ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

Student Classification using Student Diversification Algorithm

[1] Raj Tandel, [2] Salman Adhikari, [3] Sahil Shaikh, [4] Aman Pinjara, [5] Bushra Shaikh

[1] [2] [3] [4] Student, SIES Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra 400706
[5] Assistant Professor, SIES Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra 400706
Corresponding Author Email: [1] rajtit120@gst.sies.edu.in, [2] adhikarijit120@gst.sies.edu.in,
[3] mohammadsahilsit120@gst.sies.edu.in, [4] amanpit120@gst.sies.edu.in, [5] bushras@sies.edu.in

Abstract*— A comprehensive understanding of student achievement in education extends beyond traditional measures and encompasses many dimensions. This research uses advanced clustering and classification techniques to reveal micropatterns in student data, focusing on attendance, course markers, and achievement Primary goals include identifying specific groups of students, develop predictive segmentation models, and provide actionable insights for instructional interventions tailored to individual Specific needs groups of diverse learners Provides a means for tailoring strategies to address them. The importance of this research lies in its potential to reshape educational practices, creating an environment where every student can thrive. Through an extensive analysis of methods, data, and findings, this paper makes a valuable contribution to educational data analysis and improves data-driven decision-making to improve student outcomes.

***Index Terms*— Algorithm, Analysis, Clustering Classification, Diversification, Mentoring, Machine Learning, Students.**

I. INTRODUCTION

In an ever-changing educational landscape, trying to understand and improve student performance is a perennial challenge. As educational institutions collect unprecedented amounts of data, the need for sophisticated methods to gain meaningful insights becomes increasingly apparent This research seeks to develop advanced clustering and classification will unravel the complex fabric of students' learning lives.

By focusing on key indicators such as attendance, grades, and achievement, we aim to divide students into homogeneous groups, giving teachers understanding of learning processes of various types. Education is a dynamic industry that is constantly exploring new ways to enhance student achievement and learning experiences. In recent years, the combination of advanced technology, and data-driven approaches has transformed educational transformation. Learner classification, a key component of educational content analysis, plays a vital role in understanding and addressing the diverse needs and aspirations of learners.

Traditional approaches to classifying students tend to focus on performance measures and historical data, providing valuable insights but potentially overlooking the rich diversity in student populations.

The student diversity program is designed to go beyond traditional classification methods by considering different aspects of student diversity. This algorithm takes advantage of a broad range of criteria including academic achievement, learning style, socioeconomic factors, extracurricular activities. This paper examines the conceptual design, implementation, and potential impact of student diversity categorization policies on student categorization. By exploring the challenges of this process, we aim to contribute to the ongoing discourse on educational data analysis, providing teachers and administrators with a powerful tool to improve and control decision making respond effectively to the diverse needs of students

II. NEED AND OBJECTIVE

Accessibility to coaching: Coaches are invaluable, but many people have difficulty finding a good coach due to limited opportunities and conflicting coaches. The app meets the foundation's need to provide more education to a wider audience.

Effective Communication: Effective communication is essential for effective education. The initiative addresses the need for a platform that facilitates communication between instructors and supervisors, facilitating collaboration and knowledge sharing.

Relationship Management: Managing teacher relationships over time is a complex task that requires coordination and follow-up. The practice is crucial to helping instructors and supervisors have a good rapport and training relationship and to keep them focused on their goals.

Professional Development Goals: Many educational relationships lack clear goals and methods for monitoring progress. The app needs to provide a framework for setting goals and tracking changes in social education, allowing participants to measure their own progress and make adjustments as necessary. Growth and Personal: Coaches are important drivers of personal growth, both professional and personal. The program meets the need for a platform that provides access to mentors to help people realize their potential in work, education or personal development.

III. LITERATURE SURVEY

Hierarchical clustering is a widely-used technique for grouping data points based on their similarities, and its application to categorical data presents unique challenges. [1] Alalyan et al. (2019) introduced a model-based approach to address these challenges, emphasizing the need for tailored methods in the clustering of categorical data. While hierarchical clustering has been extensively applied to numerical data, the literature on its application to categorical data is relatively limited. The work by Alalyan et al. contributes to filling this gap by proposing a model-based hierarchical clustering method, offering insights into the advancements in clustering techniques for categorical data.

Predictive modelling in the context of academic performance has gained significant attention in recent years, with researchers exploring various techniques to enhance accuracy and effectiveness. Ensemble learning, a technique that combines multiple models to improve predictive performance, has emerged as a powerful approach in this domain. [2] Akinpelu et al. (2020) conducted a study focusing on predictive modelling for students' academic performance, specifically in an Information Technology course. While the use of machine learning for academic prediction is not novel, the application of ensemble learning in this specific context adds a valuable dimension to the existing literature. Ensemble learning techniques, such as bagging and boosting, have shown promise in enhancing the robustness and accuracy of predictive models. This aligns with the broader trend in educational data mining, where researchers are increasingly turning to advanced machine learning methodologies to gain deeper insights into student performance.

Clustering techniques play a vital role in data engineering and machine learning applications, facilitating the identification of inherent structures within datasets. [3] Altinigneli et al. (2020) proposed a novel approach, Hierarchical Quick Shift Guided Recurrent Clustering, which introduces a distinctive combination of hierarchical and recurrent clustering methodologies. Traditional clustering algorithms, such as K-Means and hierarchical clustering, have been extensively utilized in various domains. K-Means efficiently partitions data into distinct clusters, while hierarchical clustering organizes data in a tree-like structure. However, the evolving landscape of data engineering demands innovative clustering techniques capable of handling complex and dynamic datasets. [3] Altinigneli et al.'s (2020) work introduces a new paradigm by integrating the Quick Shift algorithm, which is known for its efficiency in mode-seeking, with recurrent clustering principles. Recurrent clustering, with its ability to capture temporal patterns, adds a dynamic dimension to the clustering process. This amalgamation of hierarchical, quick shift, and recurrent clustering methodologies reflects a nuanced understanding of data structures in time-varying environments.

[10] Veluri et al. (2022) contribute to this domain with their paper on "Learning analytics using deep learning techniques for efficiently managing educational institutes." The study addresses the growing need for advanced analytical tools in educational settings, where the volume and complexity of data necessitate sophisticated approaches for meaningful analysis. Traditional learning analytics often relies on statistical and machine learning methods, but the advent of deep learning introduces a new paradigm. Deep learning algorithms, particularly neural networks, excel at capturing intricate patterns within large datasets, making them well-suited for the complexities inherent in educational data.

The field of predictive modelling for e-learning performance has witnessed significant growth, with researchers exploring various methodologies to enhance accuracy and provide valuable insights. Ensemble machine learning, which combines multiple models to improve overall prediction performance, has emerged as a promising approach in this domain. [11] Saleem et al. (2021) contribute to this field with their paper on an "Intelligent Decision Support System for Predicting Student's E-Learning Performance Using Ensemble Machine Learning." The application of ensemble techniques in predicting student performance in an e-learning environment represents an extension of traditional predictive modelling approaches. The study aligns with the broader trend of leveraging advanced machine learning methods to gain deeper insights into student behaviour and performance.

IV. PROPOSED SYSTEM

The proposed student segmentation and diversity represents an innovative solution for educational institutions seeking to enhance student support and create a personalized learning environment. This well-designed program is recommended to enhance students' broad skills, tasks and exam results in integrating the various components and benefit from machine learning and cluster algorithms, and further measures values of the specific groups. and to plant, the corrupt, educational institutions. The user-friendly interface, with a parameter weighting mechanism customizable to specific preferences, enables teachers to gain real-time insights, enabling informed decision-making and interventions if targeted is easy. Expected outcomes include improved student support, early identification of students with additional needs, and the attainment of an inclusive educational environment that celebrates diversity and individual strengths Overall, this proposed program is in line with the evolving needs of education and provides a practical and inclusive educational system.

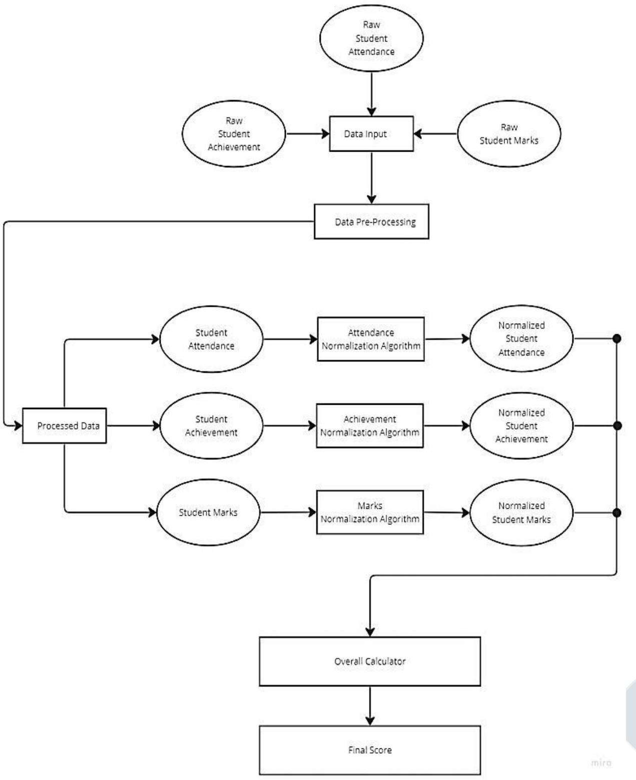


Fig. 4.1 Working of Proposed System

Implementation Details

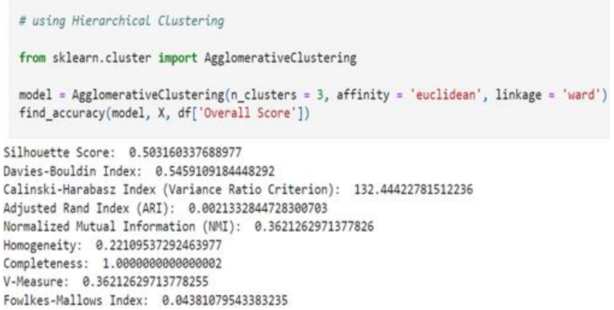
The aim of this algorithm was to create diversity algorithm for student. For students with the help of existing marks, achievement, And attendance. We have provided excel files for marks. We applied algorithm of classification and Clustering

A. Evaluation

In traditional teaching systems, student evaluations are usually limited to mathematical grades and occasional accolades. However, this traditional approach ignores the many dimensions of student engagement and achievement. Utilizing the power of data analysis, this study seeks to overcome these limitations, accessing multiple databases including attendance records, academic performance records, and progress a acquired post-class by mirroring clustering and classification processes Opens the way for more effective individual academic interventions

Hierarchical clustering

Using hierarchical clustering in our approach works by implementing overall scores derived from the careful analysis of students data involving marks attendance and achievements and by classifying them based on clusters derived from the similarity and the closeness of the scores of each individual student.



```
# using Hierarchical Clustering

from sklearn.cluster import AgglomerativeClustering

model = AgglomerativeClustering(n_clusters = 3, affinity = 'euclidean', linkage = 'ward')
find_accuracy(model, X, df['Overall Score'])
```

```
Silhouette Score:  0.503160337688977
Davies-Bouldin Index:  0.545910918448292
Calinski-Harabasz Index (Variance Ratio Criterion):  132.4442278151223б
Adjusted Rand Index (ARI):  0.0021332844728300703
Normalized Mutual Information (NMI):  0.3621262971377826
Homogeneity:  0.22109537292463977
Completeness:  1.0000000000000002
V-Measure:  0.36212629713778255
Fowlkes-Mallows Index:  0.04381079543383235
```

Fig. 4.2 Hierarchical clustering

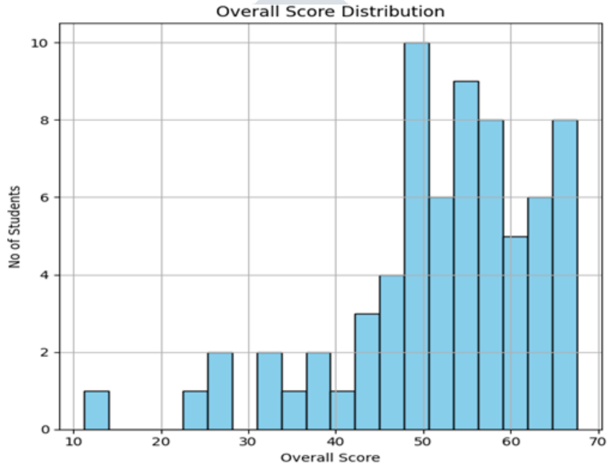
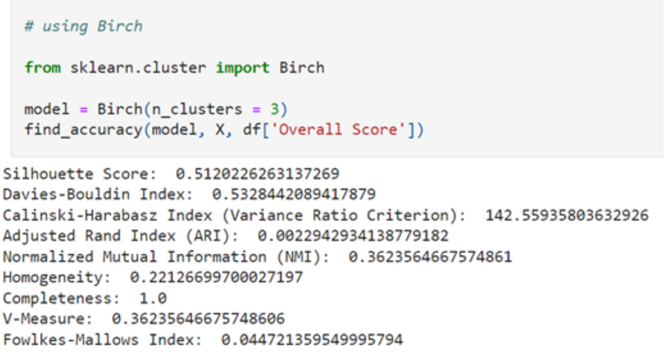


Fig. 4.3 Hierarchical clustering Graph

Birch

BIRCH, which stands for [14] Balanced Iterative Reducing and Clustering using Hierarchies, is a data clustering algorithm. It's designed for large datasets and focuses on memory efficiency and scalability. BIRCH builds a hierarchical structure of clusters using a tree-like data structure known as the Cluster Feature Tree (CF Tree). Our implementation of BIRCH is carried out by having the larger dataset of students overall scores, with each score reflecting the students performance based on their individual data which is then being summarized into a smaller datasets to be further clustered instead of the entire dataset. Here 2 indicates Weak, 0 indicates Medium and 1 indicates Bright



```
# using Birch

from sklearn.cluster import Birch

model = Birch(n_clusters = 3)
find_accuracy(model, X, df['Overall Score'])
```

```
Silhouette Score:  0.5120226263137269
Davies-Bouldin Index:  0.5328442089417879
Calinski-Harabasz Index (Variance Ratio Criterion):  142.55935803632926
Adjusted Rand Index (ARI):  0.0022942934138779182
Normalized Mutual Information (NMI):  0.3623564667574861
Homogeneity:  0.22126699700027197
Completeness:  1.0
V-Measure:  0.36235646675748606
Fowlkes-Mallows Index:  0.044721359549995794
```

Fig. 4.4 Birch

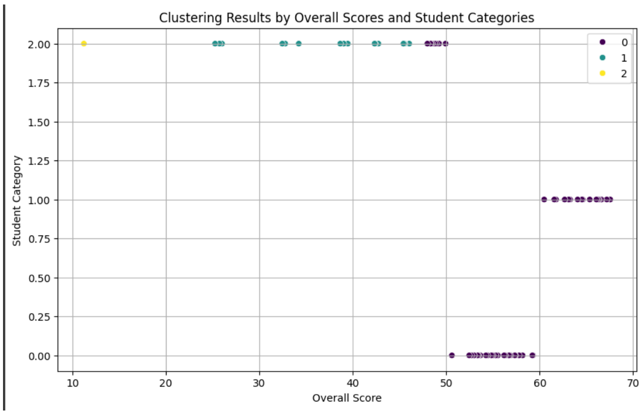
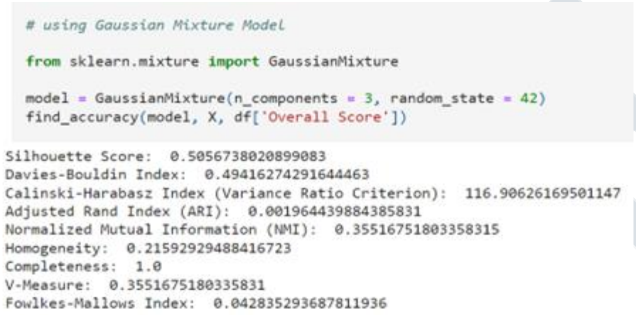


Fig. 4.5 Birch Scatterplot

Gaussian Mixture Model

[15] A Gaussian Mixture Model (GMM) is a probabilistic model used for clustering and density estimation in data analysis and machine learning. It assumes that data points are generated from a mixture of multiple Gaussian distributions. Each component of the mixture represents a cluster in the data. GMMs aim to estimate the parameters of these Gaussian distributions, including means and covariances, to model the underlying data distribution accurately.



```
# using Gaussian Mixture Model

from sklearn.mixture import GaussianMixture

model = GaussianMixture(n_components = 3, random_state = 42)
find_accuracy(model, X, df['Overall Score'])
```

```
Silhouette Score:  0.5056738020899083
Davies-Bouldin Index:  0.49416274291644463
Calinski-Harabasz Index (Variance Ratio Criterion):  116.90626169501147
Adjusted Rand Index (ARI):  0.001964439884385831
Normalized Mutual Information (NMI):  0.35516751803358315
Homogeneity:  0.21592929488416723
Completeness:  1.0
V-Measure:  0.3551675180335831
Fowlkes-Mallows Index:  0.042835293687811936
```

Fig. 4.6 Gaussian

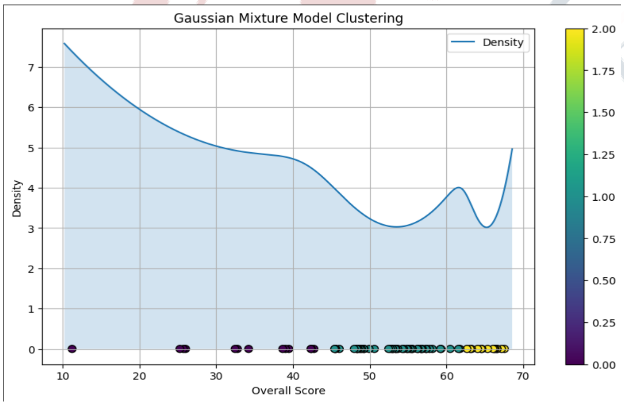


Fig. 4.7 Gaussian Mixture Model Density Contours

Algorithm Code

Attendance

```
total_classes = 255
sem_6['SEM6-Normalized'] = ((sem_6['Total'] / total_classes) - 0.5) / 0.5
sem_6['SEM6-Normalized'] = sem_6['SEM6-Normalized'] * 100
sem_6.head()
```

Marks

```
def calculate_2(df,col_1,col_2):
    roll_no_and_name = df.loc[:, ['Roll no', 'Name of student']]
    def calculate_normalized_marks(df,col_name):
  df['total'] = df.sum(axis=1)
        df[col_name] = ((df['total']/5 - 8) / (20 - 8)) * 100
        return df[col_name]
    ia_1 = calculate_normalized_marks(df.loc[:, df.columns.str.contains('IA1')],col_1)
    ia_2 = calculate_normalized_marks(df.loc[:, df.columns.str.contains('IA2')],col_2)
    ia_1_df = pd.concat([roll_no_and_name, ia_1,ia_2], axis=1)
    return ia_1_df
```

Overall Classification

```
t = kmeans.cluster_centers_.tolist()
sorted_t = sorted(t)

t, sorted_t
ans = []
for i in range(len(t)):
    pos = sorted_t.index(t[i])
    if pos == 0:
        ans.append('Weak')
    elif pos == 1:
        ans.append('Average')
    else:
        ans.append('Bright')
ans
required_cluster_names = list(map(lambda x: ans[x], kmeans.labels_))
# required_cluster_names

df['required_cluster_names'] = required_cluster_names
```

Equations

Normalized Marks Formula:

$$NormalizedMarks = \frac{(AverageMarks - MinimumMarks)}{(MaximumMarks - MinimumMarks)} * 100$$

Normalized Attendance Formula:

$$NormalizedAttendance = \frac{(AverageAttendance - MinimumAttendance)}{(MaximumAttendance - MinimumAttendance)} * 100$$

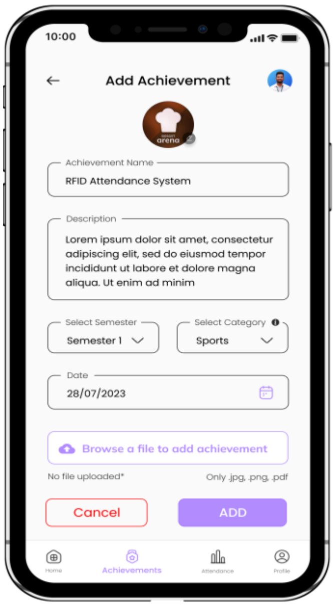
Normalized Achievement Formula:

$$NormalizedAchievements = \frac{(TotalNumberofAchievements - MinimumAchievements)}{(MaximumAchievements - MinimumAchievements)} * 100$$

Overall Score

$$Overall = \frac{(NormalizedMarks + NormalizedAttendance + NormalizedAchievement)}{3}$$

B. User Interface



Add Achievement

Achievement Name: RFID Attendance System

Description: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim

Select Semester: Semester 1

Select Category: Sports

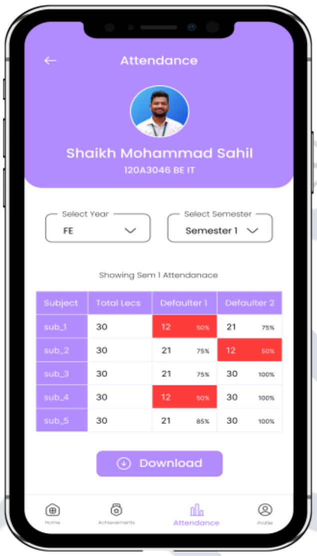
Date: 28/07/2023

No file uploaded*

Only .jpg, .png, .pdf

Fig. 4.5 Add Achievements

Students can update the Achievements which they got during their curriculum and add all the achievements which respected dates and also can add certificates as a proof and upload it.The mentees (students) can see their attendance for the respected Semesters so that they can know their status if they are in defaulter for a particular subject or not and also can download it Similarly, they can see their marksheets and can verify their marks So these are the parameters which we are taking into consideration for our algorithm



Attendance

Shaikh Mohammad Sahil

120A3046 BE IT

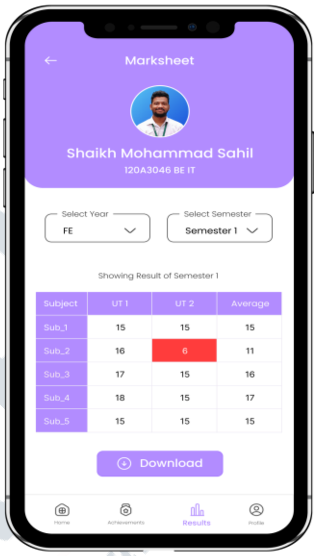
Select Year: FE

Select Semester: Semester 1

Showing Sem 1 Attendanace

Subject	Total Lecs	Defaulter 1	Defaulter 2
sub_1	30	12 50%	21 75%
sub_2	30	21 75%	12 50%
sub_3	30	21 75%	30 100%
sub_4	30	12 50%	30 100%
sub_5	30	21 85%	30 100%

Fig. 4.6 Attendance



Marksheet

Shaikh Mohammad Sahil

120A3046 BE IT

Select Year: FE

Select Semester: Semester 1

Showing Result of Semester 1

Subject	UT 1	UT 2	Average
Sub_1	15	15	15
Sub_2	16	6	11
Sub_3	17	15	16
Sub_4	18	15	17
Sub_5	15	15	15

Fig. 4.7 Marksheet

V. RESULTS

Table 1.1 Comparison of Algorithm

	K-means	**Hierarchical Clustering**	**Gaussian Mixture Model**	**Birch**	**Spectral Clustering**
Silhouette Score	0.528	0.503	0.505	0.512	0.491
Davies-Bouldin Index	0.553	0.546	0.494	0.532	0.563
Calinski-Harabasz Index	152.507	132.444	116.906	142.559	77.270
Adjusted Rand Index	0.002	0.002	0.002	0.002	0.003
Normalized Mutual Information	0.367	0.362	0.355	0.362	0.409
Homogeneity	0.224	0.221	0.216	0.221	0.257
Completeness	1.000	1.000	1.00	1.00	1.00
V-Measure	0.367	0.362	0.355	0.362	0.409
Fowlkes-Mallows Index	0.045	0.043	0.043	0.044	0.050

The table provides the evaluation metrics for diverse clustering algorithms, including K-means, Hierarchical Clustering, Gaussian Mixture Model (GMM), Birch, and Spectral Clustering. The metrics encompass the Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, Adjusted Rand Index, Normalized Mutual Information, Homogeneity, Completeness, V-Measure, and Fowlkes-Mallows Index. These metrics are usually used to evaluate the best and effectiveness of clustering algorithms. The Silhouette Score measures how well-defined the clusters are, with better values indicating higher-described clusters. The Davies-Bouldin Index evaluates cluster compactness and separation.

The Calinski-Harabasz Index assesses the ratio of between-cluster to inside-cluster variance. The Adjusted Rand Index, Normalized Mutual Information, Homogeneity, Completeness, and V-Measure measure the accuracy and purity of clustering results. The Fowlkes-Mallows Index evaluates the similarity between real and anticipated clusters. Interpretation of these metrics enables in selecting the maximum appropriate clustering algorithm for a given dataset, considering elements like cluster brotherly love, separation, and universal clustering overall performance.

VI. CONCLUSION

Overall, this study is an important step in solving today's educational problems in the academic and professional world. Improving mentor-mentee connection and management practice promises to transform the way mentors and mentees connect, communicate, and manage mentoring relationships. The system will eliminate the shortcomings of the current system, saving time, reducing inefficiencies and promoting more effective training. Additionally, the integration of various student algorithms provides personalized education that will meet the needs and desires of each student. We aim to improve the education and career of countless people by implementing this solution. This program meets the needs of our society, education is the driving force of personal and professional development. In doing this, we not only support personal development, but also the prosperity of our communities and businesses. This project effectively uses machine learning algorithms to segment students based on attendance and academic achievement. The inclusion of different grouping and classification strategies provides teachers with a nuanced perspective, helping to identify at-risk students and identify shared characteristics across groups. User-friendly interfaces ensure effective implementation, empowering teachers to make informed decisions and implement targeted interventions. With scalability and adaptability at its core, the project stands as a valuable tool to improve student management and learning support systems in a variety of educational settings

VII. SUMMARY

The number one goal of this project revolves across the category of students, a method driven by way of their attendance records and academic overall performance, leveraging a numerous set of machine mastering algorithms. The records series section involved accumulating complete facts encompassing attendance information and earlier coursework. In the clustering area, we employed more than a few algorithms together with Birch, Gaussian Mixture Model, and Hierarchical Clustering, aimed at grouping students based totally on shared characteristics.

The next tiers of the venture centered at the schooling and trying out of those machine gaining knowledge of fashions. This worried the utilization of diverse algorithms, consisting of the ones designed for type such as Decision Trees, Support Vector Machines, Random Forest, and Logistic Regression. Additionally, clustering algorithms like K-Means were applied to become aware of patterns and group students based totally on similarities.

To ensure sensible application, the task incorporated these fashions right into a consumer-pleasant format. This consumer interface serves as a platform to present actionable insights to educators. The system aids instructors in making knowledgeable choices via providing a clean know-how of student overall performance styles and capability areas of development.

One of the key strengths of our challenge lies in its adaptability. The machine is designed with flexibility in thoughts, permitting it to seamlessly combine into one of a kind educational settings. This adaptability ensures that the insights and hints generated by using the system can be efficiently implemented throughout various contexts, contributing to progressed student management and getting to know help.

REFERENCES

[1] Alalyan F, Zamzami N, Bouguila N (2019) Model-based hierarchical clustering for categorical data. In: 28th IEEE international symposium on industrial electronics, ISIE 2019, Vancouver, BC, June 12–14, 2019. IEEE, pp 1424–1429. https://doi.org/10.1109/ISIE.2019.8781307

[2] H. O. Akinpelu, O. A. Akanbi, O. A. Akande, Predictive Modeling for Students' Academic Performance Using Ensemble Learning: A Case Study of an Information Technology Course, Computers & Education 151 (2020) 103854.

[3] Altinigneli MC, Miklautz L, Böhm C et al (2020) Hierarchical quick shift guided recurrent clustering. In: 2020 IEEE 36th international conference on data engineering (ICDE). pp 1842–1845. https://doi.org/10.1109/ICDE48307.2020.00184

[4] Prasad, Abeer Alsadoon and Angelika Maag, "A systematic review: machine learning based recommendation systems", e-learning Shristi Shakya Khanal1 & P.W.C., 2019, [online]

[5] Barton T, Bruna T, Kordík P (2019) Chameleon 2: an improved graph-based clustering algorithm. ACM Trans Knowl DiscovData 13(1):10. https://doi.org/10.1145/3299876

[6] Arshi Naim and Fahad Alahmari, "Reference model of e-learning and quality to establish interoperability in higher education systems (2020/1/29)", International Journal of Emerging Technologies in Learning (iJET), vol. 15, no. 02, pp. 15-28, [online] Available: https://onlinejour.journals.public knowledgeproject.org/index.php/i-jet/article/view/11605

[7] A. Naim, S. M. Alshawaf, P. Kumar Malik and R. Singh, "Effective E-Learning Practices by Machine Learning and Artificial Intelligence," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 491-495,

[8] U. Bin Qushem, A. Christopoulos, S. S. Oyelere, H. Ogata and M. J. Laakso, "Multimodal technologies in precision education: Providing new opportunities or adding more challenges?"https://www.mdpi.com/2227-7102/11/7/338

[9] M. Injadat Moubayed, A. B. Nassif, H. Lutfiyya and A. Shami, "E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics", IEEE Access, vol. 6, pp. 39117-39138, 2018

[10] R. K. Veluri, I. Patra, M. Naved, V. V. Prasad, M. M. Arcinas, S. M. Beram, et al., "Learning analytics using deep learning techniques for efficiently managing educational institutes", Materials Today: Proceedings, vol. 51, pp. 2317-2320, 2022.

[11] F. Saleem, Z. Ullah, B. Fakieh and F. Kateb, "Intelligent Decision Support System for Predicting Student's E-Learning Performance Using Ensemble Machine Learning", Mathematics, vol. 9, no. 17, pp. 2078, 2021.

[12] S. Bharara, S. Sabitha and A. Bansal, "Application of learning analytics using clustering data Mining for Students' disposition analysis", Educ. Inf. Technol., vol. 23, no. 2, pp.

957-984, Mar. 2018

[13] [Online]Available:https://www.sciencedirect.com/topics/computer-science/hierarchical-clustering

[14] [Online]Available:https://www.geeksforgeeks.org/ml-birch-clustering/

[15] [Online]Available:https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/