International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Method for Unsupervised Ensemble Clustering to Examine Student Behavioral Patterns

[1]K. Shyam Sunder Reddy, [2]P. Rajya Lakshmi, [3]Dr. D. Maruthi Kumar, [4] P. Naresh, [5]Y. N. Gholap, [6] Dr. K. Gurnadha Gupta

Submitted: 02/12/2023 **Revised**: 10/01/2024 **Accepted**: 26/01/2024

Abstract: Identification of student behavior time to time for projected throughput in terms of performance in academics is the primary goal of any educational organization. Prediction of unconventional behavioral patterns may useful to the goal. Based which educational institutes build the learning modules and the respective support for the development of student performances. Many existed studies worked on it by means of conducting surveys, taking reports and used questionnaire are not sufficient for the objective. Hence, we proposed a frame work that can be integrated with the advanced algorithms for getting hidden patterns too. The proposed method used unsupervised clustering method and results can be refined with ensemble algorithms. We collected real time data when students are in the campus to getting better behavioral patterns. For we developed two approaches for extracting features of patterns with DBSCAN and K-Means algorithm. And also adapted density based spatial clustering techniques concepts based on statistical and entropy approaches. For the experimental purposed various types of patterns produced by the student behavior is used. With the final experimental results, we concluded that our frame work is better that the accuracy rate of 96.3% in abnormal students' behavioral patterns. With which the educational organizations improve the academic targets as per their goals. Empirical research shows a significant correlation between these behavioural characteristics and academic achievement. We also examine the relationship between each student's academic performance and that of other students who exhibit behaviors that are similar to his or her own, prompted by the social impact idea. The association is substantial, according to statistical testing.

Keywords: *Density based clustering, DBSCAN, k-means algorithm, entropy based approach, statistical analysis, student behavior, behavioral patterns.*

1. Introduction

In order to generate complete students who are performing fine academically and possess the skills they would need after completing their studies, it is crucial to forecast student outcome perfectly. This will assist reduce the symptoms of kids who are at a high risk of failing. It is crucial to be able to forecast student performance, especially early on, so that professors and universities may identify high-risk kids earlier. In order to provide a proper method for teaching and learning resources, works like Stapa et al. [1] have concentrated on analysing students learning environment. Therefore, as demonstrated by Zainuddin et al. [2], preventative interventions and early solution-giving to pupils are possible. A suitable rehabilitation programme, particularly in the area of programming, can be provided by the responsible party to students likely to earn a semester grade point average (SGPA) less than 2.5, for instance in the Frequently, this programme bases enrollment decisions only on SGPA. Berens et al. [3] examined the effects of developing an early detection system and found that it might assist universities and instructors better understand their students. In order to generate holistic kids who are academically brilliant and have wonderful traits, a strategy based on Industrial Revolution 4.0 technologies, such as ML and statistics, has to be planned. Universities' main goal is to increase graduates' employability by utilising the information they have learned, not merely to assure academic achievement for students [4,5].

One of the goals of national education is to produce holistic graduates in order to ensure high-quality national resources.

[1]*Department of CSE, Maturi Venkata Subba Rao (MVSR) Engineering College, Hyderabad.*
[2]*Assistant Professor, Department of CSE, TKR College of Engineering and Technology, Hyderabad*
[3]*Associate Professor, Dept of ECE, Srinivasa Ramanujan Institute of Technology, Ananthapuramu*
[4]*Assistant Professor, Dept of IT, Vignan Institute of Technology and Science(A),Hyderabad*
[5]*Assistant Professor, Dept of Information Technology, Army Institute of Technology, Pune*
[6] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522302,India.*
** Corresponding Author Email: changalaravindra@gmauil.com*

2. Related Works

Determining student behavioural designs and pleasing the necessary steps to optimise the edifying process are important tasks in the field of education. For instance, finding different behavioural issues with bold correlations with academic outcome[6], analysing student education attitudes to enable tutors to alter teaching schedules for good results. Related research has demonstrated that these actions can considerably raise educational standards. Many researchers employ a opinion poll survey approach to obtain data from certain students in particular situations in order to finish their studies. The technique of data collection does, however, have significant drawbacks. There might be severe repercussions if pupils with atypical behavioural tendencies are not found in a timely way.

Second, abnormally behaving students could purposefully give false information to look normal, whereas typically behaving students might not thoroughly complete the survey. As a result, the obtained data may hold polluted or incorrect material that biases the analytic findings.

Third, in order to create a feedback form that can gather enough data to fully analyse students' behavioural patterns, substantial expert knowledge is required. These drawbacks render this data collecting technique ineffective and expensive. As information technology has advanced, numerous sorts of precise student behavioural data generated on campuses are now maintained in databases, offering a more dependable and complete source for real-time behavioural analysis. The widely used methods are based on ML algorithms, which are divided into supervised, semi-supervised, and unsupervised techniques. To identify which class a hidden student belongs to, supervised techniques require branded student data and the exercise of a classification model. In semi-supervised techniques, a model is created to learn the characteristic traits of students who only belong to one class. When a student's traits deviate significantly from the typical features of the class, that student is labelled as not belonging to the group. Data on labelled students, particularly those of anomalous kids, is unavailable due to privacy issues. Additionally, student labels are always changing, thus any model must be continuously updated. These elements make it challenging to use supervised and semi-supervised techniques in real-world settings.

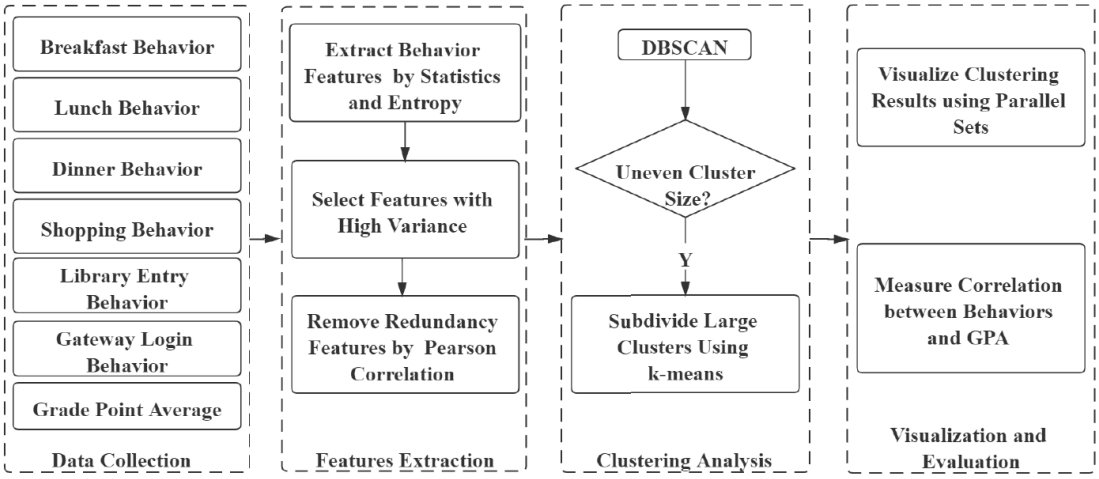


Fig 1. Working mechanism of the frame work.

Unsupervised techniques, in contrast, do not require labels and fully take advantage of the ability of datasets to cluster instances, making them popular in real-world settings. As stated above, using unsupervised clustering algorithms to manage behavioural data generated on campus is a potential route for examining students' behavioural patterns. Clustering algorithms must uncover multiple common behavioural patterns for focused management in addition to detecting anomalous behavioural patterns for exception alerts in order to fulfil the demands for specialised services and management. They must also be simple to use. Two traditional unsupervised clustering techniques that are frequently utilised in various disciplines are DBSCAN (density based spatial clustering) of applications with noise [14] and k-means algorithms. DBSCAN is appropriate in situations where it is uncertain how the data space is distributed since it can automatically remove noise from samples and locate clusters of any shape. However, DBSCAN produces clusters of varying sizes; in certain extreme circumstances, the biggest cluster contains nearly all of the samples, which prevents further exploration of the data space. Additionally, the k-means algorithm is a distance-based partition clustering technique that excels in spherical data spaces and requires the number of clusters to be predetermined in

accordance with the application's needs or partition metrics. Using the extract-transform-load method, six distinct forms of behavioural data generated on campus were gathered from various information management systems. These data are typical time series data made up of time-stamped occurrences. The central tendency and dispersion of the distribution of behavioural data are represented by statistical information, while the regularity of behaviour is represented by entropy. These two elements are used to extract the characteristics of each type of behaviour. characteristics with little variation and duplicate characteristics ought to be eliminated in order to relieve the effects of dimensionality.

References	Attributes	Prediction Model Output
[15]	Academic	CGPA, Student Category
[25]	Academic, Demographic, Psychological	Mean Score for Course, Student Category
[3]	Academic, Demographic, Institutional	Complete Studies, Did Not Complete Studies
[22]	Demographic, Institutional, Psychological	Student Category
[26]	Demographic, Institutional, Psychological	Fail, Pass
[27]	Academic, Psychological	Course Marks
[28]	Academic, Demographic	CGPA
[29]	Academic, Demographic, Institutional, Psychological	Fail, Pass
[12]	Demographic, Institutional, Psychological	Student Category
[21]	Academic	Low, High Performance
[20]	Academic, Demographic, Psychological	Low, High Performance
[18]	Academic, Institutional	Final CGPA
[24]	Demographic	Fail, Pass
[23]	Psychological, Institutional	Academic standing
[14]	Academic, Demographic, Institutional	Drop out or complete the course
[19]	Academic, Psychological	Final CGPA dan Student Category
[8]	Psychological	Student Category
[16]	Academic	Finish Studies or Drop-out

Table 1. Summary of various works for prediction of student behavioral patterns performance.

The major contributions of our work are as follows.

1) Various behavioral patterns of real time data belong students were gathered then extracted features of each pattern. For which we used entropy, central tendency and dispersion perspectives of behavioral patterns.

2) Developed a framework based on unsupervised clustering based ensemble techniques. For which we adapted the concepts of K-means and DBSCAN algorithms. The end results were satisfiable in predicting abnormal behavioral patterns of the students.

3) Correlation was made based on GPA metrics along with various patterns and performance in academics.

3. EXAMINATION OF STUDENT BEHAVIORAL PATTERNS

Three kinds of research on student behavior may be made: supervised, semi-supervised, and unsupervised techniques. The goal of supervised techniques is to analyze behavioral data in order to pinpoint student categories like academic achievement and mental health.

To identify anomalous patterns that significantly deviate from typical patterns, semi-supervised techniques are typically utilised. Unsupervised techniques, which do not need labelled data, are frequently employed in practise because they attempt to extract knowledge from vast volumes of behavioural data to improve decision-making. For instance, [18] put out a graph-based strategy to comprehend students' movement behaviours on campus. By using the DBSCAN algorithm to extract dwell points, this method creates a behaviour graph where the nodes are dwell points and the edge values are the times between dwell points. The k-core technique is used to identify students' typical behaviour based on the graph, and the proximity centre degree is used to identify anomalous behaviour. To fully characterise the data, Reference [19] suggested using the outlier preserving clustering algorithm (OPCA), a kernel-based clustering approach, to find both significant and deviant behaviours. The challenge with hierarchical algorithms is determining the best circumstances for merging or split.

Consumption, library usage, and gateway login behaviours are among the student behavioural variables

that were used in this study. The extract-transform-load (ETL) tools were used to get these data from several databases. Each type of behavioural data consists of a set of sequentially indexed recordings. The four elements of the consumption behaviour records are time, location, transaction amount, and transaction type. Despite the fact that there are several other forms of consumption behaviours, we only track eating and shopping habits because they are the most common and provide a wealth of data.

The two qualities of time and place are recorded in the learning action of entering a library.

We delete the location element because the dataset we utilised only contains one library. A protocol converter that is installed between the Internet and the campus local network is the gateway system. When using the campus network to access the Internet, students must log into the gateway system so that the gateway system can keep track of the login time, logout time, login location, length of Internet access, and amount of network traffic. We also gather the kids' GPAs to indicate their academic success in addition to the behavioural information.

3. Behavior Features

The extraction of characteristics from a vast quantity of behavioural data poses a significant obstacle to the clustering analysis of behavioural patterns. In this study, we extract features using statistics and entropy, and then we choose features using variance and correlation analysis.

No.	Feature Names	Description of Features
1	bf_frequency	Frequency of breakfast
2	bf_Loc_entropy	Shannon entropy of locations of breakfast
3	bf_Time_entropy	Shannon entropy of time of breakfast
4	bf_Time_mean	Average time of breakfast
5	bf_Time_mode	Most frequent time for breakfast
6	bf_Time_range	Difference between the earliest time and the latest time of breakfast
7	bf_Time_min	Earliest time of breakfast
8	bf_Time_Q1	25% of the breakfast time values are earlier than this time
9	bf_Time_median	50% of the breakfast time values are earlier than this time
10	bf_Time_Q3	75% of the breakfast time values are earlier than this time
11	bf_Time_max	Latest time of breakfast
12	bf_Trans_mean	Average transaction amount of breakfast
13	bf_Trans_mode	Mode of transaction amount of breakfast
14	bf_Trans_range	Difference between the minimum and maximum transaction amounts of breakfast
15	bf_Trans_min	Minimum transaction amount of breakfast
16	bf_Trans_Q1	25% of the transaction amount values of breakfast are less than this value
17	bf_Trans_median	50% of the transaction amount values of breakfast are less than this value
18	bf_Trans_Q3	75% of the transaction amount values of breakfast are less than this value
19	bf_Trans_max	Maximum transaction amount of breakfast
20	bf_Trans_sd	Standard deviation of the transaction amount of breakfast

Table 2. Student behavioral patterns features after breakfast.

Nominal and numeric qualities can be used to categorise behavioural data. The only nominal element is behavioural location; all other attributes are numerical. The range, mode, and mean are used to convey the distribution of values for numerical characteristics, while the minimum, Q2, median, Q3, and maximum are used to indicate the distribution's dispersion. We compute the Shannon entropy for nominal attribute behavioural location to assess the regularity of behaviour from the spatial dimension. We also calculate the entropy from the temporal dimension because behavioural time has been converted to an integer index. The definition of the Shannon entropy is (1):

$$K = -\sum_{j}^{n} m(j) log m(j) \quad \text{----------Eq (1).}$$

For instance, the student eats breakfast in day day same time more which leads zero entropy. We instead utilise the standard variance to assess the stability of variables with constant values, such as operation amounts, Internet access times, and system traffic, because it is challenging to calculate the Shannon entropy for these qualities. A smaller variance denotes a more stable state, similar to entropy. We also calculate the frequency to show how frequently each behaviour occurs.

The 20 breakfast behaviour traits are shown in Table 1 above.

The characteristics of breakfast behaviour are shared by the other three categories of consumption, including lunch, supper, and shopping. Prefixes 'lu', 'di', and'sp' are used to denote their characteristics, respectively. We disregard the entropy of location for shopping behaviour because there aren't many retail venues in our dataset. In addition to frequency, the library entrance behaviour contains nine time-related characteristics prefixed with 'lib', such as lib_frequency and lib_time_entropy, that have the same meaning as those of breakfast behaviour.

We extract features for login time, logout time, location, length of Internet access, and network traffic ow characteristics for the gateway login behaviour. These features are prexed with 'gw_intime', 'gw_outtime', 'gw_loc', 'gw_dura_acces', and 'gw_traf_ow', respectively. There are 38 characteristics in all, including frequency and location entropy.

4. Feature Selection with Analysis of Variance and Correlation

For every behaviour, there are hundreds of characteristics, which can lead to the expletive of dimensionality in cluster algorithms since the distance used with the algorithm to determine sample resemblance may not be accurate for highdimensional data. We choose the best attributes by examining their

variation and association in order to get over this challenge.

A) VARIANCE

Variance, a metric for data dispersion, shows how skewed a distribution of data is. Values for a characteristic with limited variation typically fall extremely near to the mean, which can offer very little helpful clustering information. We eliminate the characteristics with low variance as a result. Fig. 2 displays the variation in the characteristics of consuming behaviour.

We discovered three phenomena by examining these features: (1) There aren't many differences in the students' consumption patterns, as seen by the variance of all characteristics being less than 0.1. (2) The transaction amount-related characteristics have a smaller variance than other features. The variation of the quantity-related characteristics of breakfast behaviour is about nil among the four categories of consuming behaviours, which makes sense given how similarly priced breakfast meals are. Because there are more lunch options and a wider range of costs than there are for breakfast, the variance of the amount-related characteristics for lunch behaviour is larger than it is for morning behaviour. The menu is often the same for lunch and dinner, but what's really remarkable is that the variation of the amount-related characteristics of dinner behaviour is much smaller than that of lunch behaviour. (3) Because of their relatively high variance, the frequency, location entropy, and time-related characteristics may be employed to depict different behavioural patterns. The differences between the characteristics of gateway login behaviour and library entrance behaviour are shown in Fig. 3.

The variations for nine out of the eleven characteristics of library entering behaviour are more than 0.03, which suggests that these characteristics have significantly distinct behavioural patterns. The characteristics gw_intime_mode, gw_intime_range, gw_intime_min, and gw_outtime_mode have a great degree of variability in how a gateway logs in.

The number of characteristics that have been chosen is 10, 6, 7, 6, 9 and 4, respectively, for the breakfast, lunch, dinner, shopping, library admission, and gateway login behaviours. We have set the variance threshold for these features at 0.02.

B) CORRELATIONS

If a feature can be derived from another feature, it may be redundant. In this part, we quantify how strongly one characteristic implies another using the Pearson correlation coefficient and then get rid of the duplicate features. Figures 4 and 5 provide the coefficient matrix of the features chosen in section IV-B1 so that we can easily comprehend how any two features are correlated with one another. As seen in Fig. 4(a), the characteristics bf_time_median and bf_time_mode, for instance, have a coefcient of 0.9 and exhibit strong correlation. We delete bf_time_median to decrease duplication since it has a smaller variance than feature bf_time_mode (0.023 vs. 0.03).

5. Proposed Work

This section outlines the suggested process for categorising and grouping student data in order to forecast academic achievement. Data preparation, data cleansing, clustering, and classification are the key stages. Below is an explanation of each phase of the recommended method.

As seen in Figure 1, there are six pre-processing techniques used: data integration, filtering, cleaning, transformation, and attribute selection. Following the collection of the data, the various datasets were combined depending on the number of students, and the resulting dataset had a total size of 200. The student was eliminated from the dataset during data cleaning if there were too many missing values for particular attributes, since this was likely the case if the student failed to provide the information by failing to respond to one of the surveys. As a result, there are several missing values in a certain area. Except for the learning programme result features, most of the values in the dataset are nominal. As a result, the nominal value was transformed into numerical values for data processing. This is significant since the prediction model in machine learning needs numerical numbers. In order to verify that the individual characteristics reflect regularly distributed data, this data was also normalised using StandardScaler. To guarantee that machine learning works successfully, this is crucial before beginning.

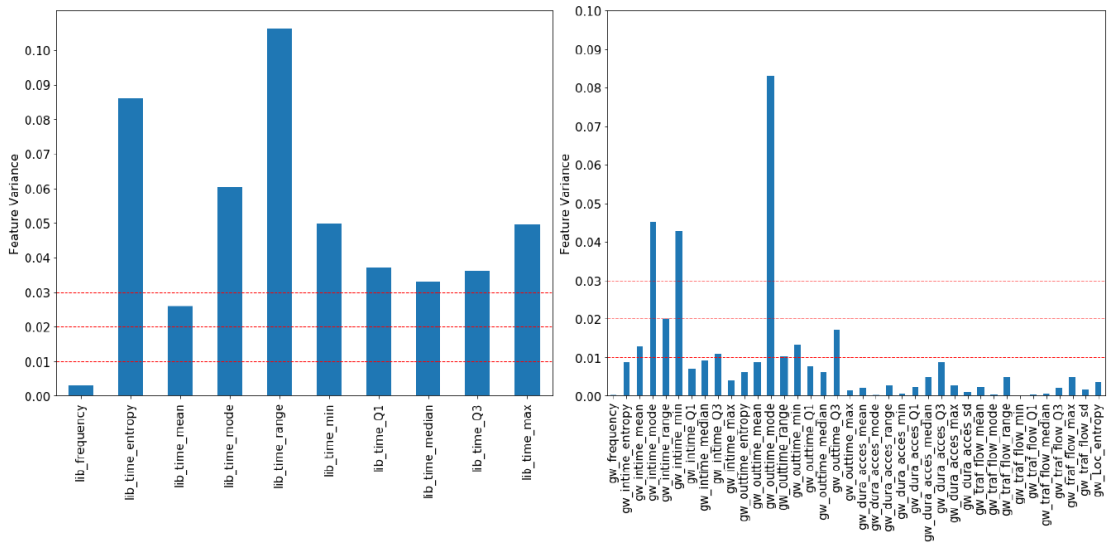
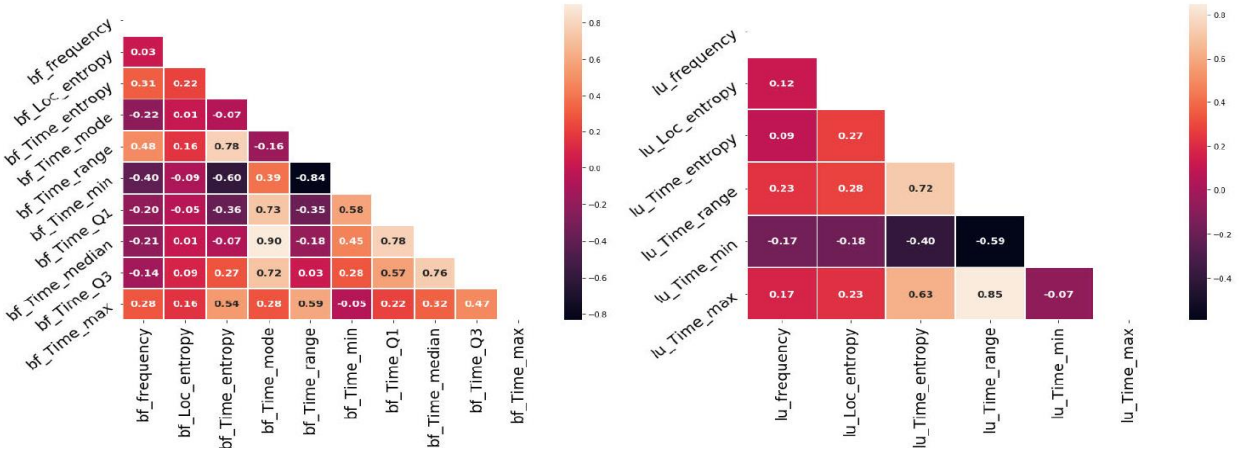


Fig 2. Variances of the features of library entry and gateway login.

Since clustering algorithms don't require students to provide labelled data, they are highly useful for identifying students' behavioural tendencies. The three categories of currently used and widely used clustering algorithms include partitioning, hierarchical, and density-based techniques. A data space is divided into k clusters using partitioning techniques, but these procedures are noise-sensitive and all the clusters have a convex form. Although nonconvex clusters can be found using hierarchical algorithms, these algorithms have a very difficult time defining a termination condition for when to stop the merge or division process. With just two parameters, density-based algorithms can find clusters of any shape and easily remove noise, but unevenly sized clusters, especially extremely big clusters, might not be suitable for student services and management.

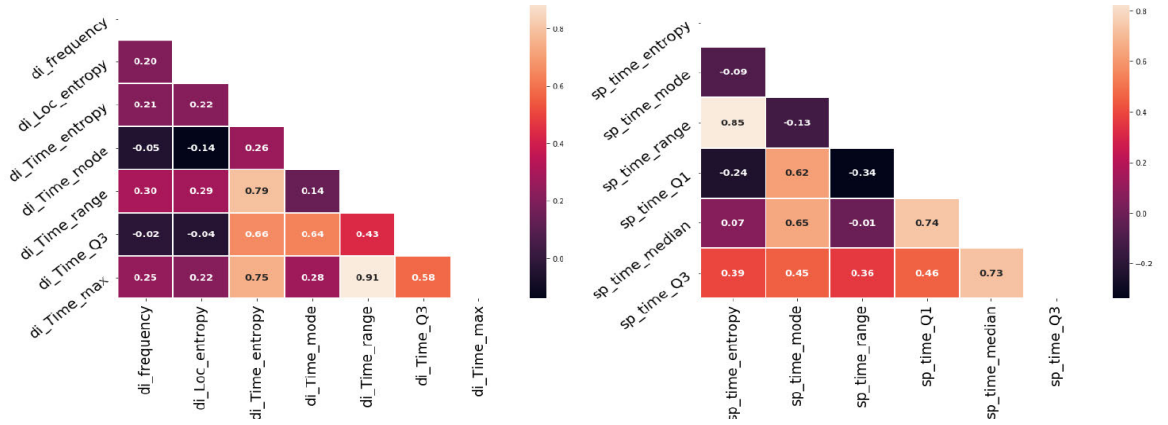
Ensemble clustering has tired a lot of interest as a practical technique to raise the standard of clustering. In order to identify student behavioural trends, we provide an ensemble clustering approach in this research. The essential concept is that the background first employs the density-based method DBSCAN to remove sound and create the first clustering, and then employs the k-means partition algorithm to separate the substantial clusters created by DBSCAN to provide the final result of clustering.



(a) Student behavior at breakfast

(b) Student behavior at lunch.

We may choose the number of subclusters in the subdivision process while concurrently taking the four metrics and application needs into account. Notably, additional subdivision is not required if the DBSCAN clustering result satisfies the application's criteria. To reach the final findings, the extremely big DBSCAN clusters can be substituted by the subdivided subclusters after subdivision.



(c) Student behavior at dinner (b) Student behavior at shopping mall

Fig 3. Correlation Summary of student behavioral patterns at various actives.

Students may utilise technology to help them develop better study habits and methods. We utilised a statistical procedure to determine which study habits had the most effects on academic success because there are many actions that may help students learn. Two models that addressed various aspects of performance self-perception and outcomes obtained were developed. After modelling the data using both current and former students and alumni, we found a general tendency towards actions and variables for current students that offer immediate advantages and behaviours and factors for alumni that foster higher-order thinking. The models presented in this study serve as the foundation for developing persuasive systems that enhance learning outcomes because they paint a more complete picture of how students' learning practises organically emerge. Aiming to influence the natural evolution of good students for those kids who are failing, potentially owing to behavioural shortcomings, persuasive methods for education can now be developed using this understanding.

6. Experimental Results and Analysis

A) DBSCAN ALGORITHM

Student behavioral patterns can be constructed by using MinPts distance graph with the parameters between 2 to 24, which can be measured by DBSCAN. But all graphs can not be altered when MinPts values is 8 and can be taken as parameters for the breakfast(0.14), lunch (0.125), dinner (0.124), shopping(0.08) and gateway(0.09) respectively is represented in Fig 6 and 7. The clusters noise numbers as 0,1 mentioned in bar chart.

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\} \text{--------Eq (2).}$$

From the Fig 6(b), it is clear that the similarities in cluster for the patterns lunch and breakfast. The findings of the other four forms of behaviour clustering are less than ideal, as can be seen in Figure 6(c) and (d) and 7(a) and (b), where clusters 0 comprise above 91% of the students. Any how this phenomena suggests the majority of scholars have comparatively same behavioural patterns, it is still important to more split these clusters in order to fully comprehend behavioural patterns. To decide which clusters need to be separated, there isn't a single criteria that can be used for all applications. It should be specified in accordance with the needs of the particular application; in this case, we make it as 82%.

B) K-MEANS ALGORITHM

The amount of clusters k may be stated beforehand by looking at the arcs of the four metrics, as well as the organization needs and past information, which is why we utilise k-means for subdivision. Additionally, because DBSCAN has cleaned out the noise and tiny clusters, this technique can find more typical behavioural patterns than applying k-means directly to the unique dataset.

Here, we use meal behaviour as an example to show how to calculate how many subclusters there are. Therefore, the suggested range of cluster sizes is 2 to 10. The curve of the silhouette score is seen in Fig. 8(b). As their silhouette scores are greater than others, the suggested k values vary from 2 to 6. Figure 8(c) depicts the CHI curve; it resembles the inertia metric in form, and values between 2 and 10 can serve as candidates for k. The two lowest values of the DBI curve are reached when k is equal to 6 or 10. The DBI curve fluctuates significantly with regard to k. The ideal number of subclusters, as determined by simultaneous evaluation of the four metrics, is six; in these graphs, the relevant metric values are denoted by a red vertical line. Similar results are obtained for cluster 0 of shopping behaviour, library entry behaviour, and gateway login behaviour, where the ideal number of subclusters is found to be six, ve, and four, respectively. To find the ideal number in practise, we may additionally include managerial criteria.The final solution separates the major clusters into essentially

uniform subclusters while also retaining the noise and minor clusters.

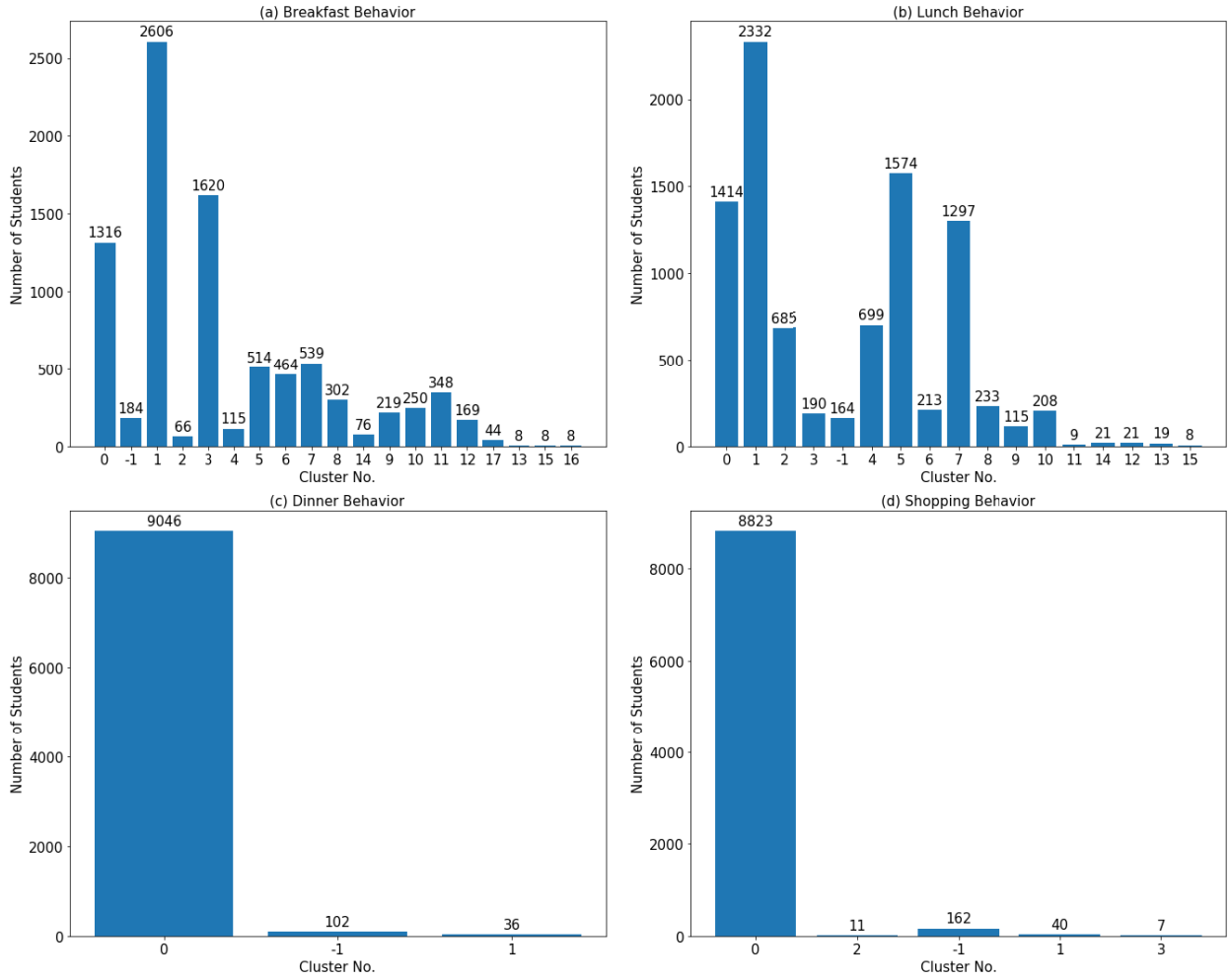


Fig 4. Various patterns result of Initial clustering using DBSCAN.

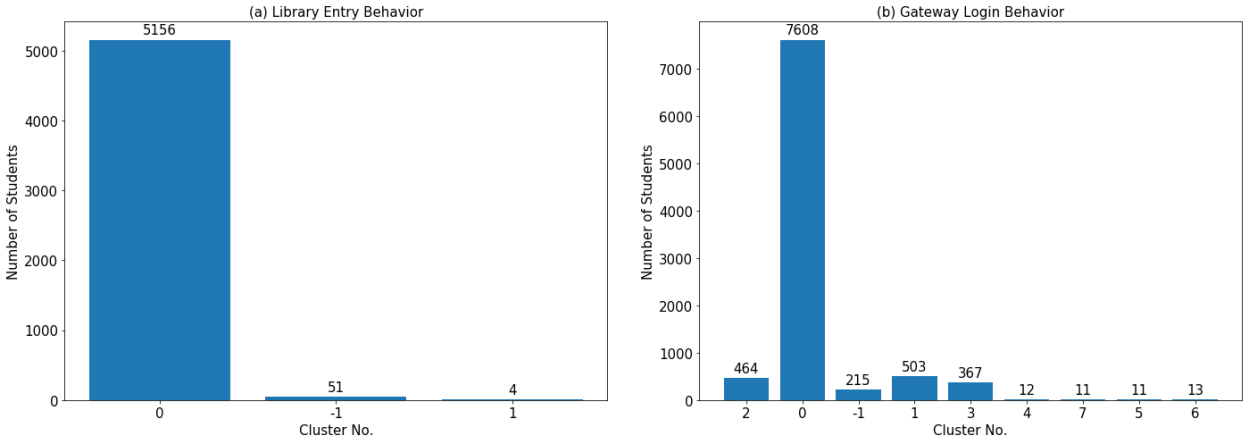


Fig 5. DBSCAN results for (a) library entry and (b) gateway login.

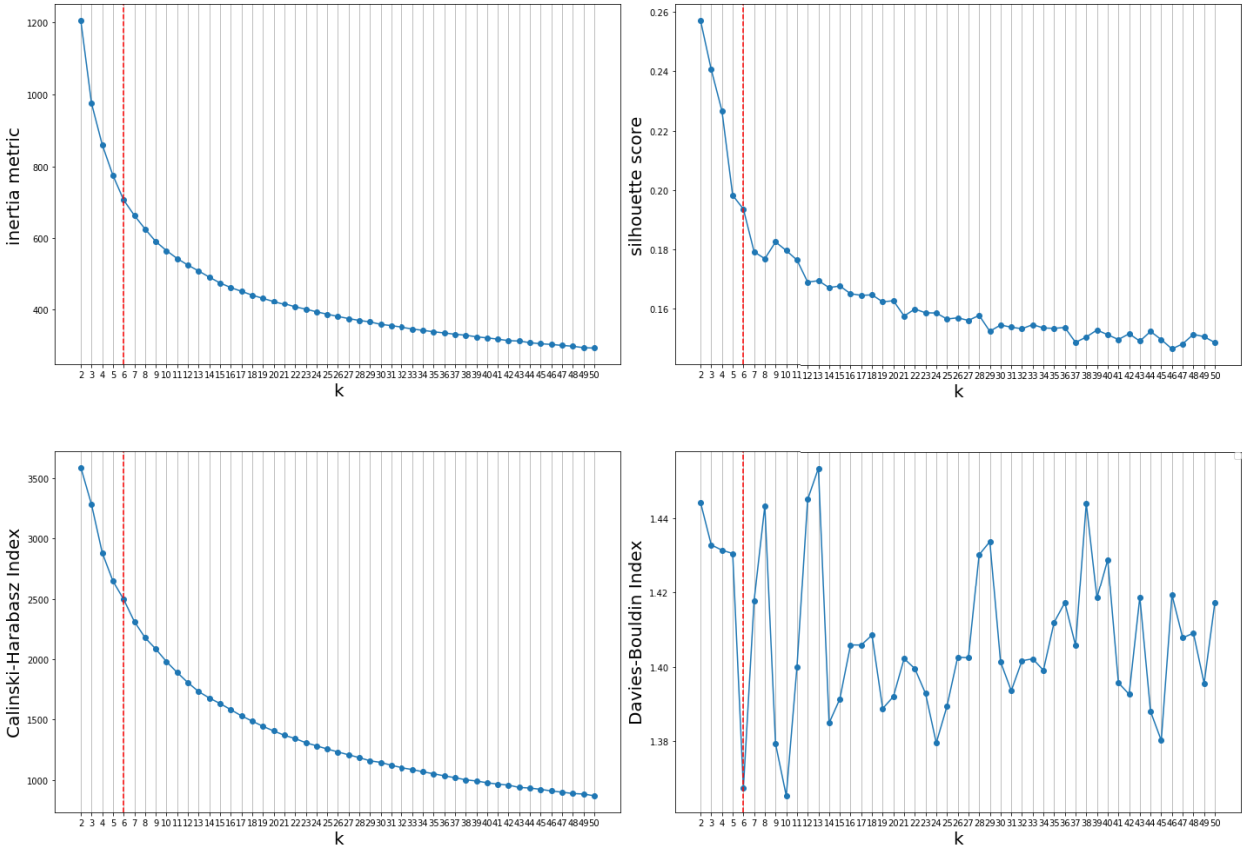


Fig 6. Line diagrams of the four metrics used to count the number of dining behavior subclusters.

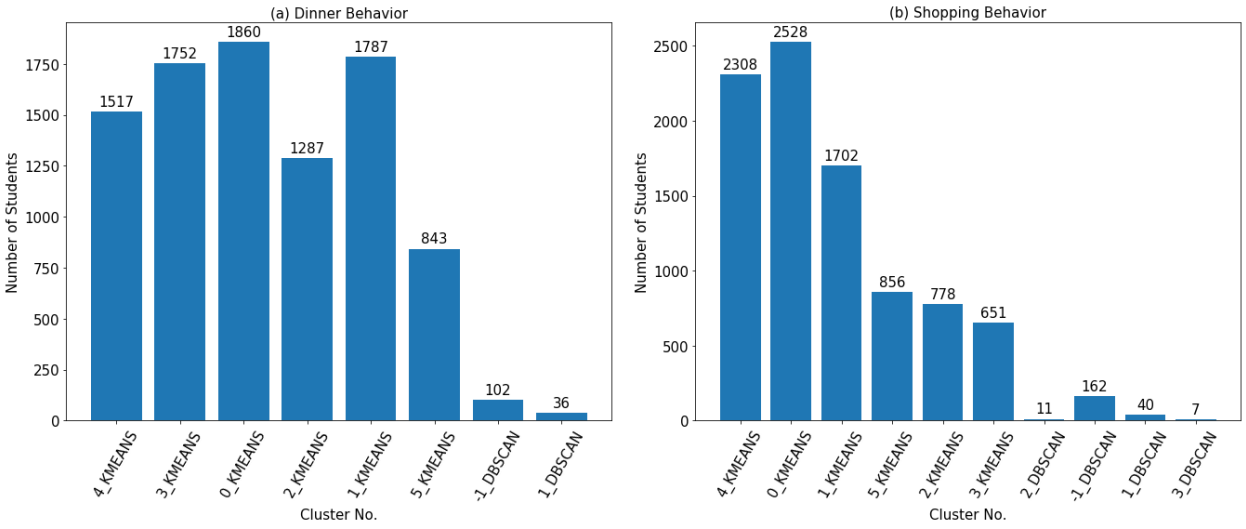


Fig 7. End clustering results of dinner and shopping behavior.

BIRCH is a multiphase hierarchical clustering technique that works well with incremental and dynamic data because the clustering characteristics are stored in a feature tree (CF-tree). Typically, BIRCH comprises two stages. The macroclustering stage is this step. A straightforward grid-based technique called CLIQUE (CLustering In QUEst) locates density-based clusters in subspaces. Clustering is carried out in two processes. CLIQUE divides each dimension into nonoverlapping intervals in the first phase, dividing all data items into cells.

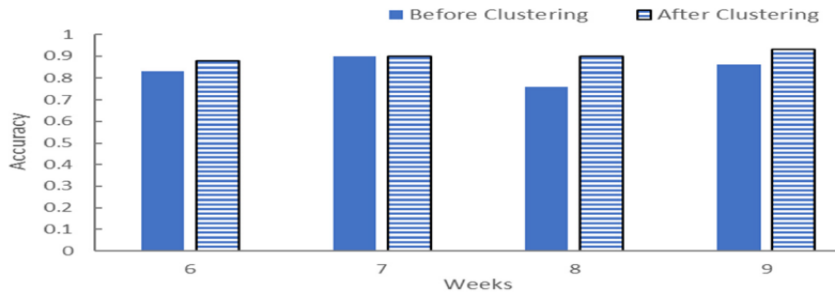


Fig 8. Accuracy rates before and after clustering.

In order to distinguish between dense and sparse cells, CLIQUE applies a density threshold. If the number of items in a cell exceeds the density threshold, the cell is considered dense. The maximum areas are utilised by CLIQUE in the second phase to cover associated dense cells. Cells that are not part of any cluster may be seen as noise, whereas the maximum areas can be seen as clusters. The EM method is a model-based approach that iteratively performs stages E and M until the clustering cannot be made any better. It then calculates the maximum likelihood parameters of the statistical model. Each item is grouped according to its posterior distribution in the E step, and the maximisation of likelihood is used to re-estimate the parameters in the M step.

These pupils exhibit various actions taken by those groups throughout the course of the weeks. It demonstrates that a student's behaviour early in the semester has little bearing on how well the student does at the conclusion of the semester. However, individuals with strong levels of self-efficacy and learning preferences performed well during the last week. The student who performed poorly had a learning style and self-efficacy that ranged from average to low. Week 9 behaviour is crucial for students since it has the most impact on their success.

Student Behavior at	Breakfast	Lunch	Dinner	Shopping	Library	Gateway
Breakfast	0.356	-	-	-	-	-
Lunch	0.065	0.310	-	-	-	-
Dinner	0.064	0.182	0.401	-	-	-
Shopping	0.011	0.101	0.061	0.621	-	-
Library	-0.003	0.095	0.058	-0.015	0.768	
Gateway	0.016	0.013	0.025	0.019	-0.005	0.185

Table 3. Coefficient correlation between patterns with Pearsons method.

Student Behavior at	AMI	ARI	Homogeneity	Completeness	V-measure	FMI
Breakfast	0.005	0.011	0018	0.006	0.009	0.44
Lunch	0.010	0.003	0.005	0.005	0.005	0.315
Dinner	0.008	0.015	0.018	0.009	0.015	0.39
Shopping	0.004	0.010	0.010	0.006	0.008	0.40
Library	0.005	0.008	0.011	0.005	0.004	0.395
Gateway	0.006	0.011	0.015	0.005	0.010	0.412
Multicore	0.016	0.009	0.020	0.014	0.019	0.555

Table 4. Academic result and student behavioral patterns correlation summary.

The experimental findings demonstrate that the ensemble technique put forth in this study is more stable and adaptable due to its constant ability to spot anomalies and its ability to determine the mainstream patterns by manually adjusting the degree of clustering.

Student Behavior at/ Technique	Dinner		Shopping		Library		Gateway	
	Anomalies	Main stream	Anomalies	Main stream	Anomalies	Main stream	Anomalies	Main stream
K-mean	No	Yes	No	Yes	No	Yes	No	Yes
DBSCAN	Yes	No	Yes	No	Yes	No	Yes	No
BIRCH	No	No	No	Yes	No	Yes	Yes	Yes
CLIQUE	Yes	No	Yes	No	Yes	No	Yes	No
EM	Yes	Yes	No	Yes	No	Yes	No	No
Proposed Works	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 5. Summary of detection of anomalous and main stream patterns.

Institutions of higher learning must recognise trends in their students' behaviour and identify at-risk pupils early. By determining what traits at-risk students exhibit and when it is most important to detect at-risk students, the university and lecturers may then develop preventative measures. Students may begin the semester with a positive mindset but conclude it with poor performance, or the opposite may be true. Consequently, it is beneficial to have a better knowledge of student behaviour. Using K-means clustering, this study was able to pinpoint student trends over the course of a semester's worth of weeks. After cleaning, the dataset utilised had 140 students overall, with 41 variables including demographic, self-efficacy, learning style, and programme result. The learning programme output became the primary attribute contributing to the clusters formed by K-means clustering. According to the mean values of the low, average, and high learning programme result, three clusters were formed.

The three groups were contrasted between weeks six and nine. Week 9 had the greatest influence on learning programme outcomes; individuals with high levels of self-efficacy and learning preferences placed a high value on this outcome. The dataset was labelled prior to clustering using the student's semester grades. A new prediction model was created after the student dataset was categorised by cluster using K-means clustering. The prediction model's performance has been compared using the semester grade as a label and the learning programme result as a label. The prediction model that used the clusters label was more accurate. K-means clustering is therefore valuable in recognising a student's pattern throughout the course of a semester and in forecasting a student's achievement. Future study will improve K-means clustering to further improve the results by streamlining the preprocessing process and experimenting with various cluster sizes. Future research in relation to the unimportant outcome across clusters may further expand the dataset size to improve the predictive model's accuracy. Additionally, using a feature selection approach can assist in reducing the quantity of features and determining the crucial aspects that affect a student's performance.

7. Conclusion

This study focused identifying student behavioral patterns to outstanding academic performance by using DBSCAN and K-means algorithms and the results can be improved with the ensemble supervised clustering methods. Our study examined 9024 students in the campus in real time data with the help of statistical methods and entropy mechanism to extract behavioral patterns of the students to suggest an effective method. The results of the experiments show that the suggested strategy is able to more precisely identify normal behavioral patterns as well as detect abnormal behavioral patterns. Clustering can be performed based on the departments. For this purpose, we used DBSCAN and K-means techniques in identifying behavioral patterns. Feature space and high dimensional multisource behavioral patterns particularly can be calculated with k-means algorithm and then out come can be correlated among the patterns obtained. We can extend this work for patterns from various sources with consistent meaning, generating new distance measure for improving feature space for high dimensional data. Correlation can be done on psychological patterns and working communities.

References

[1] Othman, N.T.A.; Misnon, R.; Abdullah, S.R.S.; Kofli, N.T.; Kamarudin, S.K.; Mohamad, A.B. Assessment of programme outcomes through exit survey of Chemical/Biochemical Engineering students. Procedia-Soc. Behav. Sci. 2011, 18, 39–48.

[2] Ka Yuk Chan, C.; Luo, J. Investigating student preparedness for holistic competency assessment: Insights from the Hong Kong context. Assess. Eval. High. Educ. 2022, 47, 636–651.

[3] Mohamed, H.; Judi, H.M.; Jenal, R. Soft Skills Assessment Based On Undergraduate Student Perception. Asia-Pac. J. Inf. Technol.Multimed. 2019, 8, 27–35.

[4] Sabri, M.Z.M.; Majid, N.A.A.; Hanawi, S.A. Model Development in Predicting Academic Performance of Students Based on Self-Efficacy Using K-Means Clustering. J.Phys. Conf. Ser. 2021, 2129, 012047.

[5] Amin, H.M.; Mustafa, N.H. The Undergraduate Learning Style and Achievement in Programming Language Amongst Undergraduates at FTSM. Asia-Pac. J. Inf. Technol. Multimed. 2014, 3, 1–12.

[6] Ghazali, A.S.M.; Noor, S.F.M.; Mohamed, H. E-Hospitality and Tourism Course Based on Students' Learning Styles. Asia-Pacific J.Inf. Technol. Multimed. 2021, 10, 100–117.

[7] Naresh, P., & Suguna, R. (2021). Implementation of dynamic and fast mining algorithms on incremental datasets to discover qualitative rules. Applied Computer Science, 17(3), 82-91. https://doi.org/10.23743/acs-2021-23.

[8] Francis, B.K.; Babu, S.S. Predicting academic performance of students using a hybrid data mining approach. J. Med. Syst. 2019,43, 162.

[9] Khan, I.; Ahmad, A.R.; Jabeur, N.; Mahdi, M.N. A Conceptual Framework to Aid Attribute Selection in Machine Learning Student Performance Prediction Models. Int. J. Interact. Mob. Technol. 2021, 15, 4–19.

[10] Aggarwal, D.; Mittal, S.; Bali, V. Significance of non-academic parameters for predicting student performance using ensemble learning techniques. Int. J. Syst. Dyn. Appl. (IJSDA) 2021, 10, 38–49.

[11] Asif, R.; Merceron, A.; Ali, S.A.; Haider, N.G. Analyzing undergraduate students' performance using educational data mining. Comput. Educ. 2017, 113, 177–194.

[12] Abdelhafez, H.A.; Elmannai, H. Developing and Comparing Data Mining Algorithms ThatWork Best for Predicting Student Performance. Int. J. Inf. Commun. Technol. Educ. (IJICTE) 2022, 18, 1–14.

[13] M. I. Thariq Hussan, D. Saidulu, P. T. Anitha, A. Manikandan and P. Naresh (2022), Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People. IJEER 10(2), 80-86. DOI: 10.37391/IJEER.100205.

[14] Nagesh, C., Chaganti, K.R. , Chaganti, S. , Khaleelullah, S., Naresh, P. and Hussan, M. 2023. Leveraging Machine Learning based Ensemble Time Series Prediction Model for Rainfall Using SVM, KNN and Advanced ARIMA+ E-GARCH. International Journal on Recent and Innovation Trends in Computing and Communication. 11, 7s (Jul. 2023), 353–358. DOI:https://doi.org/10.17762/ijritcc.v11i7s.7010..

[15] G. Kostopoulos, S. Kotsiantis, N. Fazakis, G. Koutsonikos, and C. Pierrakeas, ``A semi-supervised regression algorithm for grade prediction of students in distance learning courses," Int. J. Artif. Intell. Tools, vol. 28, no. 4, Jun. 2019, Art. no. 194000.1

[16] D. Hooshyar, M. Pedaste, and Y. Yang, ``Mining educational data to predict Students' performance through procrastination behavior," Entropy, vol. 22, no. 1, p. 12, Dec. 2019.

[17] I. HarwatiR Virdyanawaty and A. Mansur, ``Drop out estimation students based on the study period: Comparison between naive Bayes and support vector machines algorithm methods," in Proc. ICET4SD, Yogyakarta, IN, USA, 2015.

[18] V. Krishna, Y. D. Solomon Raju, C. V. Raghavendran, P. Naresh and A. Rajesh, "Identification of Nutritional Deficiencies in Crops Using Machine Learning and Image Processing Techniques," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 925-929, doi: 10.1109/ICIEM54221.2022.9853072.

[19] T. Aruna, P. Naresh, A. Rajeshwari, M. I. T. Hussan and K. G. Guptha, "Visualization and Prediction of Rainfall Using Deep Learning and Machine Learning Techniques," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 910-914, doi: 10.1109/ICTACS56270.2022.9988553.

[20] R. Kosara, F. Bendix, and H. Hauser, ``Parallel sets: Interactive exploration and visual analysis of categorical data," IEEE Trans. Vis. Comput. Graph-ics, vol. 12, no. 4, pp. 558_568, Jul. 2006.

[21] Y. Cao, J. Gao, D. Lian, Z. Rong, J. Shi, Q. Wang, Y. Wu, H. Yao, and T. Zhou, ``Orderliness predicts academic performance: Behavioural analysis on campus lifestyle," J. Roy. Soc. Interface, vol. 15, no. 146, Sep. 2018, Art. no. 20180210.

[22] B. Narsimha, Ch V Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi and P. Naresh (2022), Cyber Defense in the Age of Artificial Intelligence and Machine Learning for Financial Fraud Detection Application. IJEER 10(2), 87-92. DOI: 10.37391/IJEER.100206..

[23] R.D. Cox, It was just that I was afraid promoting success by addressing students' fear of failure, Community Coll. Rev. 37 (1) (2009) 52–80.

[24] Crowe, C. Dirks, M.P. Wenderoth, Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology, CBE-Life Sci. Educ. 7 (4) (2008) 368–381.

[25] M. Devlin, Taking responsibility for learning isn't everything: a case for developing tertiary students'

conceptions of learning, Teach. Higher Educ. 7 (2) (2002) 125–138.

[26] V.Krishna, Dr.V.P.C.Rao, P.Naresh, P.Rajyalakshmi " Incorporation of DCT and MSVQ to Enhance Image Compression Ratio of an image" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 03 | Mar-2016.

[27] E. Fitkov-Norris, A. Yeghiazarian, Measuring study habits in higher education: the way forward? Paper Presented at the Journal of Physics: Conference Series (2013).

[28] Naresh, P., & Suguna, R. (2021). IPOC: An efficient approach for dynamic association rule generation using incremental data with updating supports. Indonesian Journal of Electrical Engineering and Computer Science, 24(2), 1084. https://doi.org/10.11591/ijeecs.v24.i2.pp1084-1090.

[29] B. Fogg, Creating persuasive technologies: an eight-step design process, Paper presented at the Proceedings of the 4th International Conference on Persuasive Technology (2009).

[30] P. Naresh, S. V. N. Pavan, A. R. Mohammed, N. Chanti and M. Tharun, "Comparative Study of Machine Learning Algorithms for Fake Review Detection with Emphasis on SVM," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 170-176, doi: 10.1109/ICSCSS57650.2023.10169190.

[31] J.C. Hilpert, J. Stempien, K.J. van der Hoeven Kraft, J. Husman, Evidence for the latent factor structure of the MSLQ a new conceptualization of an established questionnaire, SAGE Open 3 (4) (2013) (2158244013510305).

[32] Hussan, M.I. & Reddy, G. & Anitha, P. & Kanagaraj, A. & Pannangi, Naresh. (2023). DDoS attack detection in IoT environment using optimized Elman recurrent neural networks based on chaotic bacterial colony optimization. Cluster Computing. 1-22. 10.1007/s10586-023-04187-4.

[33] C.R. Hynd, Teaching students to think critically using multiple texts in history, J. Adolesc. Adult Lit. (1999) 428–436.