ISSN: 2278-4632
Vol-13, Issue-11, November 2023

Study and Forecasting of Student's Academic Achievement using Educational Data Mining

M Ramchander, Sriram Nikhitha

Assistant Professor, Department of MCA, Chaitanya Bharathi Institute of Technology (A), Gandipet,

Hyderabad, Telangana State, India

MCA Student, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

ABSTRACT: Intelligent technology development is gaining attraction in the sphere of education. The increasing rise of educational data suggests that standard processing techniques may be limited and distorted. As a result, recreating data mining research technology in the education area has become more important. To prevent erroneous assessment findings and to anticipate students' future performance, this research analyses and predicts students' academic achievement using applicable clustering, discriminating, and convolution neural network theories. To begin, this work suggests that the clustering-number determination be optimized by using a statistic that has never been employed in the K-means approach. The clustering impact of the K-means method is next assessed using discriminant analysis. The convolutional neural network is presented for training and testing with labelled data. The produced model may be used to forecast future performance. Finally, the efficacy of the constructed model is tested using two metrics in two Cross validation procedures in order to verify the prediction findings. The experimental findings show that the statistic not only addresses the objective and quantitative problem of determining the clustering number in the K-means method, but also enhances the predictability of the outcomes.

Keywords *– Academic performance, clustering analysis, convolutional neural networks, discriminant analysis, educational data mining.*

1. INTRODUCTION

Data mining (DM) may find hidden information in massive amounts of unstructured data. Educational data mining (EDM) is a data mining study topic that focuses on the use of data mining, machine learning, and statistical methodologies. The implementation of data mining technologies in the educational environment has been an active study subject in recent decades. It has grown in prominence in recent years as a result of the availability of online datasets and learning systems [1]. EDM is the creation and implementation of data mining

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

algorithms that allow the study of large amounts of data from varied educational backgrounds. Academic achievement is one of the most essential factors for higher education institutions. As a result, anticipating the learning process and measuring student performance are regarded as important responsibilities in the area of EDM [2].

Fig.1: Example figure

EDM is a discipline that is constantly expanding, focusing on the advancement of self-learning and adaptive ways to expose hidden patterns or internal relationships in educational data. In the sphere of education, heterogeneous data is participating and expanding in the big data paradigm. Some particular data mining approaches are required to adaptively extract valuable information from enormous educational data sets [3]. Because data mining technologies enable the utilization of enormous amounts of student data to examine useful patterns of student learning behavior, EDM application research is progressing quickly. Many facets of educational data processing have benefited from the use of data mining technologies, including student retention, dropout prediction, academic data analysis, and student behaviour analysis [4]. EDM has always placed a premium on assessing and forecasting student academic achievement.

2. LITERATURE REVIEW

A systematic review of deep learning approaches to educational data mining:

Currently, educational institutions collect and retain massive amounts of data, such as student enrollment and attendance records, as well as test results. Mining such data generates exciting knowledge that is beneficial to its users. The rapid rise of educational data suggests that distilling huge volumes of data need a more sophisticated collection of algorithms. This problem gave rise to the subject of Educational Data Mining (EDM). Traditional data mining techniques, which may have a particular aim and function, cannot be directly applied to educational challenges. This means that a pretreatment procedure must first be implemented, and only then can specialized data mining approaches be applied to the issues. Clustering is one such EDM preprocessing

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

method. Many EDM research have focused on the application of different data mining methods to educational qualities. As a result, this research presents a thorough literature assessment spanning over three decades (1983-2016) on clustering algorithms and their application and usefulness in the context of EDM. Based on the literature analysis, future insights are presented, and possibilities for additional study are indicated.

Implementing Auto ML in educational data mining for prediction tasks:

Over the past two decades, Educational Data Mining (EDM) has evolved, concerned with the development and use of data mining techniques to ease the analysis of massive volumes of data emanating from a broad range of educational settings. One of the most essential jobs in the EDM sector is predicting students' development and learning outcomes, such as dropout, performance, and course grades. As a result, both educators and data scientists must use proper machine learning techniques to develop reliable prediction models. Given the high-dimensional input space and the complexity of machine learning algorithms, the process of developing correct and robust learning models necessitates significant data science skills and is, in most situations, time-consuming and error-prone. Furthermore, selecting the appropriate approach for a particular issue formulation and establishing the ideal parameter values for a certain model is a challenging undertaking, and the resulting findings are sometimes difficult to grasp and explain. The primary goal of this work is to investigate the possible usage of sophisticated machine learning algorithms in educational contexts from the standpoint of hyperparameter optimization. We especially study the efficacy of automated Machine Learning (autoML) in predicting students' learning outcomes based on their engagement in online learning platforms. Simultaneously, in order to provide visible and interpretable results, we restrict the search space to tree-based and rule-based models. A variety of trials were conducted to this goal, indicating that auto ML tools routinely provide better outcomes. Hopefully, our work can assist nonexpert users in the area of EDM (e.g., educators and instructors) in conducting experiments with proper automated parameter setups, resulting in extremely accurate and intelligible findings.

Integration of data mining clustering approach in the personalized E-learning system:

Educational data mining is a developing field that focuses on improving self-learning and adaptable approaches. It is used to discover hidden patterns or inherent structures in educational data. In the realm of education, heterogeneous data is involved and constantly rising in the big-data paradigm. Some particular data mining approaches are required to extract valuable information from large amounts of educational data in an adaptable manner. This

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

study describes a clustering strategy for categorizing students into groups or clusters based on their learning behavior. Furthermore, the customized e-learning system architecture is shown, which recognizes and reacts to instructional materials based on the learning capacity of the students. The major goal is to identify ideal circumstances in which learners may increase their learning skills. Furthermore, the administration can uncover critical hidden trends in order to implement successful adjustments in the current system. Using educational data mining, the clustering techniques K-Means, K-Medoids, Density-based Spatial Clustering of Applications with Noise, Agglomerative Hierarchical Cluster Tree, and Clustering by Fast Search and Finding of Density Peaks through Heat Diffusion (CFSFDP-HD) are investigated. It has been discovered that replacing current approaches with CFSFDP-HD yields more robust findings. Data mining methods are equally useful in analyzing massive data to improve education systems.

The use of tools of data mining to decision making in engineering education—A systematic mapping study:

In recent years, there has been an increase in theoretical and practical research on educational data mining. Learning analytics is a subject that employs methodologies, methods, and algorithms to enable users to uncover and extract patterns in recorded educational data in order to improve the teaching-learning process. However, many needs connected to the application of new technologies in teaching-learning processes go largely ignored by learning analytics. An examination of the literature reveals the absence of a comprehensive review of the use of learning analytics in the area of engineering education. The study presented in this article gives researchers an overview of the progress achieved so far and suggests areas where more research is needed. To that purpose, a comprehensive mapping study was conducted with the goal of categorizing publications based on the kind of research and contribution. The findings indicate a tendency toward case study research, which is primarily aimed at software and computer science engineers. Furthermore, subjects such as student retention or dropout prediction, analysis of academic student data, student learning evaluation, and student behavior analysis illustrate developments in the use of learning analytics. Although the emphasis of this systematic mapping research was on the use of learning analytics in engineering education, some of the findings may be applicable to other educational settings.

Data mining in educational technology classroom research: Can it make a contribution?:

The study covers and clarifies some of the important concerns about the use of data mining in classroom research in educational technology.

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

Two studies, one in Europe and one in Australia, are shown as examples of the application of data mining methods, notably association rules mining and fuzzy representations. Both of these studies look at how students learn, behave, and experience computer-supported classroom activities. The approach of association rules mining was utilized in the first research to better understand how learners with various cognitive types interacted with a simulation to solve a problem. Association rules mining was discovered to be an effective way for acquiring accurate data regarding the simulation's usage and performance by learners. The research shows how data mining may be utilized to improve educational software assessment procedures in the area of educational technology. In the second investigation, fuzzy representations were used to inductively investigate questionnaire data. The research shows how educational technologists might utilize data mining to guide and evaluate school-based technology integration projects. The study's ramifications are examined in terms of the need to build instructional data mining tools that can show findings, information, explanations, comments, and suggestions to non-expert data mining users in relevant ways. Finally, data privacy concerns are handled.

3. METHODOLOGY

The conventional absolute score has certain drawbacks in terms of accurately portraying the learning context. The reasons for this include that the difficulty of various courses varies, as do the grading standards of different professors in the same course. To assure the quality of talents, colleges and universities should not only assess students based on grades, but also study students' learning impacts, estimate students' academic performance in the future based on the analyzed findings, and then issue academic warnings in time. This effort will not only assist colleges and universities in enhancing educational quality, but will also assist students in improving their overall performance, hence boosting educational resource management.

This paper's study issue is to objectively assess students' academic accomplishment from the standpoint of clustering and forecast future achievement based on present achievement.

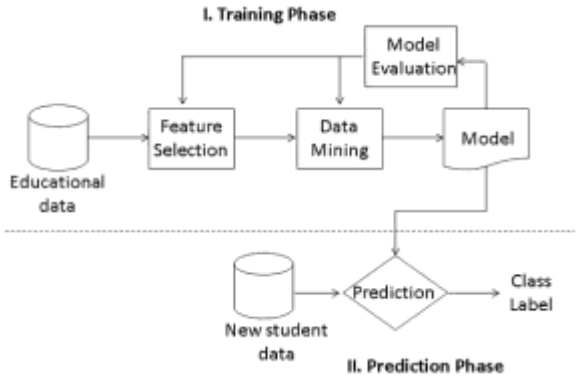

Fig.2: System architecture

4. IMPLEMENTATION

ALGORITHMS:

K-means stands for K-means Algorithm. The K-means method is an iterative technique that attempts to split the dataset into K unique non-

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

overlapping subgroups (clusters), with each data point belonging to just one group. The K-means clustering technique is used to detect groupings in data that have not been explicitly categorized. This may be used to validate business assumptions about the sorts of groups that exist or to find unknown groups in large data sets.

Discriminant analysis: A flexible statistical tool used by market researchers to categorize data into two or more groups or categories is discriminant analysis. To put it another way, discriminant analysis is used to allocate things to one of many recognized categories. Discriminant analysis is a statistical approach that uses scores on one or more quantitative predictor variables to classify data into non-overlapping categories. A clinician, for example, may use discriminant analysis to identify patients who are at high or low risk of having a stroke.

Random forest: Data scientists utilize random forest on the job in a variety of sectors, including banking, stock trading, medical, and e-commerce. It's utilized to forecast factors like consumer behavior, patient history, and safety, which help these businesses function smoothly. The random forest method is a classification system made up of numerous decision trees. When creating each individual tree, it employs bagging and feature randomization in an attempt to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree.

The k-nearest neighbors method, often known as KNN or k-NN, is a non-parametric, supervised learning classifier that employs proximity to create classifications or predictions about an individual data point's grouping. Because it delivers very precise predictions, the KNN algorithm can compete with the most accurate models. As a result, the KNN method may be used for applications that need high accuracy but do not require a human-readable model. The accuracy of the forecasts is determined by the distance measure.

SVM: The "Support Vector Machine" (SVM) is a supervised machine learning technique that may be used for classification or regression tasks. SVM works by mapping data to a high-dimensional feature space in order to classify data points that are otherwise not linearly separable. A separator between the categories is discovered, and the data are processed such that the separator may be drawn as a hyperplane.

Classifier for voting: A voting classifier is a machine learning estimator that trains many base models or estimators and predicts by aggregating their results. Aggregating criteria may be coupled voting decisions for each estimator output.

CNN+LSTM: A CNN is a kind of network architecture for deep learning algorithms that is primarily utilized for image recognition and pixel data processing jobs. There are different forms of neural networks in deep learning, but

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

CNNs are the network design of choice for identifying and recognizing things.

LSTM is an abbreviation for long short-term memory networks, which are utilized in Deep Learning. It is a kind of recurrent neural networks (RNNs) that may learn long-term dependencies, particularly in sequence prediction tasks.

K-fold validation for CNN: K-Fold is a validation strategy in which we divide the data into k-subsets and repeat the holdout procedure k-times, with each of the k subsets serving as the test set and the other k-1 subsets serving as the training set. The average error from all k trials is then determined, which is more trustworthy than the traditional handout technique.

5. CONCLUSION

The following are the consequences of the aforementioned study for the education sector:

1) Leverage the great group to propel overall growth.

2) Targeted modifications to the training program.

to meet the goal of educating pupils based on their ability.

3) Look at more efficient teaching approaches to support student growth.

Given the degree of irrationality and subjectivity in the results of the school's evaluation, the paper begins with data mining by using the K-means algorithm in unsupervised learning to perform clustering analysis on student performance, and then using the clustering results as the category label of CNN. It is eventually discovered that the model has a higher ideal forecast accuracy, which is important to ensuring objective and fair student assessment by school. Furthermore, it is accessible to quickly recall students who are on academic probation. When examining data labels, the label value selection range must be considered, and the label value selection range is connected to the clustering number. The K-means method has a well-known flaw: the value of k is selected arbitrarily. To enhance the method, the study employs an objective statistic to maximize k-value selection and substitutes subjective assessment with quantitative analysis, resulting in more strong clustering findings. The persuasiveness also makes CNN training and prediction outcomes more dependable, and the model's success is automatically assured. Although the clustering findings are acquired after a thorough examination of the current situation and the application of quantitative analysis, the initial clustering center is chosen at random, which may have an influence on the accuracy of the clustering results. Although the suggested statistic improves CNN results over those obtained without it, we do not compare it to other classifiers. In the age of big data, EDM has several potential in terms of policy, resources, and technology. EDM research is

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

important to the advancement and innovation of education as well as society as a whole. Because of the complexity of educational challenges and the multidisciplinary nature of EDM, it stands apart in terms of data sources, data features, research techniques, and application aims. EDM's goal has been to reveal and solve research issues in the education sector by using a number of data mining methods to evaluate educational data and leverage current data to uncover new information, ultimately increasing the quality of education and the learning process. The student dataset is analyzed using a hybrid model that blends data mining approaches with current education data processing technologies.

6. FUTURE SCOPE

It may be improved in the future by merging association models or certain integration-based technologies. Furthermore, EDM may be used to medical data processing, sports data processing, and other sectors. Future study material might include using educational data mining tools to uncover ideas to encourage discipline development, learning analysis in a virtual learning environment, technology-assisted teaching approaches, and monitoring student mental health. The importance of data mining technology in forecasting academic achievement and boosting learning ability motivates us to go further with our study.

REFERENCES

[1] H. B. Antonio, H. F. Boris, T. David, and N. C. Borja, ‘‘A systematic review of deep learning approaches to educational data mining,’’ Complexity, vol. 2019, May 2019, Art. no. 1306039.

[2] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, ‘‘Implementing AutoML in educational data mining for prediction tasks,’’ Appl. Sci., vol. 10, no. 1, pp. 90–117, Jan. 2020. [3] S. Kausar, X. Huahu, I. Hussain, W. Zhu, and M. Zahid, ‘‘Integration of data mining clustering approach in the personalized E-learning system,’’ IEEE Access, vol. 6, pp. 72724–72734, 2018.

[4] D. Buenaæo-Fernandez, W. Villegas, and S. LujÆn-Mora, ‘‘The use of tools of data mining to decision making in engineering education—A systematic mapping study,’’ Comput. Appl. Eng. Educ., vol. 27, no. 3, pp. 744–758, May 2019.

[5] C. Angeli, S. K. Howard, J. Ma, J. Yang, and P. A. Kirschner, ‘‘Data mining in educational technology classroom research: Can it make a contribution?’’ Comput. Educ., vol. 113, pp. 226–242, Oct. 2017.

[6] B. A. Javier, F. B. Claire, and S. Isaac, ‘‘Data mining in foreign language learning,’’ WIREs Data Mining Knowl. Discov., vol. 10, no. 1, Jan./Feb. 2020, Art. no. e1287.

[7] C. Romero and S. Ventura, ‘‘Data mining in education,’’ Wiley Interdiscipl. Rev., Data

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

Mining Knowl. Discovery, vol. 3, no. 1, pp. 12–27, 2013.

[8] C. Romero and S. Ventura, ‘‘Educational data mining and learning analytics: An updated survey,’’ WIREs Data Mining Knowl. Discov., vol. 10, no. 1, May 2020, Art. no. e1355.

[9] S. Wang, ‘‘Smart data mining algorithm for intelligent education,’’ J. Intell. Fuzzy Syst., vol. 37, no. 1, pp. 9–16, Jul. 2019.

[10] M. J. James, S. H. Ganesh, M. L. P. Felciah, and A. K. Shafreenbanu, ‘‘Discovering students’ academic performance based on GPA using K-means clustering algorithm,’’ in Proc. World Congr. Comput. Commun. Technol., Trichirappalli, India, 2014, pp. 200–202.

[11] A. Ani, L. Nicholas, and S. B. Ryan, ‘‘Enhancing the clustering of student performance using the variation in confidence,’’ in Proc. Int. Conf. Intell. Tutoring Syst. Cham, Switzerland: Springer, 2018, pp. 274–279.

[12] R. G. Moises, D. P. P. R. Maria, and O. Francisco, ‘‘Massive LMS log data analysis for the early prediction of course-agnostic student performance,’’ Comput. Educ., vol. 163, Apr. 2020, Art. no. 104108.

[13] J. N. Walsh and A. Rísquez, ‘‘Using cluster analysis to explore the engagement with a flipped classroom of native and non-native Englishspeaking management students,’’ Int. J. Manage. Educ., vol. 18, no. 2, Jul. 2020, Art. no. 100381.

[14] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, ‘‘Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation,’’ Soft Comput., vol. 24, no. 24, pp. 18477–18487, Dec. 2020.

[15] L. M. Crivei, G. Czibula, G. Ciubotariu, and M. Dindelegan, ‘‘Unsupervised learning based mining of academic data sets for students’ performance analysis,’’ in Proc. IEEE 14th Int. Symp. Appl. Comput. Intell. Informat. (SACI), Timisoara, Romania, May 2020, pp. 11–16