investigating behavioral patterns and student labels in order to improve clustering analysis.

In a related study by [18], the researcher in this study computes the correlation of student grades between pairs of courses in a university, based on academic performance, in an

attempt to replicate the course groupings. Courses are represented as nodes in the course network graphs created for the study, and courses are connected if their grade correlation is high. Graph mining and network analysis tools were used in conjunction with the clique algorithm to visualize course networks and detect cliques and clusters within them, where the Pearson correlation was used to determine how similar two courses are to one another. Recall that the Clique concept was used in this study to visualize the results and obtain graphical Cliques. A k-clique is a collection of k nodes that are all connected to one another directly by an edge. In this study, the 0.5 correlation threshold was exceeded by 25% of course pairs and the 0.7 correlation threshold by 5% of pairs when at least 20 common students were examined, these results were evaluated against the metrics of accuracy, precision and recall. Which indicated a uniform performance of 75% for the three metrics. The study found a high correlation between student performance and a number of course pairs. Within the course correlation networks, cliques and modularity classes were recognized as course clusters. Course pairings and course groupings based on academic department were examined. According to the study, there is a significant grouping of courses with strong similarities across scientific disciplines, pre-health courses, and computer science subfields. Notably, the researcher stated that no study that used the concept of course similarity had ever been carried out using the clustering technique in a predictive way.

According to a study by [23] on a free relative density-based clustering method in nonlinear feature relationship patterns. The method proposed in this study allowed finding clusters based on feature relationships, even if the relationships are nonlinear. Since the proposed method did not require any input from the user, it could be applied to datasets from different domains. Fifteen simulated datasets were used for the experiments and eleven different clustering algorithms were used to compare their performance. Among them was the Clique clustering algorithm. For most simulated datasets used to identify behavioral patterns in learning environments, the proposed method appeared to provide better results. After several clustering operations on different simulated datasets, the Clique model showed an average performance of 93.9% accuracy, 94.3% precision and 93.5% recall; this was the best Clique performance in all of the researcher's experiments. The need for research in various dimensions is explained by the fact that the study did not achieve 100% accuracy even after using different data sets.

We can quickly identify a few issues based on the observations made in the various literature in the study for the use of Clique subspace clustering in predicting learners' progress. According to [25] study, Clique was able to perform clustering when used to predict learning progress in Aerobics. However, the algorithm was viewed with disdain due to its ambiguous explanation of data management, feature selection, and dimensionality reduction in relation to cross-validation. In a separate study by [24] Encompassing grid-based clustering algorithms and the use of well-known Clique algorithms in the experiment, it was reiterated that although the Clique model was efficient in forming clusters when clustering data, it required input parameters, a density threshold and the number of intervals required what was difficult to calculate. [17] used the Clique clustering approach to examine patterns of student behavior in a related study. Based on the findings, the study suggested that future research concentrate on developing new distance measures, dimensionality reduction, and behavioral pattern analysis to enhance clustering analysis. Similar to the course groupings created by [18], the researcher in this study used the Clique algorithm to calculate the correlation of student grades between pairs of courses in a university based on academic performance. The researcher noted that no study that used the notion of course similarity had ever employed the clustering technique in a predictive manner. Consequently [23] conducted a study on nonlinear feature relationship patterns. They proposed a free relative density-based clustering method and compared it directly to the Clique algorithm. The research used fifteen different data sets, but it did not establish a clear connection between the research and students' academic progress. However, the literature reveals significant deficiencies in each of the reviewed studies, including shortcomings in input parameters, dimensionality reduction, density threshold, data management, performance metrics and specific areas of application, resulting in inaccuracies in cluster analysis. This study adopts a unique strategy to address these issues and enhance the performance of the Clique algorithm in predicting learners' progress in a multidimensional dataset.

C. Existing Prediction Models

A popular clustering technique for predicting leaner behavior is the K-Means algorithm. This was the main choice made by researchers in a study by [29] titled “Identifying student behavior patterns in higher education using K-Means clustering.” This method is mainly used because it is very simple and gives better, easy-to-understand results. The data set used in the study was obtained from a university database. The dataset includes data derived from student files according to their academic, behavioral and demographic characteristics. Various tuning parameters were used to optimize the model. After several analyses, the results revealed an accuracy rate of 93%, a recall rate of 94%, and a precision rate of 93%. According to the study, universities must recognize student behavior patterns and identify students at risk early on. For example, students may have a positive attitude at the beginning of the semester but perform poorly towards the end, or the opposite. However, by optimizing data preprocessing activities and experimenting in various data sets with different characteristics, the researcher found that K-means clustering should be enhanced to improve the results.

In a study on the BIRCH algorithm (Balanced Iterative Reducing and Clustering Using Hierarchies) to analyze the integration of core competencies in sports and health courses at universities based on data mining techniques by [28]. This study conducted research to examine the state of courses in universities using data mining techniques. The study focused on identifying patterns of information on the essential competencies of university health and physical education courses. Various experiments were conducted to determine the ability of the BIRCH algorithm in predictive analysis. The aftermath experiment outlined the result as follows using the metrics of accuracy (91.1%), recall (91.1%) and precision (91%). The algorithm made credible predictions, but with

limitations in the accuracy and reliability of the results, for instance BIRCH can only process metric attributes. A metric attribute is any attribute whose values can be represented in Euclidean space, i.e. there should be no categorical attributes, which the researcher believed could be addressed by improving the algorithm.

In research conducted by [31], Density-based spatial clustering of applications with noise (DBSCAN) was used in academic performance analysis at higher education institutions with educational data mining in a normally detection manner. Academic data in the form of study findings over a specific time period makes up the data set that was used. It was discovered that the DBSCAN algorithm can identify academic data anomalies with accuracy of up to 91.0%, precision of 90%, and recall of 90% after a series of experiments were carried out to identify data anomalies from academic data. In other words, the variation leaves a gap that can be filled by additional study. Consequently, in order to enhance the effectiveness and dependability of the clustering algorithm, DBSCAN is highly sensitive to the epsilon parameter setting; a low value will result in the clusters being classified as noise. The clusters will merge and become denser simultaneously with the shift to a higher value [30]. One potential issue with DBSCAN could be the global constant parameter needed to determine the neighborhood's radius, since it could have an impact on the accuracy of the findings.

D. Research Gaps

In summary, we can quickly identify the gaps in the literature that led to this study. A number of studies conducted to predict learner behavior or analyze using the Clique model or in comparison with other models have drawbacks, including the initialization of probabilities or the creation of ratios for starting experiments, which become apparent and later affects the accuracy of the results. The choice of data set, whether linear or nonlinear, categorical or numerical, is challenging for some algorithms in such a context. The problem with parameter tuning, which sometimes leads to biased results, has been highlighted in a number of studies. Challenges in reducing dimensionality due to multidimensional data that some of the methods examined in this study could not address. The density threshold limitation for density-based clustering methods, the scope and specific focus of application sometimes does not ensure reliable results, lack of appropriate evaluation metrics caused by inconsistency in metric evaluation for the same algorithm in a similar study area is itself a major gap and finally ambiguity in clustering objectives that affect cluster analysis.

However, in order to predict learners' academic progress using multidimensional data, this study adopted a different strategy and thought about how best to solve these problems. For example, feature engineering is used to process data and is considered effective in feature selection and data labeling. Principal component analysis was used to address data dimensionality, reducing the dimensions to a more manageable set of attributes. The rationale for appropriate parameter tuning and metrics for evaluating results have been addressed in detail in this study to ensure consistency. To confirm the validity of the results, data validation was performed and our work was compared with that of other researchers to ensure new contribution. This study used academic performance data from a university, which represents the main factor in learners' progress and makes the results legitimate and relevant, as opposed to other studies that widely examine general behavioral analysis and not progression. For example, to exploit how unique the application area is from other studies, when a learner completes a course of study and, on the other hand, that learner is good at co-curriculum activities such as sports, the institution, on the assumption that sports is not a subject of study, takes into account academic performance when it comes to the transition to the next academic level, even though sport can have an influence on students' academic performance.

III. Methodology

A. Introduction

The data collection, preprocessing, methodology design, methods and tools for the research are outlined in this section.

B. Data Acquisition and Preprocessing

The study examined the subspace clique algorithm in predicting students' academic progress through interesting pattern analysis. This study retrieved data from a university database consisting of students' average grades over the past five years. In total, the data from 3153 students were examined. The dataset contained the following fields: student number marked D for privacy policy, diploma year 1, diploma year 2, average and grade as indicated in Table I below. A diploma course will probably only take two years. After this, the student is either expected to advance to BSc or not.

TABLE I. Sample Working Data

S/No	diplomaY1	diplomaY2	Average	Grade
D1	69.88	53.42	61.65	Credit
D2	53.33	63.40	58.36	Pass
D3	61.36	67.13	64.24	Credit
D4	60.89	59.90	60.39	Credit
D5	52.09	61.88	56.98	Pass
D6	55.93	53.61	54.77	Pass
D7	72.94	69.72	71.33	Distinction
D8	54.87	52.37	53.62	Pass
D9	61.97	68.13	65.05	Credit
D10	62.82	61.97	62.39	Credit

In the dataset, the grading is divided into the categories “Pass”, “Credit”, “Distention” and “Fail”. Students who achieve “Credit” or above are allowed to progress to BSc, while those who achieve “Pass” and “Fail” are not allowed to progress to BSc. Cleaning the data was done by trying to fill in the missing values, identifying and removing the outliers, which in the end solved the problems of inconsistencies that existed in the data before, this was achieved by Python's Jupyter Notebook. This study employed future engineering in grade assignment, ration creation, and binary label creation to predict student behavior patterns.

C. Design Methodology, Methods and Tools

The study used an experimental research design methodology to validate the use of the Clique subspace clustering model to predict learner's academic progress. Subspace clustering uses a new approach to the traditional clustering technique that aims to find clusters in different subspaces within a dataset (a subspace is a space that is

completely contained in another space or whose points or elements are all in one location in another room). It can also be defined as a vector space that is completely contained in another space. In this context, cluster analysis is about discovering groups or clusters of similar objects. The objects are usually represented as a measurement vector or point in a multidimensional space. The similarities between objects are usually determined by an observable distance measure from different dimensions in a data set. Subspace clustering uses all dimensions selected before clustering performance [27]. The study relied on the following expressed Subspace clique mathematical assumptions;

D. General Subspace Notations

1) Considering data matrix X: $X = \left(x_1, x_2, ..., x_n\right) \in R^{dxn}$ where $x_i \in R^d$ represents a data point in d-dimensional space, and n is the number of data points.

2) A subspace $S_i \subseteq R^d$ is spanned by a subset of the dimensions. For example, if S_i is spanned by dimensions $\left[j_1, j_2, ...j_k\right]$, then: $S_i = Span\left[\ell_{j1}, \ell_{j2}, ..., \ell_{jk}\right]$ where ℓ_j is the unit vector along dimension *j*.

3) The projection of a data ¬oint *x* onto a subspace S_i is given by: $Proj_{S_i}\left(x\right) = p_i\left(x\right)$ where p_i is the projection matrix corresponding to S_i.

4) The goal of subspace clustering is to find a partition of the data points and corresponding subspaces that minimize some form of error. A common objective function is:

$$C_1, C_2, ..., C_k^{min}, s_1, s_2, ...s_k \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \Pr oj_{S_i}\left(x\right) \right\|^2$$

where:

- C_i represents the set of data points assigned to the *i*-th cluster.
- S_i represents the subspace corresponding to the *i*-th cluster.

E. CLIQUE (Clustering In QUEst) Algorithm

CLIQUE is a density and grid-based approach of subspace clustering model. It reflects a grid-based approach thus represents the data space through grid and examines the density by counting the number of points in a grid cell. By density-based approach a cluster refers to a maximal set of co-joined dense units in a subspace, a unit therefore is dense if the fraction of total data points contained in the unit exceeds the input model parameter.

Subspace cluster describes a set of neighboring dense cells in an arbitrary subspace, it does unveil some minimal descriptions of the clusters. It systematically recognizes subspaces of high dimensional data space that allow better clustering than original space by concept of a priori algorithm. Mathematically the steps in the Clique algorithm are as follows:

Step 1: Data preparation

1) Let X be a dataset with m rows (also called data size/observation) and n columns (also called number of features)

Step 2: Finding 1-dimensional Dense Units

We define D_1 as the set of 1-dimensional dense units thus a 1-dimensional dense unit is a subset of data points within a feature that satisfies a certain density threshold denoted by τ

2) Let F_i denote the $i-th$ column/feature in the dataset X, for $i = 1, 2, ..., n.$

3) Let u_{ij} denote the $j-th$ in the feature F_i

4) Then, the set D_1 is defined as: $D_1 = \left\{u_{ij} \in F_i : \left\|u_{ij}\right\| > m\tau, \forall i, j\right\}$

Step 3: Candidate Generation for Higer Dimensions

5) Let k be the number of higher dimensions (i.e., $k = 2, 3, ..., n$), C_K be the set of k-dimensional dense units. For $k > 1$, we generate candidate dense units C_k by performing a self-join operation on D_{k-1} where the conditions ensure that the units share ($k-2$) dimensions.

Step 4: Finding Higher dimensional Dense Units

From the candidate dense units C_K,we filter out the units D_k such that ($k-1$) projections of a unit are in D_{k-1}.

Step 5: finding clusters

6) For each feature set $f = \left\{F_1, F_2, ..., F_k\right\}$ containing dense units C_K,we build a graph G where dense units are nodes and connection between dense units (having a common face) are edges. We then find connected components in G to identify clusters.

7) Let ρ_c be the density of combination $c \in C_k$, calculated as:

$$\rho_c = \frac{\text{Number of data points containing all features in c}}{\text{Total number of data points in X}}$$

.

8) The i-th dense unit in the feature set *f* is obtained by the formula:

$$U_i = \arg\max_{c \in C_k}(\rho_c)$$

This formula finds the combination *c* from C_k that maximizes the density.

9) To check whether dense units V and W are connected based on sharing at least one feature can be expressed as follows:

$$\begin{cases} 1, \text{ if } V \cap W \neq \phi \\ 0, \text{ otherwise} \end{cases}$$

10) Let V be the set of vertices in the graph G. The number of connected components in G is obtained by the formula:

$$p = \left| \{ \mathrm{DFS}(v) : v \in V \} \right|$$

where DFS(v) denotes the Depth-First Search traversal algorithm starting from vertex v in V.

F. The Choice for Subspace Clique Algorithm

Because of its distinct benefits in addressing the problems identified in previous research concerning the prediction of learner behavior and academic progress, the Clique Subspace method was selected for this study. Clique is particularly good at handling multidimensional, high-dimensional, and heterogeneous data, which is frequently present in educational datasets that contain a variety of learner information types, including academic performance and extracurricular activities, in contrast to traditional clustering techniques. The following explains why Clique is a good fit in this situation:

- Subspace Clustering for Multidimensionality: Clique is an algorithm for subspace clustering that is specifically tailored to detect clusters within high-dimensional spaces by identifying dense areas in subspaces instead of the entire dimensional space. This capability is particularly important in educational data analysis, where datasets frequently encompass various dimensions such as test scores, participation levels, background information, and extracurricular activities. By concentrating on relevant subspaces, Clique facilitates a natural reduction in dimensionality, effectively tackling the challenges associated with dimensionality reduction in a way that surpasses traditional clustering techniques.
- Density-Based Clustering Advantages: Clique employs a density-based technique that allows for the identification of clusters within subspaces by establishing adjustable density thresholds. This capability facilitates the exclusion of noise and irrelevant information. Such an approach effectively mitigates challenges encountered by other density-based clustering techniques, which often face limitations related to thresholds, resulting in inconsistent clusters when used with multidimensional educational data.
- Robustness to Mixed Data Types: Educational datasets generally comprise both categorical and numerical data. Clique effectively handles mixed data types by partitioning the data into grid cells, which facilitates efficient clustering while minimizing reliance on assumptions regarding the data distribution, whether linear or nonlinear. This characteristic renders Clique especially robust and dependable for analyzing various attributes of learner data.
- Minimization of Bias from Parameter Tuning: Clique stands out from certain clustering techniques that necessitate extensive parameter adjustments, which may lead to bias and inconsistency. Its parameters, specifically density thresholds and grid size, are relatively simple and can be determined either empirically or through domain expertise. This approach mitigates the tuning bias often highlighted in earlier research, thereby promoting stability and replicability in the results.
- Efficient for Large Datasets: The study makes use of extensive academic data, and Clique's grid-based clustering approach which provides a highly computational efficiency, making it well-suited for the large datasets commonly found in educational settings. This effectively resolves the scalability challenges faced by certain other clustering techniques, especially when managing substantial, multidimensional educational datasets.

The purpose of the study was to analyze complex and multidimensional educational data in order to make predictions about learner behavior and academic progression. The traditional clustering techniques have several challenges when applied in this scenario. For example, the K-Means approach requires a predetermined number of clusters, assumes spherical cluster forms, and is sensitive to outliers. On the other hand, BIRCH is only capable of supporting numerical data, it is however, not ideal for the mixed categorical and numerical data that is typically present in educational datasets. In spite of the fact that DBSCAN can recognize clusters of any shape, it is extremely sensitive to the epsilon parameter, which has an effect on the consistency of the clusters.

Importantly, the study considered the following procedures to ensure that the full capability of the experimental design methodology was maximized and the desired outcome was achieved:

a) Data partitioning: The dataset was divided into two separate folds, fold1 and fold2, where fold1 was used for the training dataset and fold2 was used for validation. The training dataset was used to build the Clique model, while the validation dataset was used to perform hyperparameter tuning to optimize the model. The two partitions can be seen in Table II below;

TABLE II. DATA PARTITIONS

Partition	Number of records
Fold 1	1884
Fold 2	1269

b) Benchmark model: The study conducted two different attempts to train the model. The first experiment was used as the main model training, in which the output results of the second experiment were compared.

c) Performance comparison: Precision, accuracy, and recall were used as performance metrics to compare how effective the enhanced Clique model predicted learners' academic progress with the main model.

d) Sensitivity analysis: The study examined the performance of the developed model across a range of datasets and conditions by exploiting variations for the initial state probabilities.

IV. Clique Model for Predicting Learners Academic Progress

A. Introduction

This section explains the basic steps to develop the model, as well as the required parameters and model architecture.

B. Background

Learning is a broad concept that varies depending on the area of application. The common learning environment familiar to many is school, from primary school to secondary school to colleges and universities. In the case of university, students participate in their studies in different categories, from the lowest level (certificate, diploma, bachelor's degree, master) to the highest level (doctoral degree). All of these categories require learners to progress gradually to the highest level after completing the requirements of a particular category. Several factors can determine learners' progress from one category to another, including academic skills, co-curricular activities, government scholarships and many others. The common determinant of a student moving from one category to another is the student's academic performance, usually determined through examination and grading in accordance with the academic policy of the university system. In this study, student performance based on academic grade was considered the primary determinant of progress.

The study examined various grades from students, which are divided into the following categories: A pass is considered to be a performance that is greater than or equal to 50 percent but less than 60 percent; a credit is considered to be an achievement that is greater than or equal to 60 percent but less than 70 percent. A performance of less than 50 percent is considered a failure; an award is given for achievements of 70 percent or more. Each category was assigned an observation that indicates the desired learning progression behavior. Table III below illustrates the state symbols for grade observation.

TABLE III. Grade Observation States

Grade	Observation state	Progress State to BSc
Distinction	Yes	Progress
Credit	Yes	Progress
Pass	No	No Progress
Fail	No	No Progress

For any student who scores a distinction grade automatically qualifies to progress to the BSc Degree, again when a student scores a credit grade the student meets the threshold to proceed to BSc Degrees, on the other hand when a student scores a Pass that student does not qualify to proceed to BSc Degree, consequently when a student gets a Fail that student does not Progress to the Category of learning that is BSc Degree. The mapping can be represented in Fig. 1 below.

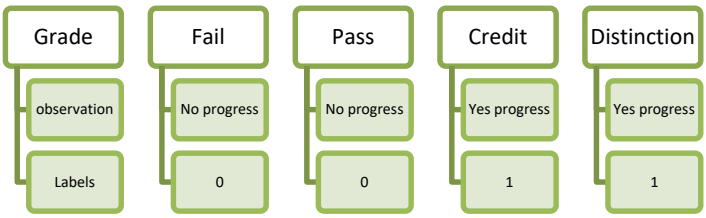

Fig. 1. Observation state mappings.

In this study, the correct assignment of different data categories was carried out, as shown in Fig. 1 above. Grades are assigned to the observed states and indicated by the labels, where “no” means (0) and “yes” means (1). For example, “Fail” means a learner has not made progress and “Credit” means progress. The states “No” and “Yes” are the hidden patterns that need to be discovered through prediction.

C. Predicting Learners Academic Progress

We used a Python script to implement the learner's academic progress in the study. This script is a comprehensive data processing and machine learning pipeline that includes data preparation, feature engineering, dimensionality reduction, clustering, evaluation, and visualization. Below is a detailed step-by-step explanation of how this script works:

- Data preparation is the first step in this process. The very first activity in this phase is to create a dictionary called Data that contains four categories of grades: Credit, Distinction, Fail, and Pass. and the corresponding number of students who answered “no” and “yes” to a particular condition. This data is then converted into a Pandas Data Frame “df” for easier editing and analysis.
- The second step included the task of Feature engineering, feature engineering involves selecting, modifying, and transforming raw data into features suitable for application in machine learning algorithms. Feature engineering was performed in three main stages;

a) Categorical to numerical conversion: The script maps the categorical grades (`Fail`, `Pass`, `Credit`, `Distinction`) to numerical values using a dictionary (`grade mapping`). This is crucial because machine learning models typically work with numerical data.

b) Creating a new feature (`Ratio`): The ratio of "Yes" responses to the total number of students (sum of "No" and "Yes") is calculated. This feature might represent the likelihood or propensity of students to answer "Yes."

c) Label creation: A new binary label, `Label`, is created where 1 indicates that more students answered "Yes" than "No," and 0 indicates the opposite. This served as the ground truth for model evaluation.

- In the third step, the script performed feature selection, where the script selects relevant features (“Grade”, “No”, “Yes”, “Ratio”) from the Data Frame to form the feature matrix “X”. The labels (0 or 1) are stored in the variable “y”.

- In step four we performed data standardization, which involved standardizing the features in "X" using the "Standard Scaler." This ensures that each feature contributes equally to the model and avoids bias due to scale differences.
- In the fifth step, the experiment involved dimensionality reduction using principal component analysis (PCA). This technique reduces the dimensionality of Scaled from 4 to 2 dimensions, which makes the data easier to visualize and also reduces noise. The transformed data is stored in "X_pca".
- In the sixth step, the experiment performed clustering using the clique subspace for cluster analysis and cluster prediction. The script applied clique subspace clustering with 2 clusters ("n_clusters=2") to the PCA-transformed data ("X_pca"). The algorithm attempts to divide the data into two groups based on the input features. To predict clusters, the Clique subspace algorithm assigns each data point to one of the two clusters, and these assignments are stored in y_pred.
- In step seven, cluster label assignment and model evaluation were implemented as follows; since clustering of clique subspaces is unsupervised, the clusters may not correspond directly to the original labels ("y"). The script checks the accuracy of the initial clustering. If it is less than 50%, the cluster labels ("y_pred = 1 – y_pred") are inverted to match the actual labels. During model evaluation, the script calculated key performance metrics including: accuracy (the percentage of correct predictions), precision (the proportion of actual correct positive identifications), and recall (the proportion of actual positive identifications that were correctly identified).
- In the last step the experiment performed results display and visualization. The script adds the predicted cluster labels (`y_pred`) to the Data Frame `df` under the column `Predicted Progression'. It then prints the Data Frame to show the original data along with the predicted progression. Through Visualization a scatter plot of the PCA-reduced data (`X_pca`) is generated, where data points are colored according to their cluster assignments. Horizontal and vertical gridlines are added to the plot to enhance readability. The plot includes labeled axes, a title, and a color bar that indicates cluster labels. This process can be seen in Fig. 2 of the architectural diagram presentation.

Finally, the research examined key points to strengthen its contribution to the existing scientific knowledge base. Key aspects included hyperparameter optimization using density thresholds and input parameters, as well as experimental validation of results through training. Feature engineering was used to create accurate grade assignments, ratios and labels. By standardizing performance metrics, bias reduction and data scalability were achieved. Principal component analysis was used to eliminate noise and prevent overlapping clusters in multidimensional data, allowing clear presentation of study results. However, research highlights a notable strength of the Clique model in predicting learners' academic progress, making it an attractive choice for scholarly contributions. These are presented below:

- This behavioral analysis investigation provided excellent prediction and significantly improved the model analysis by applying the subspace clustering technique and fine-tuning hyperparameters until optimal results were achieved.
- Unlike numerous previous studies, this research adopted a unique strategy by predicting learners' academic progress based on their academic performance, which generally reflects their progress at different stages of academic learning. The scope is more focused and relevant.
- There are various characteristics associated with a situation, such as a student achieving one credit often triggers the transition to a BSc, driven by an invisible urge for advancement. In contrast, a student who receives a passing grade under the policy cannot advance, but there is a compelling reason in this policy that educational institutions should recognize. This unspoken pattern becomes clear through precise mapping in this study.
- A number of multidimensional data sets were used as part of managing multidimensional data for this study. These datasets were manipulated and analyzed using feature engineering techniques and principal component analysis, demonstrating their effectiveness and applicability in cluster analysis. This contributes positively to the existing scientific knowledge.

D. Feature Engineering

This study used feature engineering to ensure appropriate data management, which contributes to the accuracy and reliability of the results. The dataset initially contained multiple columns, which were transformed and used to create new features. Here's a step-by-step breakdown of the feature engineering process:

Step 1. Categorical Conversion (Mapping Grades to Numerical Values):

- Objective: To convert categorical grade information into numerical form, which is necessary for further numerical analysis.
- Implementation: A mapping dictionary (grade_mapping) was created where each grade (e.g., 'Fail', 'Pass', 'Credit', 'Distinction') was assigned a corresponding numerical value (0, 1, 2, 3). The 'Grade' column in the DataFrame was then mapped to these numerical values. Grade_mapping = {'Fail': 0, 'Pass': 1, 'Credit': 2, 'Distinction': 3}.

Step 2. Creation of a New Feature: Ratio of 'Yes' to Total Students:

- Objective: To introduce a new feature that captures the ratio of 'Yes' responses (e.g., passing students) to the total number of students for each grade category.

- Implementation: The ratio was computed by dividing the 'Yes' count by the total number of students ('No' + 'Yes'). This new ratio feature was added as a column ('Ratio') in the DataFrame. df['Ratio'] = df['Yes'] / (df['No'] + df['Yes']).

Step 3. Creation of a Label Column:

- Objective: To generate a binary label indicating whether the number of 'Yes' responses exceeds the number of 'No' responses for each grade category.
- Implementation: A binary label was created using np. where, assigning a value of 1 if 'Yes' responses were greater than 'No' responses, and 0 otherwise. This label was stored in the 'Label' column of the DataFrame. df['Label'] = np. where(df['Yes'] > df['No'], 1, 0).

Step 4. Feature Selection:

- Objective: To select specific columns as features for further analysis.
- Implementation: The script selected the columns 'Grade', 'No', 'Yes', and 'Ratio' as features (X), and the 'Label' column as the target (y). X = df [['Grade', 'No', 'Yes', 'Ratio']] y = df['Label'].

Step 5. Standardization of Features:

- Objective: To standardize the features, ensuring they have a mean of 0 and a standard deviation of 1. This step is crucial for algorithms that are sensitive to the scale of input data, such as PCA and Clique subspace algorithm.
- Implementation: The features in X were standardized using StandardScaler from sklearn. The standardized data was stored in X_scaled. scaler = StandardScaler () X_scaled = scaler.fit_transform(X).

Step 6. Dimensionality Reduction Using PCA (Principal Component Analysis):

- Objective: To reduce the dimensionality of the data, potentially improving the performance of clustering algorithms and enabling visualization in a 2D space.
- Implementation: PCA was applied to the standardized data, reducing it to 2 principal components. The transformed data was stored in X_pca. pca = PCA(n_components=2) X_pca = pca.fit_transform(X_scaled).

Step 7. Clustering Using Clique subspace algorithm:

- Objective: To group the data into clusters based on the transformed features, aiming to identify patterns in the data.
- Implementation: The Clique subspace algorithm was applied to the PCA-transformed data with 2 clusters. The predicted cluster labels were stored in y_pred. Clique subspace algorithm = Clique subspace algorithm (n_clusters=2, random state=42) Clique subspace algorithm. Fit(X_pca) y_pred = Clique subspace algorithm.

Step 8. Manual Label Adjustment:

- Objective: To ensure that the cluster labels align with the actual labels, particularly when the initial clustering might have assigned the labels inversely.
- Implementation: The accuracy of the initial clustering was checked. If the accuracy was below 0.5, the labels were swapped (1 for 0, and 0 for 1). if accuracy score (y, y_pred) < 0.5: y_pred = 1 - y_pred.

Step 9. Evaluation of Clustering Performance:

- Objective: To assess the accuracy of the clustering model.
- Implementation: The accuracy of the clustering compared to the true labels was calculated and printed. accuracy = accuracy score(y, y_pred) print(f"optimized Accuracy: {accuracy * 100:.2f}%").

In summary this systematic feature engineering process effectively transformed the original dataset, enabling the use of advanced clustering techniques to analyze and predict patterns in the data.

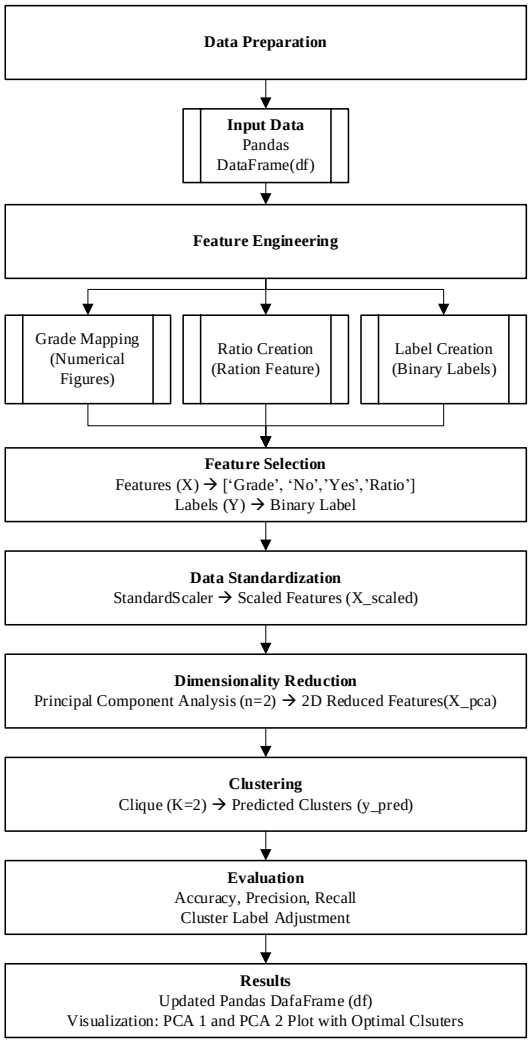

Fig. 2. Clique model for predicting learners' academic progress architectural diagram.

E. Dimensionality Reduction using PCA (Principal Component Analysis)

The experiment involving the use of principal component analysis in this study was conducted in three main phases, namely: feature reduction, clustering using clique, and cluster prediction adjustment.

In the first phase before plotting, the script uses PCA to reduce the dimensionality of the data from four features ("Grade", "No", "Yes" and "Ratio") to two main components. PCA is a linear technique that transforms the data into a new coordinate system in which the axes (principal components) correspond to the directions of maximum variance in the data. By reducing the data to two dimensions, we can visualize it in a 2D graph while retaining as much of the original variability as possible. The PCA(n_components=2) call creates a PCA object that is configured to reduce the data to two dimensions. The "fit transform(X_scaled)" method then applies this transformation to the standardized features ("X_scaled"), resulting in a two-dimensional array ("X_pca"). This array represents the data points in the new coordinate system defined by the first two principal components.

In the second phase, the Clique script applies subspace clustering to the two-dimensional PCA-transformed data. Clique Subspace is an unsupervised machine learning algorithm that divides the data into a certain number of clusters (in this case 2 clusters). The Clique Subspace (n_clusters=2, random state=42)" call initializes the Clique Subspace algorithm to form 2 clusters. The "fit(X_pca)" method is used to fit the model to the PCA transformed data ("X_pca"). The clustering process involves randomly initializing cluster centers (centroids) and iteratively adjusting them by minimizing variance within the cluster until convergence is achieved. After fitting, the model assigns each data point to one of the two clusters and the resulting cluster labels ("y_pred") are saved.

In the third phase of cluster prediction adjustment: Since Clique Subspace randomly assigns cluster labels, the script includes a step to align these predicted labels ("y_pred") with the actual labels ("y"). If the initial precision is below 50%, the script reverses the predicted labels ("y_pred = 1 – y_pred") to optimize the performance metrics (accuracy, precision, and recall). The accuracy of the initial clustering is calculated using the "Accuracy Score (y, y_pred)". If the accuracy is less than 50%, it indicates that the clusters are mislabeled. The script then reassigns the labels by flipping them to ensure that Cluster 1 is more likely to represent the positive class ("Yes" progression). The final step is to visualize the clustering results in a 2D scatterplot. This plot helps understand the distribution and separation of clusters in the transformed feature space. Fig. 3 and Fig. 4 below provide a visual representation of the diagrams.

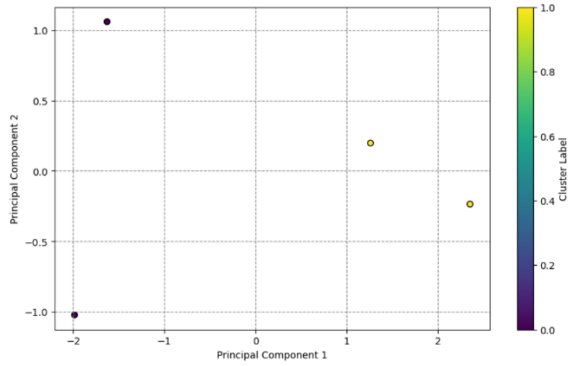

Fig. 3. Training data PCA.

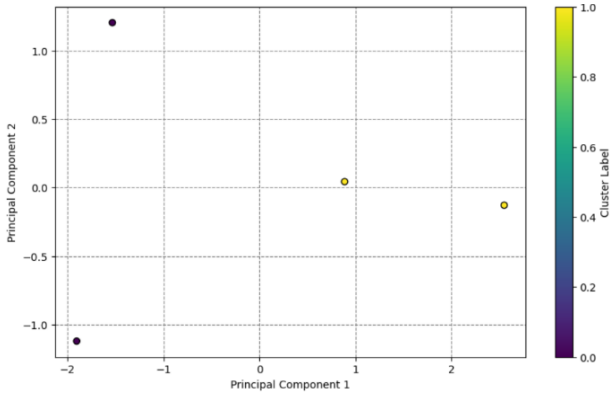

Fig. 4. Validation data PCA.

In summary, the graphs illustrate how well the Clique Subspace algorithm combined with Principal component analysis divides the data into two clusters. The spacing and distribution of points in the graph can provide insights into the natural grouping of data and the effectiveness of the clustering algorithm. Well-separated clusters with minimal overlap indicate that the algorithm has successfully identified different groups within the data. Aligning the principal component analysis axes with the inherent variability of the data ensures that clustering is based on the most informative aspects of the data even after dimensionality is reduced. This detailed process shows how machine learning techniques such as PCA and Clique Subspace are used together to group high-dimensional data and how the results can be scientifically visualized and interpreted.

V. Results

A. Introduction

This part describes the training and validation of the results using the ordinary clique model and the optimized clique model. It provides a comparative analysis of the results at different levels and subsequent discussions and conclusions.

B. Model Training and Validation

The model was trained and validated using the first and second folds of the dataset, respectively. You can see the results in Tables IV and V below.

TABLE IV. MODEL TRAINING RESULTS

Results without Modification						
S/No	Grade	No	YES	Ratio	Labels	Predicted Progression
0	2	156	127	0.449	0	1
1	3	16	269	0.943	1	1
2	0	261	0	0	0	0
3	1	1055	0	0	0	0
Performance Evaluation Metrics						
Precision	75%					
Accuracy	75%					
Recall	75%					
Results with Modification						
S/No	Grade	NO	YES	Ratio	Labels	Predicted progression
0	2	80	203	0.717	1	1
1	3	16	269	0.944	1	1
2	0	261	0	0	0	0
3	1	1055	0	0	0	0
Performance Evaluation Metrics						
Precision	98.50%					
Accuracy	98.90%					
Recall	98.50%					

TABLE V. MODEL VALIDATION RESULTS

S/No	Grade	NO	YES	Ratio	Labels	Predicted Progression
0	2	118	165	0.583	1	1
1	3	9	276	0.968	1	1
2	0	97	0	0	0	0
3	1	604	0	0	0	0
Performance Evaluation Metrics						
Precision	98.50%					
Accuracy	98.90%					
Recall	98.50%					

Training was performed using data fold 1 in Table II of the data partition, which considered the student's academic performance with random initial state probabilities. Training was performed by subjecting the data to two training trials with different hyperparameter tunings and then recording the results. In the first training attempt there was no modification, i.e. no parameter optimization because the implementation used the common and existing Clique algorithm and presented the results. Using the ordinary clique, the model showed an average performance of 75% for precision, 75% for accuracy, and 75% recall.

To confirm the consistency of our working model, we adjusted the dataset and performed some hyperparameter tuning as we observed new prediction patterns. The results showed better patterns with some common predictions from the previous model. We evaluated the performance of the model using the same clustering metrics such as precision, accuracy, and recall. The results with the model modifications showed exemplary performance with a precision value of 98.50%, an accuracy value of 98.90% and a recall value of 98.50%. This performance was consistent with the performance of the validation model. The model built in this study demonstrated efficiency and effectiveness when run on different datasets under different circumstances. During both training and validation, the model showed an accuracy rate of 98.90%. The applicability of the model developed on the entire dataset is consistent with the performance, which is well summarized.

C. Comparison with Related Work

In this study we examined various studies recently undertaken that used the same model and compared the results against our developed Model. The results are presented in Table VI as shown below:

TABLE VI. COMPARISON WITH OTHER STUDIES

Author	Algorithm	Accuracy	Precision	Recall
Oyugi et al. (2024)	Clique Model	98.90%	98.50%	98.50%
[25]	Clique Model	86.5%	85.99%	85.9%
[23]	Clique Model	93.9%	93.5%	94.3%
[15]	Clique Model	93%	93%	93%
[16]	Clique Model	98.6%	98.2%	97.5%
[18]	Clique Model	75%	75%	75%
[17]	Clique Model	92.0%	91.83%	90.12%

From Table VI above, our study is related to other studies, which can be seen from the result of this study. For example, the Clique algorithm was observed to display an overall cluster analysis with above-average performance of at least 75% for all metrics used in the evaluation, namely precision, recall and accuracy. It has also been observed that the accuracy, precision and recall displayed when using clique models are above 90% in most cases with the exception of a few instances reported by [18] in which the findings for all assessed metrics revealed 75%, and similar to [25]. which reported accuracy as 86.5%, precision as 85.99%, and 85.9% recall, this was slightly lower than the performance of clique models generated in other studies. According to our study, our model achieved 98.90% accuracy, 98.50% precision, and 98.50% recall. This was far higher than the value of all other related studies we examined in this study. However, after a careful comparison between our model and the other studies, it is entirely reasonable to conclude

that our model showed excellent performance in predictive cluster analysis. However, this can be attributed to the proper scope, use of future engineering that ensured sufficient and accurate future selection, proper management of the data, proper dimensionality reduction, and scientific cross-validation of the dataset that ultimately resulted in accurate prediction of learner progression.

TABLE VII. COMPARISON WITH OTHER MODELS

Author	Algorithm	Accuracy	Precision	Recall
Oyugi et al.	Clique Model	98.90%	98.50%	98.50%
[29]	K-Means	93%	94%	93%
[28]	BIRCH	91.1%	91.1%	91%
[31]	DBSCAN	91%	90%	90%

As shown in Table VII above, the accuracy rate for K-Means was 93%, for BIRCH was 91.1%, for DBSCAN was 91%, and for the Clique model was 98.90%. According to the other metrics, Clique scored the highest with a precision recall of 98.50 percent, while DBSACN scored the lowest with 90 percent. Clique recorded a maximum recall rate of 98.50 percent, while DBSCAN recorded a minimum recall rate of 90 percent. The above results show that our model performs better than the other models. Based on this study, it is advisable to consider using Clique.

In this study, the subspace Clique model has been analyzed and contrasted with other behavior analysis frameworks and similar methodologies employed in learner prediction research. Furthermore, to assess the challenges identified and the objectives achieved in this investigation, we engage in a thorough analysis and discussion of the pertinent literature. This research aims to enhance decision-making and learner management in educational contexts by predicting learner' academic progress based on their assessment scores. When evaluated against alternative models such as BIRCH, K-Means, and DBSCAN, the Clique subspace model demonstrated commendable performance. However, it is essential to address several issues that emerged from the studies conducted.

In the study of [28] Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm for behavior analysis prediction, the researcher observed that BIRCH was particularly effective for clustering large datasets. Nonetheless, it exhibited several limitations, especially in its ability to manage complex data distributions, overlapping clusters and sensitivity to the input parameters. In contrast, the CLIQUE algorithm demonstrated a superior capacity for handling large datasets and accommodating various data distributions, whether categorical or non-categorical, linear or non-linear which formed the features of our academic data. This adaptability positioned the CLIQUE model as a more efficient option for clustering in scenarios involving complex, multidimensional datasets within the scope of this study.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a widely utilized clustering algorithm recognized for its effectiveness in identifying clusters of arbitrary shapes and managing noisy datasets. However, it presents certain limitations, especially when applied to multidimensional data and scenarios involving varying densities [31]. Subspace Clique addresses these shortcomings by concentrating on dense subspaces, effectively managing different density levels, automatically determining critical dimensions, and offering enhanced scalability for predictive clustering.

K-means is an effective and straightforward algorithm designed for low-dimensional, approximately spherical clusters; however, it encounters difficulties when dealing with complex cluster shapes, Multidimensional datasets, and the presence of outliers [29]. The Subspace Clique model offers a solution to these challenges by considering dense subspace clustering techniques. This approach allows for the identification of arbitrary cluster shapes, the management of high-dimensional data, the adaptive determination of cluster structures, and resilience against outliers. Consequently, the subspace CLIQUE model emerges as a more appropriate option for predictive clustering in the context of academic progression as demonstrated in this study results.

Other researchers [16], [15], and [23] who employed the subspace clique method for behavior analysis reported performance levels of 98.6%, 93%, and 93.9% accuracy, respectively. These figures are marginally lower than the accuracy achieved by our model, which stands at 98.90%. Additionally, studies conducted by [25], [18], and [17] yielded accuracies of 86.5%, 75%, and 92%, all of which are again inferior to the performance of the clique model at 98.90%. The discrepancies in accuracy may be attributed to various challenges, including adaptive grid sizing, parameter optimization, and insufficient principal component analysis, all of which could enhance the accuracy of subspace CLIQUE algorithms. In contrast, our model effectively addressed these challenges through the computational efficiency of feature engineering, dimensionality reduction, and cross-validation.

D. Significance of the Research

This study used the subspace clique model to predict learners' academic progress patterns in multidimensional data. The results showed that the model was successfully able to predict learners' progress through clustering and establish hidden patterns that can help in decision making about learners. Student progress is crucial, particularly if they have started at a low level, such as a certificate or diploma in mainstream higher education, and was therefore crucial to this study. Although various studies have looked at the areas of learning in clustering, our study took a different approach and was limited to students' academic progress, which plays a key role in learners' progress as it can reflect the reality of progress in an academic environment Advice on teaching tasks as well as career guidance for the future can be used. It is right that we keep learners under control so that they do not lose track of their progress, as many students drop out of university due to poor performance, which results in them being unable to progress to the next level of learning. The behavioral patterns uncovered in

this study confirm the model's ability to perform predictive and accurate cluster analysis on an educational dataset.

VI. Conclusions and Future Work

The study aimed to predict learners' academic progress using subspace clustering in multidimensional data. However, the study was limited to a grid-based subspace clustering method using the clique model. To achieve this goal, we improve the Clique algorithm through optimization and parameter tuning. These improvements were achieved through future engineering, principal component analysis, data standardization, and cross-validation. The introduction of subspace served to reduce, if not eliminate, noise while performing clustering, which is called the "curse of dimensionality," where the number of dimensions increases as data size increases, resulting in overlapping clusters. Principal component analysis and future engineering helped find a solution to the overlapping clusters. By integrating density-based, grid-based, and subspace clustering, CLIQUE detects clusters embedded in subspaces of high-dimensional data without requiring users to select the subspaces of interest. The algorithm provided an effective and efficient method for pruning the space of dense units to counteract the inherent exponential nature of the problem. However, there was a trade-off for pruning dense units in the sparse coverage subspaces. Although the algorithm was faster, there is an increased chance of missing clusters. Furthermore, while CLIQUE does not require users to select subspaces of interest, its susceptibility to noise and ability to identify relevant attributes depend heavily on the user's choice of unit intervals and sensitivity threshold. The intent of the research was to search the data to reveal learners' progress from one stage to the next, the results clearly demonstrated that it is possible to determine learners' progress in clustering using the subspace clique technique. In summary, subspace clustering is complex but efficient and effective in performing clustering assessment of multidimensional data regardless of its size, and can be used to perform analysis for critical decisions, especially determining learner progression. Subspace clustering can also be used in dimensionality reduction to simplify the complex nature of grid-based bottom-up clustering evaluation methods. Future research can explore different methods and potentially consider a spectral clustering learning approach that leverages clique prediction capabilities to learn more about behavior analysis. Future engineering is recognized as a significant contribution to this study, yet it is sometimes perceived as biased due to its absence of a standardized execution method, consequently, further research is imperative. Given that this study focused on human behavior analysis in educational learning environments, I suggest that future studies be conducted in various settings.

References

[1] Albaity, M., Mahmood, T., & Ali, Z. (2023). Analysis and applications of artificial intelligence in digital education based on complex fuzzy clustering algorithms. *Mathematics*, *11*(14), 3184.

[2] [2] He, H., Sun, B., Yang, Y., & Chen, J. (2022). A K-means optimization algorithm suitable for fast clustering of WebGIS massive data. In *Journal of Physics: Conference Series* (Vol. 2171, No. 1, p. 012069). IOP Publishing.

[3] Mazarbhuiya, F. A. A., & Shenify, M. (2023). Finding IoT Anomaly using Rough Fuzzy Periodic Subspace Clustering Approach.

[4] Jia, H., & Cheung, Y. M. (2017). Subspace clustering of categorical and numerical data with an unknown number of clusters. *IEEE transactions on neural networks and learning systems*, *29*(8), 3308-3325.

[5] Wenz, V., Kesper, A., & Taentzer, G. (2023). Clustering heterogeneous data values for data quality analysis. *ACM Journal of Data and Information Quality*, *15*(3), 1-33.

[6] Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53-62.

[7] Grün, B. (2019). Model-based clustering. In *Handbook of mixture analysis* (pp. 157-192). Chapman and Hall/CRC.

[8] Ma, F., Wang, C., Huang, J., Zhong, Q., & Zhang, T. (2024). Key grids-based batch-incremental CLIQUE clustering algorithm considering cluster structure changes. *Information Sciences*, *660*, 120109.

[9] Shetty, N., & Shirwaikar, R. (2013). A comparative study: BIRCH and clique. *Int. J. Eng. Res. Technol*, *2*, 2019-2023.

[10] Cao, M., Hu, Y., & Yue, L. (2023). Research on variable weight CLIQUE clustering algorithm based on partial order set 1. *Journal of Intelligent & Fuzzy Systems*, *44*(6), 9461-9473.

[11] Fatehi, K., Rezvani, M., & Fateh, M. (2020). ASCRClu: an adaptive subspace combination and reduction algorithm for clustering of high-dimensional data. *Pattern Analysis and Applications*, *23*, 1651-1663.

[12] Kwale, F. M. (2014). An Overview of VSM-Based Text Clustering Approaches. *International Journal of Advanced Research in Computer Science*, *5*(1), 69-73.

[13] Nayagam, S. C. (2015). Comparative study of subspace clustering algorithms. *Int. J. Comput. Sci. Inform. Technol*, *6*(5), 4459-4464.

[14] Zhu, J., & Liu, X. (2024). An integrated intrusion detection framework based on subspace clustering and ensemble learning. *Computers and Electrical Engineering*, *115*, 109113.

[15] Madran, U., & Soyoglu, D. (2023). Compatibility of Clique Clustering Algorithm with Dimensionality Reduction. *Appl. Math*, *17*(5), 839-849.

[16] He, H., He, Y., Wang, F., & Zhu, W. (2022). Improved Clique algorithm for clustering non-spherical data. *Expert Systems*, *39*(9), e13062.

[17] Li, X., Zhang, Y., Cheng, H., Zhou, F., & Yin, B. (2021). An unsupervised ensemble clustering approach for the analysis of student behavioral patterns. *Ieee Access*, *9*, 7076-7091.

[18] Leeds, D. D., Zhang, T., & Weiss, G. M. (2021). Mining course groupings using academic performance. In *International Conference on Educational Data Mining*.

[19] Chi, D. (2021, January). Research on the application of k-means clustering algorithm in student achievement. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)* (pp. 435-438). IEEE.

[20] Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry*, *15*(9), 1679.

[21] Hu, X., & Ye, N. (2023). The Design of College Students' Mental Health Analysis System Based on Human-Computer Interaction. *Innovation in Science and Technology*, *2*(6), 34-42.

[22] Budler, L. C., Gosak, L., & Stiglic, G. (2023). Review of artificial intelligence-based question-answering systems in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1487.

[23] Jain, N., Ghosh, S., & Ghosh, A. (2024). A parameter free relative density based biclustering method for identifying non-linear feature relations. *Heliyon*.

[24] Starczewski, A., Scherer, M. M., Książek, W., Dębski, M., & Wang, L. (2021). A novel grid-based clustering algorithm. *Journal of Artificial Intelligence and Soft Computing Research*, *11*(4), 319-330.

[25] Ge, Z., & Xia, Q. (2024). Research on action analysis and guidance in aerobics blended learning based on data mining. *Applied Mathematics and Nonlinear Sciences*, *9*(1).

[26] Guo, Y., Dietrich, F., Bertalan, T., Doncevic, D. T., Dahmen, M., Kevrekidis, I. G., & Li, Q. (2021). Personalized Algorithm Generation: A Case Study in Meta-Learning ODE Integrators. *arXiv preprint arXiv:2105.01303*.

[27] Peng, X., Feng, J., Zhou, J. T., Lei, Y., & Yan, S. (2020). Deep subspace clustering. *IEEE transactions on neural networks and learning systems*, *31*(12), 5509-5521.

[28] Liu, H. (2024). Pedagogical Integration of Core Literacy in College Physical Education and Health Courses Based on Data Mining Techniques. *Applied Mathematics and Nonlinear Sciences*, *9*(1).

[29] Mohd Talib, N. I., Abd Majid, N. A., & Sahran, S. (2023). Identification of student behavioral patterns in higher education using k-means clustering and support vector machine. *Applied Sciences*, *13*(5), 3267.

[30] Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, *12*(19), 9467.

[31] Chrisnanto, Y. H., & Abdullah, G. (2021). The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI). *Matrix: Jurnal Manajemen Teknologi dan Informatika*, *11*(1), 26-35.