An Unsupervised Ensemble Clustering Approach for the Analysis of Student Behavioral Patterns

M. Sri Lakshmi[1], Seemakurthy Hupesh Naga Ketan[2]

[1]Associate Professor, Department of Computer Science and Engineering, G.Pullaiah College of Engineering and Technology, India.

[2]PG-Scholar, Computer Science and Engineering, G.Pullaiah College of Engineering and Technology, India.

a

Abstract

Performance analysis of outcome based on learning is a system which will strive for excellence at different levels and diverse dimensions in the field of student's interests. This paper proposes a complete EDM framework in a form of a rule-based recommender system that is not developed to analyze and predict the student's performance only, but also to exhibit the reasons behind it. The proposed framework analyzes the students' demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers, and parents. Seeking the highest possible accuracy in academic performance prediction using a set of powerful data mining techniques. The framework succeeds to highlight the student's weak points and provide appropriate recommendations. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones.

Keywords: EDM framework,Formatting, Cleaning, Sampling, Machine learning

I. INTRODUCTION

Identifying patterns in student behaviour and adapting teaching methods accordingly is a crucial part of improving education. Finding various behavioural factors that have strong correlations with academic performance [1][6], analysing student learning behaviours to allow teachers to adjust teaching schedules for better outcomes, and providing early warnings to students who may fail examinations are all examples. [7][10], modelling the mobility ow of students on campus to support the reasonable allocation of resources by administrators, detecting students' anomalous behaviours to allow managers to take timely preventative measures, and determining social networks from behavioural patterns.Dongxiao Yu served as associate editor and oversaw the manuscript review process, giving final approval before publication. Research in this area shows that these adjustments can have a major impact on classroom instruction.Most of these studies are conducted through questionnaire surveys directed at targeted groups of students in controlled environments. The data collection approach has some drawbacks, though. In the first place, surveys are conducted on a scheduled basis, such as once per academic year or semester, making it impossible to capture students' current state with this method. Negative outcomes could occur if students exhibiting unusual behaviour patterns are not identified quickly [11, 13].Second, the collected data may contain noise or false information that biases the analysis results, as either students exhibiting anomalous behaviours or normal students may not carefully ll out the survey. As a third point, designing a questionnaire that can capture enough

data for conducting a thorough investigation of students' routines and habits. Due to these constraints, this approach to data collection is not ideal.Ineffective and expensive. Thanks to advances in computing power, universities now have access to databases containing a wealth of information about student behaviour that can be mined for insights in real time.

Machine learning algorithms, which are the basis for many of the most common approaches, can be broken down into three broad classes: supervised, semisupervised, and unsupervised. To determine which class an unobserved student belongs to, supervised methods need labelled student data and the training of a classification model. Semisupervised methods create a model to discover the typical characteristics of individuals from a single group. When an individual's characteristics deviate too greatly from those of the class's representatives, that person is flagged as not belonging to that group.Privacy concerns prevent us from sharing labelled student data, especially that of outlier students.Furthermore, student labels are constantly changing, necessitating dynamic updates to any model. Due to these limitations, supervised and semisupervised methods are not always feasible in practise. Alternatively, unsupervised approaches are frequently used in the real world because they do not rely on labels and instead fully exploit the nature of datasets to cluster instances.To analyse student behaviour patterns, unsupervised clustering algorithms applied to campus-generated behavioural data appears to be a promising new direction. Clustering algorithms should find several mainstream behavioural patterns for targeted management in addition to detecting anomalous patterns for exception warnings, and they should be user-friendly. Both the classical unsupervised clustering method k-means algorithms and the more recent density-based spatial clustering of applications with noise (DBSCAN) [14] are widely used in a variety of fields. In situations where the distribution of the data space is unknown, DBSCAN can automatically filter noise out of samples and find clusters of arbitrary shapes. However, DBSCAN produces clusters of varying sizes, with the largest cluster sometimes containing nearly all the samples and rendering a refine of the data space impossible. The k-means algorithm, on the other hand, is a distance-based partition clustering method that excels in spherical data spaces where the number of clusters must be specified in accordance with the application's requirements or the partition's metrics. Although effective, the k-means algorithm is vulnerable to sample noise, which can cause the cluster centroids to shift toward the noise and thus become less representative. Inspired by ensemble learning, we analysed the pros and cons of the two algorithms and proposed an ensemble clustering methodology that combines DBSCAN and k-means algorithms to better meet the needs of student management. Four steps make up the proposed framework: data gathering, feature extraction, cluster analysis, and finally, visualisation and evaluation. The extract-transform-load tool was used to compile six types of campus-generated behavioural data from various information management systems. These records are time-stamped events in a traditional time series. The two aspects of statistics and entropy are used to extract the characteristics of all conceivable forms of behaviour; statistical data represents the mean and standard deviation of the distribution of behavioural data, while entropy represents the regularity of behaviour. Features with low variance and redundancy are eliminated to reduce the curse of dimensionality.

II. METHODOLOGY

MACHINE LEARNING

Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by programmer. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results.

Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input, use an algorithm to formulate answers.

Machine Learning vs. Traditional Programming

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

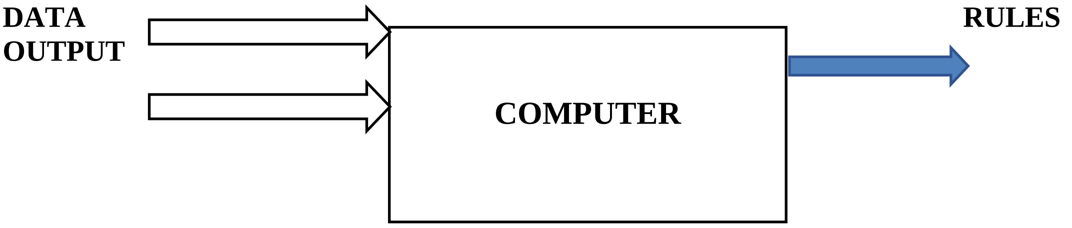

Figure 1: Machine Learning

Working:
Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation.

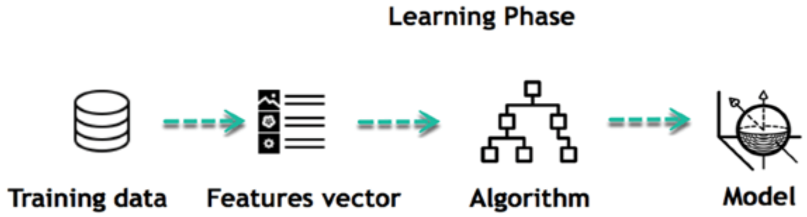

Figure 2: Working Process

Inferring
When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

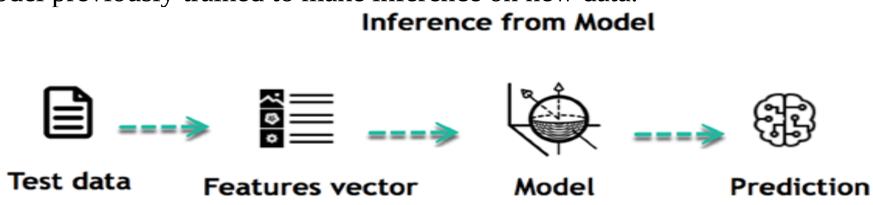

Figure 3: Inferring Process

The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

Machine learning Algorithms and where they are used:

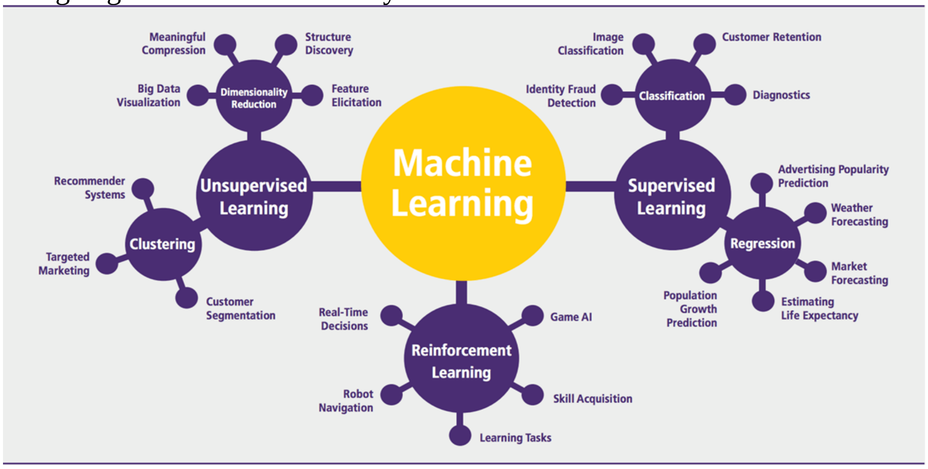

Figure 4: Machine learning Algorithms

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms

Supervised learning

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.

You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

- ✓ Classification task
- ✓ Regression task

Classification

Imagine you want to predict the gender of a customer for a commercial. You will start gathering data on the height, weight, job, salary, purchasing basket, etc. from your customer database. You know the gender of each of your customer, it can only be male or female.

The proposed framework firstly focuses on merging the demographic and study related attributes with the educational psychology fields, by adding the student's psychological characteristics to the previously used data set (i.e., the students' demographic data and study related ones). After surveying the previously used factors for predicting the student's academic performance, we picked the most relevant attributes based on their rationale and correlation with the academic performance. An Unsupervised Ensemble Clustering used in this research.

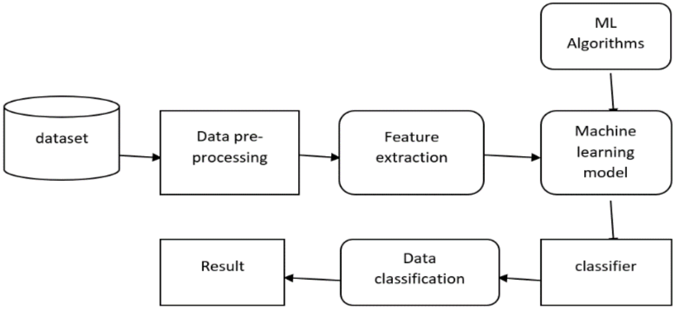

Figure 5: Proposed system Architecture

- **DATA COLLECTION**
- **DATA PRE-PROCESSING**
- **FEATURE EXTRATION**
- **EVALUATION MODEL**

DATA COLLECTION

Data used in this paper is a set of student details in the school records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called *labelled data*.

DATA PRE-PROCESSING

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

1. **Formatting**
2. **Cleaning**
3. **Sampling**

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database, and you would like it in a flat file, or the data may be in a proprietary file format, and you would like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonym zed or removed from the data entirely.

Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python.We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random Forest. These algorithms are very popular in text classification tasks.

DATA DESCRIPTION AND PRIVACY PROTECTION

i. DATASET:

This paper makes use of student behavioral data on consumption, library visits, and gateway log-ins. Using extract-transform-load tools, this information was compiled from various data stores. Each type of behavioral data consists of a list of records that are catalogued in a specific order. Time, place, dollar amount, and kind of transaction are the four data points that make up the consumption patterns. We only collect data on eating and shopping habits despite the fact that there are many other types of consumption behaviors; these two are the most common and provide the most insights. A person's eating habits can be better comprehended if they are broken down into three distinct categories based on when they eat: breakfast (6 am to 9 am), lunch (11:30 am to 2:00 pm), and dinner (4:30 pm to 8:30 pm). The act of entering a library is a significant learning activity, and the record of it includes the two attributes of time and location. Only one library was present in the dataset we used, so we nullified the location field. Deployed between the Internet and the campus LAN, the gateway system converts traffic between the two networks. When students use it to connect to the Internet from the campus network, the gateway system logs information about their logins (including when and from where they logged in), as well as the amount of time they spend online and the volume of data transferred. We also keep track of students' grade point averages (GPAs) as a proxy for their academic performance in addition to their behavior data.

ii. PRIVACY PROTECTION

In the course of our investigation, protecting your privacy is a top priority of ours. To begin, prospective students are asked during the enrollment process if they would be willing to contribute to the enhancement of the quality of their education by sharing the behavioral data they generate while attending the institution. Second, each student's ID number is hashed to protect their privacy. Third, an integer index is derived from the behavior time. We divide a day evenly into 48 bins and give each bin a numeric index between 1 and 48; the index of the bin can be substituted for the behavior time. A time of 8:10 a.m., for instance, can become 17 hours later. Last but not least, the location where the behavior occurred is transformed into a generic symbol. As a result of these steps, all of the behavior records that occurred at the same time and place are combined into a single file. The sum of the transaction amounts in the combined records provides insight into consumer behavior. Time spent logged in and bandwidth consumed at the gateway are new metrics. We get rid of duplicates in the library catalogue. In this way, the dataset is devoid of identifying information while still retaining sufficient detail to fuel a behavior clustering analysis.

EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation to avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated based on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

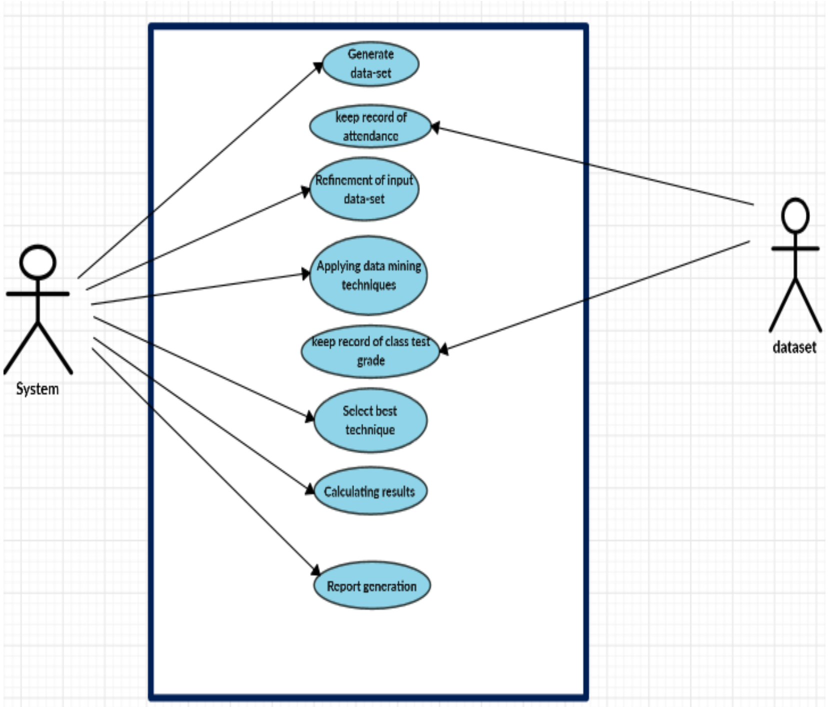
Generate data-set
keep record of attendance
Refinement of input data-set
Applying data mining techniques
keep record of class test grade
Select best technique
Calculating results
Report generation
System
dataset

Figure 6: Evaluation Model

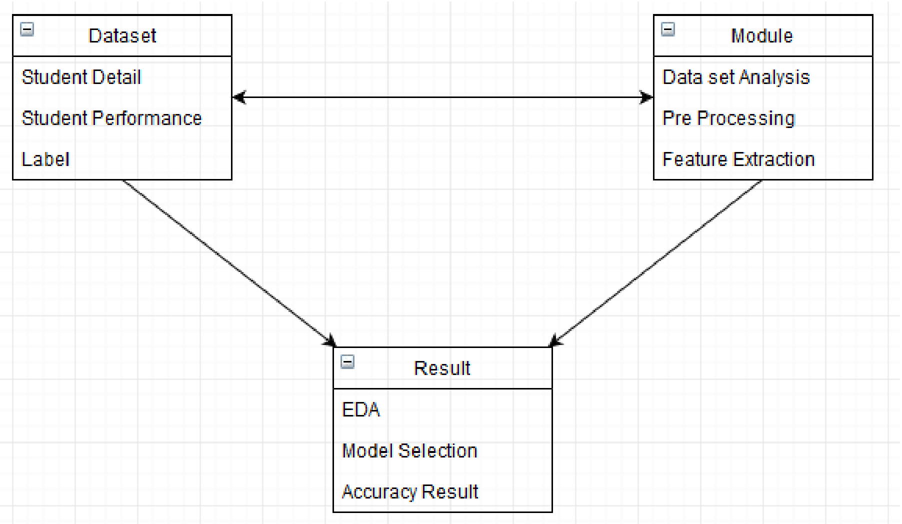

Figure 7: Evaluation of Class diagram

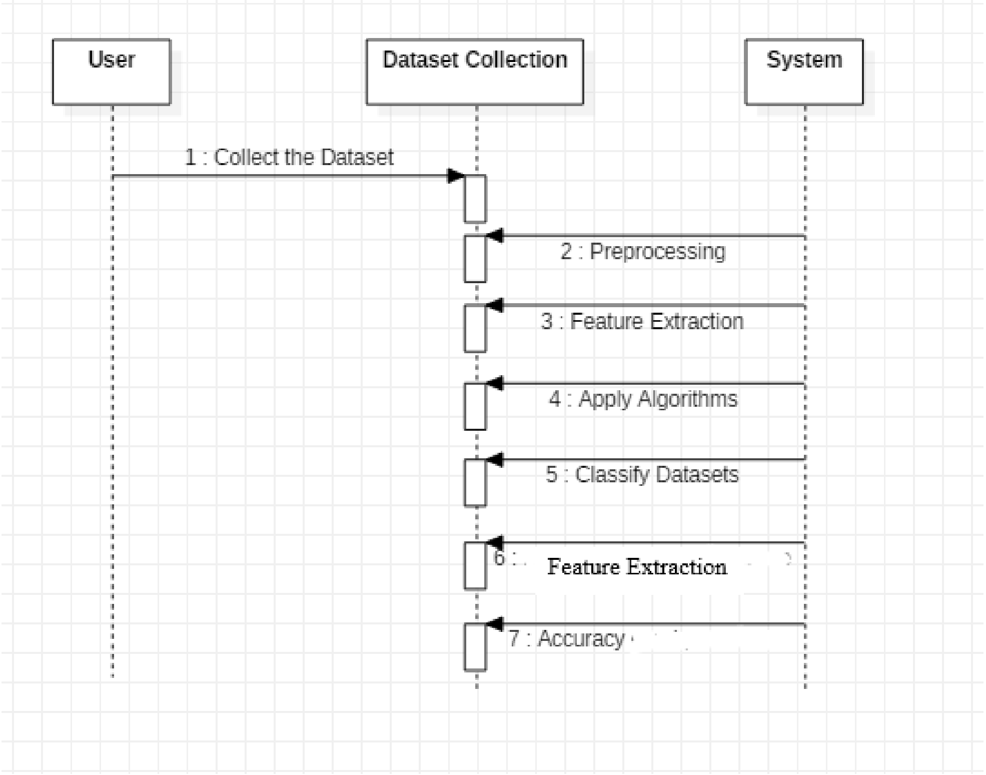

Figure 8: Sequence Diagram

ALGORITHM:

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
- The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

Types of Logistic Regression

Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, Logistic regression can be divided into following types −

Binary or Binomial

In such a kind of classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

Multinomial

In such a kind of classification, dependent variable can have 3 or more possible ***unordered*** types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.

Ordinal

In such a kind of classification, dependent variable can have 3 or more possible ***ordered*** types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3.

HOW logistic regression WORKS

Logistic regression uses an equation as the representation, very much like linear **regression**. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).

ADVANTAGES OF USING logistic regression

- ✓ Logistic regression is easier to implement, interpret, and very efficient to train.
- ✓ It makes no assumptions about distributions of classes in feature space.
- ✓ It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions.
- ✓ It not only provides a measure of how appropriate a predictor (coefficient size)
- ✓ Is, but also its direction of association (positive or negative)

Python

Python is a general-purpose, versatile and popular programming language. It's great as a first language because it is concise and easy to read, and it is also a good language to have in any programmer's stack as it can be used for everything from web development to software development and scientific applications.It has simple easy-to-use syntax, making it the perfect language for someone trying to learn computer programming for the first time.

Features of Python

A simple language which is easier to learn, Python has a very simple and elegant syntax. It's much easier to read and write Python programs compared to other languages like: C++, Java, C#. Python makes programming fun and allows you to focus on the solution rather than syntax. If you are a newbie, it's a great choice to start your journey with Python.

NUMPY

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the NumPy package, is the ndarray object.

The Numeric Python extensions (NumPy henceforth) is a set of extensions to the Python programming language which allows Python programmers to efficiently manipulate large sets of objects organized in grid-like fashion. These sets of objects are called arrays, and they can have any number of dimensions: one dimensional arrays are similar to standard Python sequences, two-dimensional arrays are similar to matrices from linear algebra. Note that one-dimensional arrays are also different from any other Python sequence, and that two-dimensional matrices are also different from the matrices of linear algebra, in ways which we will mention later in this text. All users of NumPy, whether interested in image processing or not, are encouraged to follow the tutorial with a working NumPy installation at their side, testing the examples, and, more importantly, transferring the understanding gained by working on images to their specific domain. The best way to learn is by doing – the aim of this tutorial is to guide you along this "doing."

III. RESULTS & DISCUSSION

Proposed Framework Steady

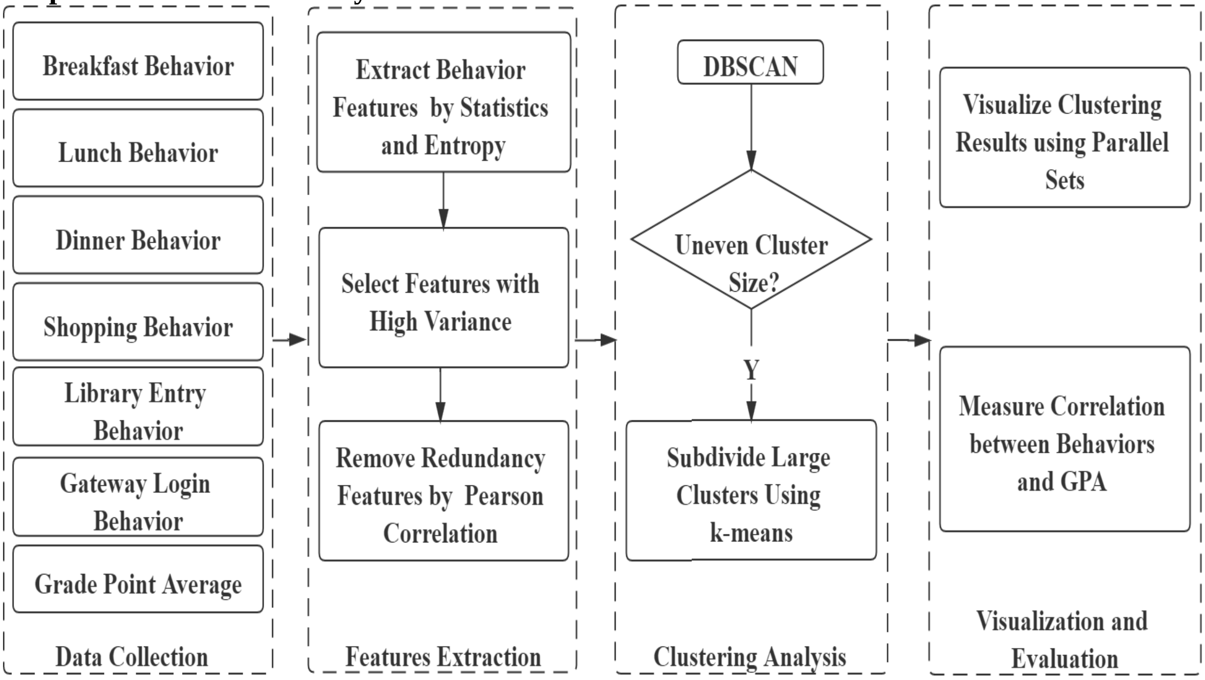
Breakfast Behavior
Lunch Behavior
Dinner Behavior
Shopping Behavior
Library Entry Behavior
Gateway Login Behavior
Grade Point Average
Data Collection
Extract Behavior Features by Statistics and Entropy
Select Features with High Variance
Remove Redundancy Features by Pearson Correlation
Features Extraction
DBSCAN
Uneven Cluster Size?
Y
Subdivide Large Clusters Using k-means
Clustering Analysis
Visualize Clustering Results using Parallel Sets
Measure Correlation between Behaviors and GPA
Visualization and Evaluation

Figure 9: Proposed Framework Steady

The six types of behavioral data analyzed in this paper were collected from 9024 undergraduates at a university in Beijing during the spring of 2019. The experiments are implemented using Python and scikit-learn libraries.

A. CLUSTERING RESULTS USING DBSCAN

To determine the parameters *Eps* and *MinPts* of DBSCAN, we plot a *MinPts*-dist graph for each type of behavior, where *MinPts* is set from 2 to 24. In the six graphs, the curves donot significantly change when *MinPts* is greater than 8, so we set *MinPts* to 8. The 8-dist graphs show that the optimal *Eps* values are 0.231 for breakfast behavior, 0.14 for lunch behav- ior, 0.175 for dinner behavior, 0.124 for shopping behavior,

0.082 for library entry behavior, and 0.09 for gateway login behavior. The clustering results of DBSCAN with the given values of *Eps* and *MinPts*, where 1 is the label of the noise cluster, the normal clusters are labeled with numbers starting from 0, and the number of students in each cluster is above its bar. For example, there are a total of -19 clusters numbered from1 to 17 for breakfast behavior, as shown in Fig. 6(a); noise cluster 1 contains 184 students who can be identified as those with unexpected behavioral patterns; clusters numbered 2, 4, 12, 13, 14, 15, 16 and 17 all contain relatively few students, less than 200, so the behavioral patterns they represent should be in the minority; clusters 0, 1, 3, 5, 6, 7, 8, 9, 10 and 11 all contain relatively large numbers of students, and they can represent students' mainstream behavioral patterns, especially clusters 0, 1, and 3. Based on the results, student services and man- agement departments should pay more attention to the noise clusters and minority clusters for early warnings and provide targeted services and management according to mainstream patterns. Lunch behavior has similar clustering results as breakfast behavior, as shown in Fig. 6(b). However, the clus- tering results of the other four types of behavior are not ideal; as shown in Fig. 6(c) and (d) and Fig. 7(a) and (b), their clusters 0 contain more than 90% of students. Although this phenomenon indicates that the behavioral patterns of the majority of students are relatively similar, it is necessary to further subdivide these clusters to understand behavioral patterns in detail. However, there is no one threshold that can be applied to all applications to determine which clusters need to be subdivided. It should be specified according to specific application requirements, here we set it to 80%.

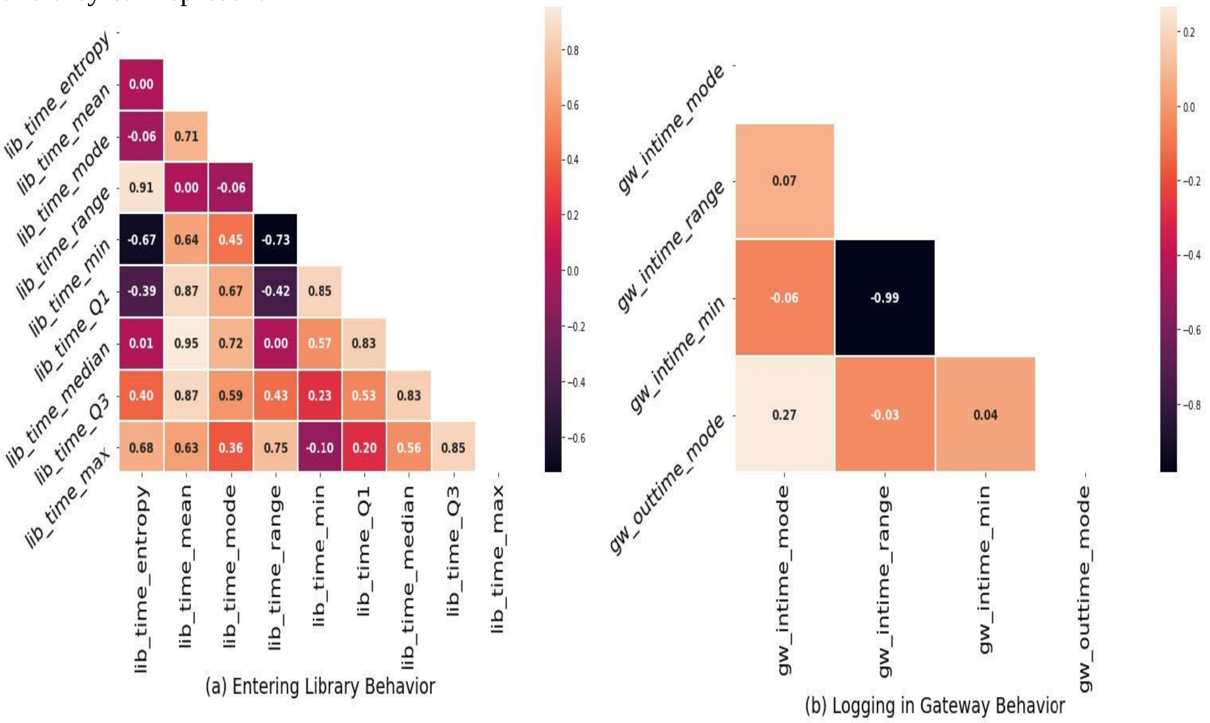

Figure10: Correlation coefficients between features of (a) library entry and (b) gateway login.

B. *SUBDIVISION RESULTS USING K-MEANS*

We use *k*-means for subdivision because the number of clus- ters *k* can be specified in advance by observing the curves of the four metrics, as well as the management requirements and prior knowledge. Additionally, this method can obtain more representative behavioral patterns than direct application of *k*-means to the original dataset because DBSCAN has filtered out the noise and very small clusters.

Here, we take dinner behavior as an example to illustrate how to determine the number of subclusters. The line charts of the four metrics are plotted as shown in Fig. 8, where *k* is set from 2 to 50. The inertia metric decreases as *k* increases, as shown in Fig. 8(a), and its scope becomes smooth when *k* is greater than 10, which indicates that it cannot significantly reduce the inertia value when the dataset is divided into more than 10 clusters, so the proposed number of clusters ranges from 2 to 10. Fig. 8(b) shows the curve of the silhouettescore. The proposed k values range from 2 to 6 since their silhouette scores are higher than others. The curve of CHI is shown in Fig. 8(c); its shape is similar to the inertia metric, and we can take the values from 2 to 10 as the candidates for *k*. The curve of DBI fluctuates considerably with respect to *k* and reaches the two lowest values when *k* equals 6 or

10. Simultaneous consideration of the four metrics shows that the optimal number of subclusters is six; the corresponding metric values are highlighted using a red vertical line in these graphs. In the same way, the optimal numbers of subclusters for cluster 0 of shopping behavior, library entry behavior, and gateway login behavior are determined to be six, five and four, respectively. In practice, we can also introduce management requirements to determine the optimal number. The final clustering results of these four types of behav- iors after subdividing cluster 0 with the given *k* are shown in Figs. 9 and 10, in which the clusters suffixed with '_DBSCAN' are noise clusters and minority clusters gener- ated by DBSCAN, while clusters suffixed with '_KMEANS' are the subclusters subdivided using *k*-means. The number of students in each cluster is above the bar. The final result not only retains the noise and small clusters but also subdividesthe large clusters into basically uniform subclusters.

C. VISUALIZATION OF THE CLUSTERING RESULTS

To intuitively understand the clustering results, parallel sets are introduced to visualize them. Parallel sets are a method for the visualization of categorical data, in which an axis represents a behavioral feature, the boxes in the axis repre- sent the feature value categories, and the thickness of each curved line represents a quantity that is repeatedly subdi- vided by category. By observing the result, we can under- stand the distribution of the behavioral features of every cluster and the difference between clusters. We take dinner behavior as an example to illustrate the visualization effect,

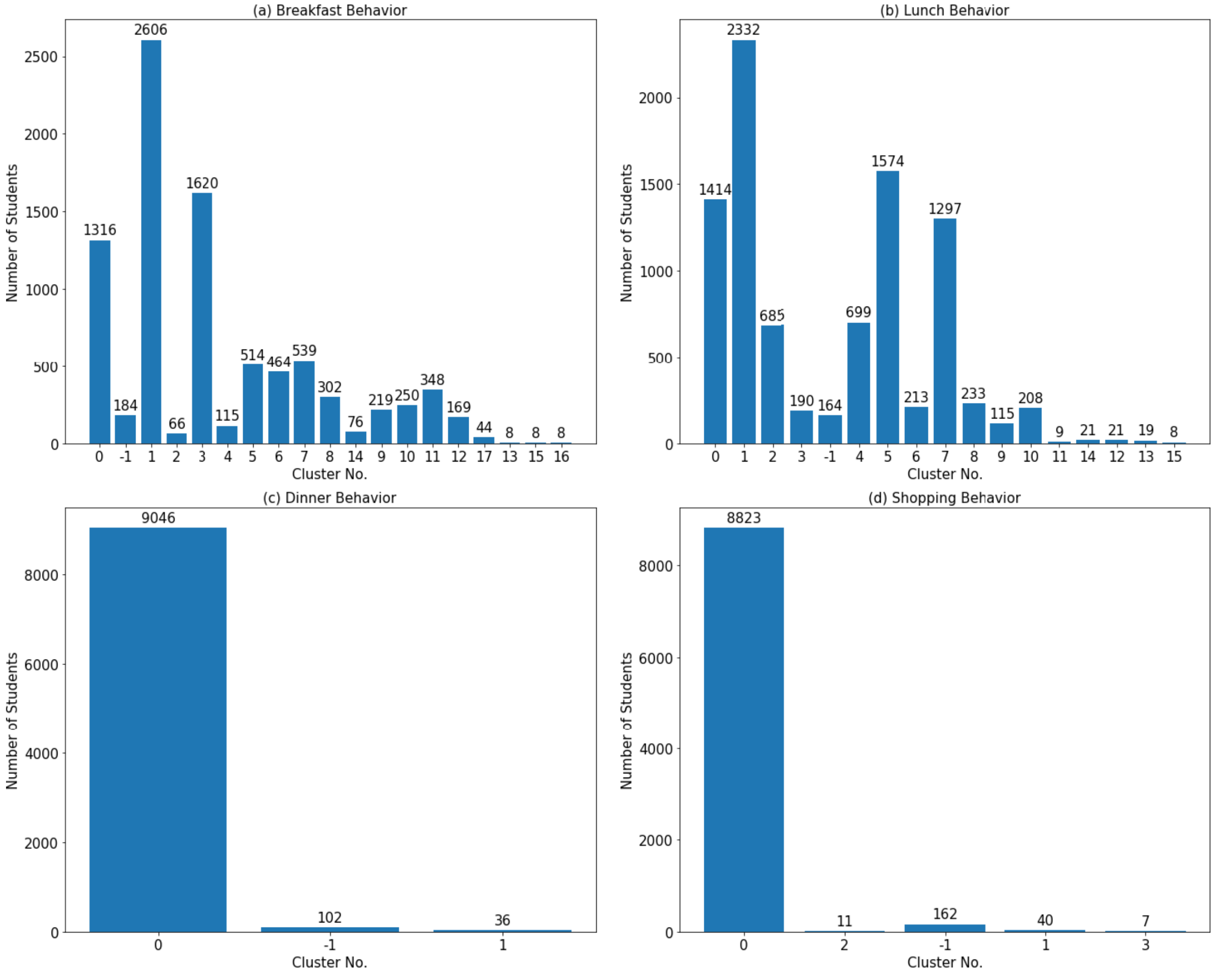

Figure11: Initial clustering results of (a) breakfast behavior, (b) lunch behavior, (c) dinner behavior, and (d) shopping behavior using DBSCAN.

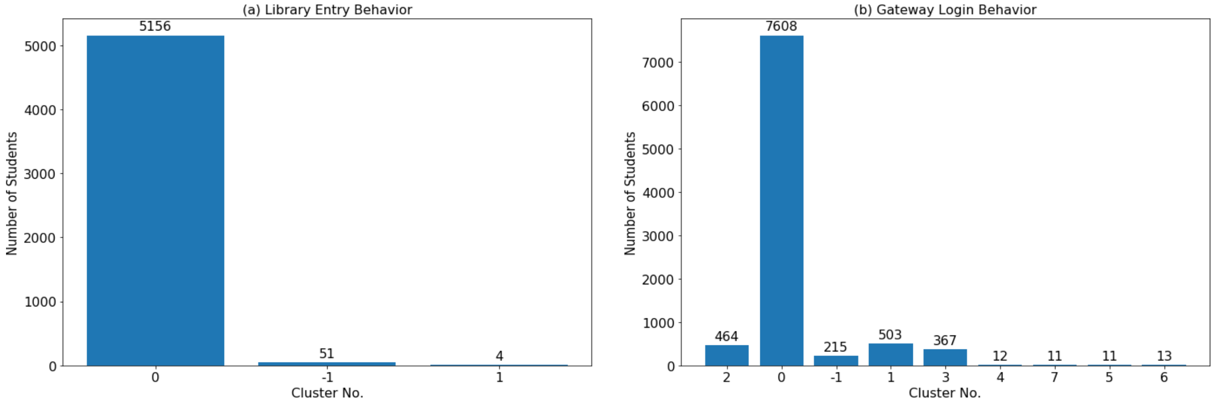

Figure12: Initial clustering results of (a) library entry and (b) gateway login using DBSCAN.

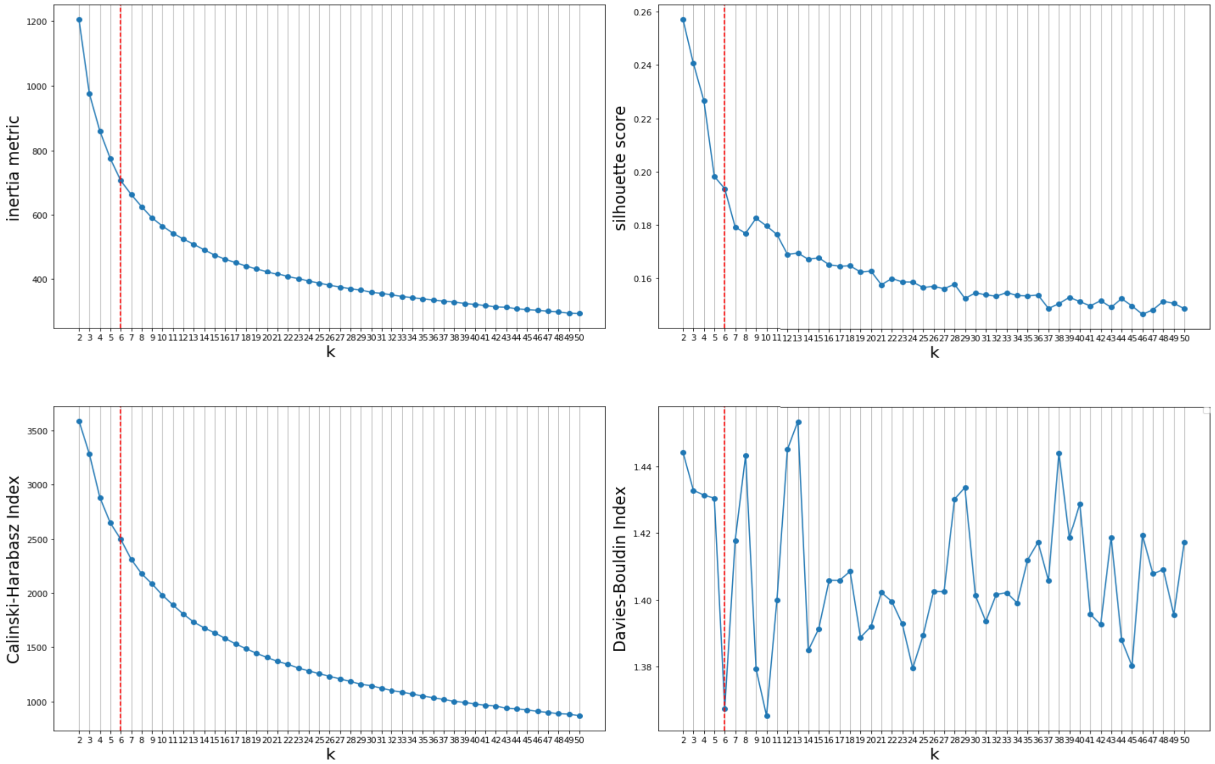

Figure13: Line charts of the four metrics for determining the number of subclusters of dinner behavior.

Table 1:Test case related to Dataset

SL #	TEST CASE NAME	DESCRIPTION	STEP NO	ACTION TO BE TAKEN (DESIGN STEPS)	EXPECTED (DESIGN STEP)	Test Execution Result (PASS/FAIL)
1	Excel Sheet verification	**Objective:** There should be an excel sheet. Any number of rows can be added to the sheet.	Step 1	Excel sheet should be available	Excel sheet is available	Pass
			Step 2	Excel sheet is created based on the template	The excel sheet should always be based on the template	Pass
			Step 3	Changed the name of excel sheet	Should not make any modification on the name of excel sheet	Fail
			Step 4	Added 10000 or above records	Can add any number of records	Pass

IV. CONCLUSION

In conclusion, the evaluation of students' performances poses a significant challenge. It is essential that we defend ourselves against them. The work that was done for this thesis and reported in it indicates that machine learning techniques with supervised learning algorithms were used to understand the performance of the algorithm with respect to student records. Specifically, we analyzed the performance of students and divided it into three categories: high, average, and low with an accuracy of 64%.

REFERENCE

[1] A. H. Eliasson, C. J. Lettieri, and A. H. Eliasson, ``Early to bed, early torise! Sleep habits and academic performance in college students,'' *SleepBreathing*, vol. 14, no. 1, pp. 71_75, Feb. 2010, doi: 10.1007/s11325-009-0282-2.

[2] X. D. Keating, D. Castelli, and S. F. Ayers, ``Association of weekly strengthexercise frequency and academic performance among students at a largeuniversity in the united states,'' *J. Strength Conditioning Res.*, vol. 27, no. 7,pp. 1988_1993, Jul. 2013, doi: 10.1519/JSC.0b013e318276bb4c.

[3] M. Valladares, E. Duran, A. Matheus, S. Duran-Agueero, A. M. Obregon,and R. Ramirez-Tagle, ``Association between eating behavior and academicperformance in university students,'' *J. Amer. College Nutrition*,vol. 35, no. 8, pp. 699_703, 2016, doi: 10.1080/07315724.2016.1157526.

[4] J. Filippou, C. Cheong, and F. Cheong, ``Modelling the impact of studybehaviours on academic performance to inform the design of a persuasivesystem,'' *Inf. Manage.*, vol. 53, no. 7, pp. 892_903, Nov. 2016, doi:10.1016/j.im.2016.05.002.

[5] S. Ghosh and S. K. Ghosh, ``Exploring the association between mobilitybehaviours and academic performances of students: A context-aware trajgraph(CTG) analysis,'' *Prog. Artif. Intell.*, vol. 7, no. 4, pp. 307_326,Dec. 2018, doi: 10.1007/s13748-018-0164-6.

[6] Z. Yang, X. Mo, D. Shi, and R.Wang, ``Mining relationships between mentalhealth, academic performance and human behaviour,'' in *Proc. IEEESmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., ScalableComput. Commun., Cloud Big Data Comput., Internet People SmartCity Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*,San Francisco, CA, USA, Aug. 2017, pp. 1_8.

[7] T. Phan, S. G. McNeil, and B. R. Robin, ``Students' patterns of engagementand course performance in a massive open online course,'' *Comput. Edu.*,vol. 95, pp. 36_44, Apr. 2016, doi: 10.1016/j.compedu.2015.11.015.

[8] I. JO, Y. Park, J. Kim, and J. Song, ``Analysis of online behavior andprediction of learning performance in blended learning environments,''*Educ. Technol. Int.*, vol. 15, no. 2, pp. 71_88, 2014.

[9] G. Kostopoulos, S. Kotsiantis, N. Fazakis, G. Koutsonikos, andC. Pierrakeas, ``A semi-supervised regression algorithm for grade predictionof students in distance learning courses,'' *Int. J. Artif. Intell.Tools*, vol. 28, no. 4, Jun. 2019, Art. no. 1940001, doi: 10.1142/
S0218213019400013.

[10] D. Hooshyar, M. Pedaste, and Y. Yang, ``Mining educational data topredict Students' performance through procrastination behavior,'' *Entropy*,vol. 22, no. 1, p. 12, Dec. 2019, doi: 10.3390/e22010012.

[11] N. Iam-On and T. Boongoen, ``Improved student dropout predictionin thai university using ensemble of mixed-type data clusterings,'' *Int.J. Mach. Learn. Cybern.*, vol. 8, no. 2, pp. 497_510, Apr. 2017, doi:10.1007/s13042-015-0341-x.

[12] I. HarwatiR Virdyanawaty and A. Mansur, ``Drop out estimation studentsbased on thestudy period: Comparison between naive Bayes and supportvector machines algorithm methods,'' in *Proc. ICET4SD*, Yogyakarta, IN,USA, 2015.

[13] P. Aparicio-Chueca, I. Maestro-Yarza, and M. Domínguez-Amorós, ``Academicpro_le of students who drop out a degree. A case study of facultyof economics and business, UB,'' in *Proc. EDULEARN*, Barcelona, Spain,Jul. 2016.

[14] M. Ester, H. Kriegel, J. Sander, and X. Xu, ``A density-based algorithm fordiscovering clusters in large spatial databases with noise,'' in *Proc. 2nd Int.Conf. Knowl. Discov. Data Mining*, 1996, pp. 226_231.
[15] R. Kosara, F. Bendix, and H. Hauser, ``Parallel sets: Interactive explorationand visual analysis of categorical data,'' *IEEE Trans. Vis. Comput. Graph-ics*, vol. 12, no. 4, pp. 558_568, Jul. 2006, doi: 10.1109/TVCG.2006.76.
[16] Y. Cao, J. Gao, D. Lian, Z. Rong, J. Shi, Q. Wang, Y. Wu, H. Yao,and T. Zhou,``Orderliness predicts academic performance: Behaviouralanalysis on campus lifestyle,'' *J. Roy. Soc. Interface*, vol. 15, no. 146,Sep. 2018, Art. no. 20180210, doi: 10.1098/rsif.2018.0210.
[17] H. Yao, D. Lian, Y. Cao, Y. Wu, and T. Zhou, ``Predicting academicperformance for college students: A campus behavior perspective,'' *ACMTrans. Intell. Syst. Technol.*, vol. 10, no. 3, pp. 1_21, May 2019, doi:10.1145/3299087.
[18] F. Li, X. Long, S. Du, J. Zhang, Z. Liu, M. Li, F. Li, Z. Gui, andH. Yu, ``Analyzing campus mobility patterns of college students by usingGPS trajectory data and graph-based approach,'' in *Proc. 23rd Int. Conf.Geoinformatics*, Wuhan, China, Jun. 2015.
[19] M. J. Lesot, ``Outlier preserving clustering for structured data through kernels,''in *Proc. 29th Annu. Conf. German-Classi_cation-Soc.*, Magdeburg,Germany, 2005, pp. 462_469.
[20] S. Fan, P. Li, T. Liu, and Y. Chen, ``Population behavior analysis of Chinese university students via digital campus cards,'' in *Proc. IEEE Int. Conf. DataMining Workshop (ICDMW)*, Atlantic, NJ, USA, Nov. 2015, pp. 72_77.