ISSN : 2455-9679
SJIF Impact Factor : 6.008

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"A COMPREHENSIVE REVIEW ON DATA MINING AND MACHINE LEARNING MODELS"

Somil Jain [1], Dr Sachin Patel [2]

[1] Research Scholar, Department of Computer Science and Engineering, Institute of Engineering and Technology, SAGE University, Indore, Madhya Pradesh, India

[2] Professor & Head, Department of Computer Science and IT, Institute of Engineering and Technology, SAGE University, Indore, Madhya Pradesh, India

ABSTRACT

The application of data mining (DM) and machine learning (ML) has revolutionized the field of education by providing tools for predicting and enhancing student performance. This study explores the use of predictive analytics, a fusion of ML, historical data, and artificial intelligence (AI), to uncover patterns and trends in educational datasets. With a focus on Educational Data Mining (EDM), the paper outlines methodologies for identifying student behaviors, modeling knowledge structures, and offering personalized educational support. Various machine learning techniques, including Artificial Neural Networks (ANN) and deep learning models, are reviewed to highlight their potential for improving accuracy in student performance predictions. Furthermore, the research evaluates the challenges faced by institutions and proposes solutions through predictive analytics for optimizing teaching strategies and improving educational outcomes. This study aims to contribute to the evolving domain of EDM by presenting a comprehensive overview of current methodologies and suggesting avenues for future research.

Key Words: Optimization, Machine Learning, Performance, Education, Data Mining

I. INTRODUCTION OF DATA MINING

In recent times, Data Mining (DM) is a process of identifying knowledge from massive databases for uncovering the patterns and trends involved in data. Massive collection of data in the information sector is transformed into meaningful data (Jothi et al., 2015). Apart from this, DM is an extraction of knowledge from large scale data that also computes the operations like Cleaning, Combination, Transformation, Mining, Estimation, and Projection. DM is classified into 3 phases namely, Descriptive, Classification, and Prediction (Vlahos et al., 2004). Initially, the descriptive process consults with common features of data while classification and prediction resolve the modules of data. Followed by, the descriptive function is composed of summarization and data mapping typically named data characterization as well as data differentiation. Predictive analytics is defined as a unification of Machine Learning (ML), historical information, and Artificial Intelligence (AI). Moreover, it helps examine the condition of recent details and examines the upcoming results. It becomes more prominent in applications like the economic sector, marketing, medical science, social media, and so on. The execution of predictive analytics is one of the trivial processes as it has massive challenging issues. The key objective of predictive analytics is to earn better profit with cost-effective and limited threats. There is a better solution for all predictive issues. It is also used for mysterious data in past, present, or future (Chandamona et al., 2016).

ISSN : 2455-9679
SJIF Impact Factor : 6.008

II. EDUCATIONAL DATA MINING (EDM)

In the last decades, EDM has attained massive attention from researchers due to the existence of massive educational details which is accessible from many sources. The main aim of EDM is to make DM models more effectively in order to safeguard the numerous amounts of educational information and to develop a protective atmosphere for the student's learning. In this approach, diverse models have been deployed for DM and its analytics (Baker *et al.*, 2014). Moreover, prediction models were used namely,
Classification, Regression, and Latent factor evaluation technologies.

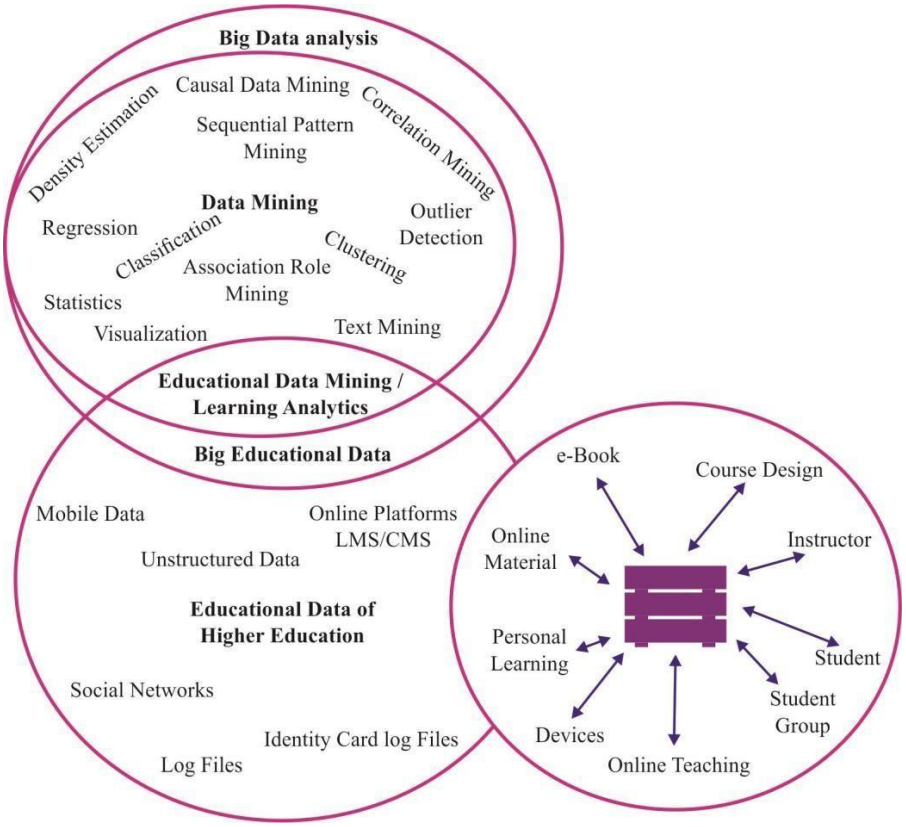
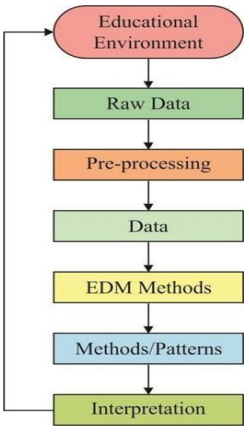


Figure 1. An illustration of DM use in HE

The DM tools are collaborated with academics in enhancing the students learning methodologies by exploring, filtering, and estimating the parameters relevant to student's features or behaviours (Baradwaj *et al.*, 2012) (Figure 1). The major challenges faced by any educational institutions lies in the number of placements it gets and the number of successful graduates it produces.

III. LIFECYCLE/ PROCESS INVOLVED IN EDM



The application of EDM models is composed of various phases (Figure 2). At the initial phase, the method is developed with the responsibility of identifying essential data. Then, data has been filtered from an accurate educational platform. Subsequently, data has to be pre-processed, as it is aroused from diverse sources with distinct templates and hierarchy levels. The identical patterns are attained while using EDM models which is interpreted. Finally, the simulation outcome recommends that using changes in the teaching process is not an appropriate solution and the analysis is carried out after changing the teaching process previously.

The EDM applications have been enhanced in recent decades. Based on the classified into 4

ISSN : 2455-9679
SJIF Impact Factor : 6.008

classes:

- Student modelling: student details (knowledge, motivations, and so on.) and EDM methods might be employed in developing a customized learning process by labelling the variations among students.
- Modelling the knowledge structural way of the field, which integrate the psychometric modelling approaches with space-searching technologies are developed in identifying the data-related applic[F]a[ig]ti[u]o[re]n[2]s:. Workflow of EDM
- Pedagogical care: Offers effective educational care.
- Scientific researches: domains are employed for developing and sample educational scientific strategies and to make novel hypotheses.

The methods are developed for detecting student efficiency, scientific investigation, giving feedback to support developers, recommending the students, developing alerts for stakeholders, student labelling, domain deployment, student grouping or profiling, introducing courseware, organization, scheduling, and parameter evaluation (Romero *et al.*, 2013).

IV. AN EDM EXAMPLE

In this method, consider that a tutor is permitted to use multidimensional information regarding the learning nature of the students in the online learning model. Especially, the student value of sign-in has been saved and the proportion of lesson accomplishments the value of student development. The input data point is represented as a black dot. This figure is drawn under the application of an online visualization system (Mohan, 2016). The main aim of a tutor is to categorize the student's behavior. But, labelling this classification is not an appropriate model. Also, it is considered to be general UL issues in ML, named as clustering (Murphy, 2012). When compared with diverse clustering models, a reputed k-means clustering mechanism has been applied. The parameter k implies the count of clusters desired. Initially, k-points for cluster centers have been selected randomly. The primary centroids are illustrated as triangles. Then, consider the points in a 2-D Cartesian coordinate plane and x coordinate point of a centroid is the mean value of x coordinate points that belongs to a cluster. The y coordinate of the cluster is estimated identically. A similar strategy uses points in n- dimension space. Thus, an applicable distance measure has to be defined for measuring distance among 2 points. Also, centroids are not considered as original data points. Followed by, every data point is divided into closest centroid relied on a distance measure. Diverse class labels of the points are represented in various colors. This categorization is not a novel clustering of points as cluster borders are slightly ineffective. Afterward, for all classes, a novel centroid has been re-determined. As the centroids were modified for students who signed for massive times; yet is not applicable in numerous progress, perhaps the short attention spans. According to the interpretation of 4 clusters, a tutor designs unique learning paths for various students. (Siemens *et al.*, 2012) indicates that EDM is used for automatic identification. Then, EDM's method in classifying the problems into tiny portions as it resolves the problems using ITS. This illustration is extended to detect learning results. Finally, (Figure 3) depicts some of the typical approaches applied in EDM, they are Clustering and Visualization.

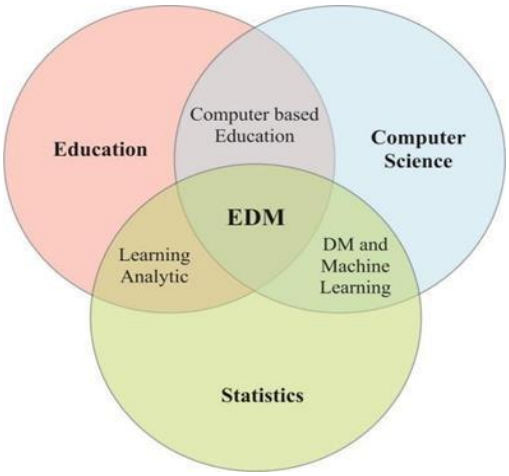


Figure 3: Main areas involved in Educational data mining

ISSN : 2455-9679
SJIF Impact Factor : 6.008

V. APPLICATIONS OF EDM IN HIGHER EDUCATION

Student modeling is defined as a way of representing the cognitive factors of student actions like examining student performance and behavior, separating the fundamental misconception, shows student aims and achievements, finding the advanced as well as acquired knowledge, retaining episodic memory, and defining personality features. In this approach, the definition of classifying EDM applications. Each domain in this class shows a technique which defines student's aims and goals. According to the study in (Chrysafiadi *et al.,* 2013) diverse features in student modeling, such as knowledge and skill (1), error and misconception (2), learning style with priority (3), affective and cognitive aspects (4) as well as metacognitive factor (5). The taxonomy of application in EDM (Figure 4).

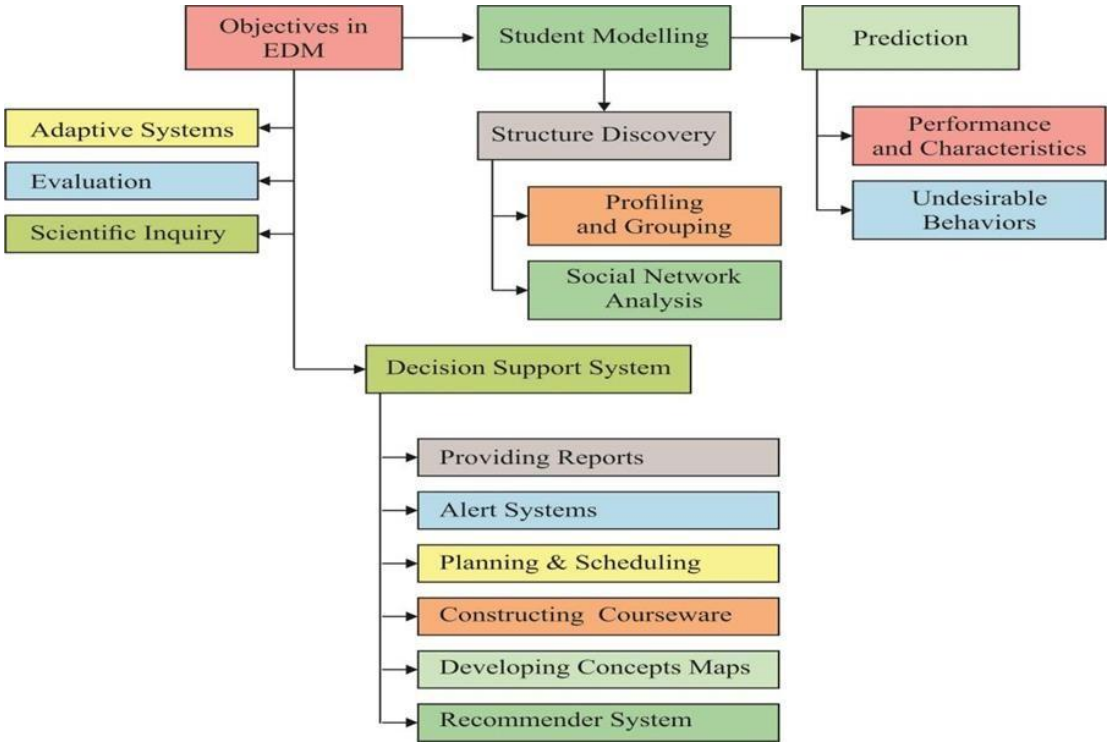


Figure 4: Classification of different EDM Applications

At the time of labeling student actions and behavior, the prediction of values shows the students identifying structures that define student's events. Consequently, 2 sub-classes in student modeling namely, Prediction and Structure discovery. In the case of a predictive process, a particular variable that desires to detect whereas in structure discovery, the specific attribute is not identified as depicted in structure, rather than using an individual feature. Therefore, it is significant to showcases that there is no clear line among 2 subcategories; however, the sufficient variations in an objective of 2 groups.

VI. BACKGROUND OF MACHINE LEARNING TECHNIQUES

The field of artificial intelligence known as machine learning gives intelligence to computers so that they can learn on their own without even being explicitly programmed to do so. This type of deep learning can be broken down into three main groups: learning that is supervised, learning that is unsupervised, and learning that is repeated. To be able to make predictions about new data, we need to get a model from a named or tested set. This is the primary objective of training set, often known as machine learning. can further divide classification model into two categories: a classification task, in which the anticipated outcome is a category value, and a regression task, in which the result is a positive value. Both of these categories fall under the umbrella of "supervised learning." The objective of relevance feedback is to design a system or perhaps an agent that can enhance its performance as a function of the way it interacts with its surrounding environment. This strategy includes both incentives and penalties, but it is very similar to supervised learning overall. An unsupervised learning task involves using data that has not been named and pulling out useful information from it

ISSN : 2455-9679
SJIF Impact Factor : 6.008

without knowing anything about it beforehand. For unsupervised learning, groupig can be used. Here, we will talk about some classification and regression methods that are related to our thesis.

Valentin Yunusov et.al. (2023) In this body of work, we look into how different educational factors affect the amount of academic success that school students achieve while getting used to the COVID-19-mandated format of remote learning. The Corona virus's outbreak affected not only the health care system but also every part of people's lives, even the school system. A big set of data from the library of the "Electronic education in Tatarstan Republic" system was used for the study. The dataset had all of the results that students got during the 2020 school year. Contemporary deep learning methods (feed-forward neural networks) and an approach for figuring out the relevance of features using the Captum framework were used to look at the data and come to conclusions. This approach not only lets us find the factors that affect the estimate the most, but it also lets us include numerical features so that we can compare and evaluate those features in a quantitative way. Among the things that were looked into were the teachers' traits, the average grades before the pandemic, the number of grades, the subject groups, and the schools' features. Because of this, we decided that the mean mark for the time before COVID19 is the most important thing. Because of this, the general trend of school grades has stayed pretty similar since then, though it's not exactly the same. We also paid a lot of attention to how students' grade point averages changed at different schools, types of schools, and teachers of different ages and genders. All of these changes can be explained by the fact that students spent different amounts of time getting used to a constantly changing setting.

Soha Ahmed et.al. (2022) Today, guessing how well a student will do in a virtual classroom is seen as a very important job. It includes many different learning activities that students do, such as signing up for classes, doing and turning in homework, taking tests, and interacting with others virtually, all of which are thought to be good places to do research. The field of prediction was also greatly affected by deep learning, a branch of artificial intelligence study. So, the study's goal was to talk about the part that AI plays in the e-learning system as a whole and, more specifically, the part that deep learning plays in guessing how well a student will do. This research found that most of the previous studies only looked at dropout prediction and ignored other success factors. These studies also didn't try to make the dataset better. So, the study was mostly about talking about these things. Using deep neural networks, the suggested model was very accurate (91.29%) and had a low loss value (0.18%) compared to other studies that used the same dataset. So, the study came up with a deep learning strategy to guess how well the student would do in school in the virtual classroom.

K. V. Deshpande et.al. (2023) The government has had to close schools in order to stop the spread of COVID-19. Because of this choice, there will be less interaction between students and teachers, which will also make it harder for them to talk to each other. The goal of this poll is to help students and teachers talk to each other more. With this poll, we wanted to learn more about the areas where the students need to improve, as well as the different things the teacher could do to help the student do better, and to find out if the above idea should be brought up again. Through reading the papers of other researchers, we learned that most of them made the same mistake in their studies. This led us to believe that the idea of AI should be looked into again, and we should try to avoid making the same mistake in our own research. With the help of Machine Learning (ML) and Deep Learning (DL), the main goal of our project is to create a "Teacher facing dashboard" that helps teachers summaries, visualise, and analyse data in the academics sector, as well as look at each student's performance.

Yuling Ma et.al. (2022) An accurate prediction of a student's success is very important in many educational settings, including academic early warning and individualized instruction. Many studies have been done to try to improve learning by using a lot of data about students. This data includes both the students' demographic details and their grades from previous classes. Many Chinese colleges have started using campus smartcards in the past few years. We can now record a lot of information about kids' behavior without bothering them, which gives us a new way to guess how well they will do in school. Contrary to most common approaches, the one we describe in this study is a new way to predict how well kids will do in school. In order to make the prediction models, HANDS (enHancing Academic Performa Nce via Deep foreSt) uses a decision tree-based deep learning method. Unlike most traditional methods, this approach is meant to be used instead. Furthermore, the HANDS method may automatically figure out how well students will do in their classes because it uses the end-to-end learning style that is linked to deep learning. As a result, it can cut down on the expensive use of human labour when compared to handcrafted feature-based methods. The outcomes of our tests using real-world data indicate that our approach is better than the most recent and widely accepted methods.

Ghaith Al-Tameemi et.al. (2021) The use of artificial methodologies by educational providers to predict students'

ISSN : 2455-9679
SJIF Impact Factor : 6.008

levels of achievement on the basis of their participation in virtual learning environments (VLE) is becoming increasingly common. These methods are being implemented in an increasing number of different types of educational settings. The Open University Learning Analytics Dataset (OULAD), which among other things includes information on student demographics, assessment scores, the number of clicks in the virtual learning environment, and final results, has been used to predict student performance in this paper. Other datasets have also been used. This dataset also contains information regarding other topics. During the pre-processing stage, a variety of different approaches have been utilized, including standardization and normalization. The degree of correlation that exists between the various kinds of activities and the overall performance of the students is quantified with the use of Spearman's correlation coefficient, which is then used to assess whether or not the activities are relevant by applying this coefficient. Deep learning was utilized to make a prediction regarding the performance of the students, with their participation in a virtual learning environment (VLE) functioning as the independent variable. The actual results demonstrate that our model had the ability to make accurate predictions regarding the academic progress of students.

Utkarsh Verma et.al. (2022) Educational Data Mining (EDM) uses big amounts of data from an education system to make predictions about how well students will do in school. Educational Data Mining is what EDM stands for. In order to make things easier for everyone in the school system, EDM tries to get useful information from the data that is collected. There are several machine learning (ML) techniques used in this study to predict how well a student will do in school. These techniques are based on real data collected from the students, such as their academic history and daily routines. A review of ML methods based on different evaluation criteria has also been given. It will be easier for students to keep track of how well they are doing in school and change the way they study based on what they find. This will help them get better grades in the future.

Pinaki Chakraborty (2020) the first few years of the 2000s saw the rise of social networking sites, which now play a big role in the daily lives of their users. Many high school and college students use Face book, which is by far the most well-known social networking site out there. As a way to learn more about how using Face book affects students' academic performance, I conducted this technological-behavioral study. I looked at the different and sometimes contradictory results of 22 research studies that were done in different countries around the world over the last ten years. Based on what I've seen, the main reasons students use Face book are to have fun and talk to people who like the same things they do. Students get worse grades when they use Face book in class or when they spend too much time on the site in general. Some students, though, use Face book to talk about school issues with their teachers and friends and to trade educational materials with each other. Some teachers also use learning management systems that they've made up on Face book by having group conversations to teach their students. The people who decide on policies for schools and the people who build social networking sites will both benefit from what I've found.

VII. CONCLUSION

This study underscores the transformative impact of machine learning and data mining on the education sector, particularly in predicting and enhancing student performance. By leveraging advanced techniques such as Artificial Neural Networks and deep learning models, significant improvements in accuracy and reliability of performance prediction have been achieved. The findings emphasize the necessity of integrating data-driven approaches into educational institutions to address challenges like dropout rates and learning variability. Moreover, the study highlights the importance of tailoring educational methodologies to individual student needs, supported by predictive analytics. Future research should focus on overcoming data challenges, improving model interpretability, and expanding the application of these methodologies to diverse educational contexts, ensuring equitable and impactful outcomes for all learners.

REFERENCES

[1.] K. AL-Dulaimi, J. Banks, K. Nugyen, A. Al-Sabaawi, I. Tomeo-Reyes, and V. Chandran, ‘‘Segmentation of white blood cell, nucleus and cytoplasm in digital haematology microscope images: A Review–challenges, current and future potential techniques,’’ IEEE Rev. Biomed. Eng., vol. 14, pp. 290–306, 2021.

[2.] L. Bigorra, A. Merino, S. Alférez, and J. Rodellar, ‘‘Feature analysis and automatic identification of leukemic lineage blast cells and reactive lymphoid cells from peripheral blood cell images,’’ J. Clin. Lab. Anal., vol. 31,

no. 2, Mar. 2017, Art. no. e22024.

[3.] Y. Liu and F. Long, ‘‘Acute lymphoblastic leukemia cells image analysis with deep bagging ensemble learning,’’ in CNMC Challenge: Classification in Cancer Cell Imaging. Singapore: Springer, 2019, pp. 113–121

[4.] J. W. Choi, Y. Ku, B. W. Yoo, J.-A. Kim, D. S. Lee, Y. J. Chai, H.-J. Kong, and H. C. Kim, ‘‘White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks,’’ PLoS ONE, vol. 12, no. 12, Dec. 2017, Art. no. e0189259.

[5.] N. Baghel, U. Verma, and K. K. Nagwanshi, ‘‘WBCs-Net: Type identification of white blood cells using convolutional neural network,’’ Multimedia Tools Appl., vol. 162, pp. 1–17, Sep. 2021.

[6.] R. M. Roy and A. P. M., ‘‘Segmentation of leukocyte by semantic segmentation model: A deep learning approach,’’ Biomed. Signal Process. Control, vol. 65, Mar. 2021, Art. no. 102385.

[7.] K. A. K. Al-Dulaimi, J. Banks, V. Chandran, I. Tomeo-Reyes, and K. N. Thanh, ‘‘Classification of white blood cell types from microscope images: Techniques and challenges,’’ Tech. Rep., 2018.

[8.] Trachtman, Joel P., The Internet of Things Cyber security Challenge to Trade and Investment: Trust and Verify? (April 18, 2019). Available at SSRN: https://ssrn.com/abstract=3374542 or http://dx.doi.org/10.2139/ssrn.3374542

[9.] Mohammad, Sikender Mohsienuddin, Security and Privacy Concerns of the 'Internet of Things' (IoT) in IT and its Help in the Various Sectors across the World (April 4, 2020). International Journal of Computer Trends and Technology (IJCTT) – Volume 68 Issue 4 – April 2020, Available at SSRN: https://ssrn.com/abstract=3630513

[10.] Valentin Yunusov;Fail Gafarov;Pavel Ustin Deep learning techniques for the study of student's academic performance during distance education caused by COVID-19 2023 17th International Conference on Electronics Computer and Computation (ICECCO) Year: 2023 |

[11.] Soha Ahmed;Yehia Helmy;Shimaa Ouf A deep learning framework for predicting the student's performance in the virtual learning environment 2022 5th International Conference on Computing and Informatics (ICCI) Year: 2022 |

[12.] K.V.Deshpande;Shubham Asbe;Akanksha Lugade;Yash More;Dipali Bhalerao;Anuradha Partudkar Learning Analytics Powered Teacher Facing Dashboard to Visualize, Analyze Students’ Academic Performance and give Key DL(Deep Learning) Supported Key Recommendations for Performance Improvement. 2023 International Conference for Advancement in Technology (ICONAT) Year: 2023 |

[13.] Yuling Ma;Huiyan Qiao;Xiwei Sheng;Xiaoli Wang;Zhen Li HANDS: enHancing Academic performaNce via Deep foreSt 2022 15th International Conference on Human System Interaction (HSI) Year: 2022 |

[14.] Ghaith Al-Tameemi;James Xue;Suraj Ajit;Triantafyllos Kanakis;Israa Hadi;Thar Baker;Mohammed Al-Khafajiy;Rawaa Al-Jumeily A Deep Neural Network-Based Prediction Model for Students’ Academic Performance 2021 14th International Conference on Developments in eSystems Engineering (DeSE) Year: 2021 |

[15.] Utkarsh Verma;Chetna Garg;Megha Bhushan;Piyush Samant;Ashok Kumar;Arun Negi Prediction of students’ academic performance using Machine Learning Techniques 2022 International Mobile and Embedded Technology Conference (MECON) Year: 2022 |

[16.] Pinaki Chakraborty, Effects of Using Facebook on Academic Performance of Students: A Review2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) Year: 2020 |

[17.] Eluri Roopa;B.Eswara Reddy Predicting JNTUA CEA Student's Academic Performance Using Deep Neural Networks 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) Year: 2023 |

[18.] Yuanyi Zhen;Jar-Der Luo;Hui Chen Prediction of Academic Performance of Students in Online Live Classroom Interactions—An Analysis Using Natural Language Processing and Deep Learning Methods Journal of Social Computing Year: 2023 |

[19.] Bui Ngoc Anh;Nguyen Hoang Giang;Ngo Quang Hai;Trinh Nhat Minh;Ngo Tung Son;Bui Dinh Chien An University Student Dropout Detector Based on Academic Data 2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA) Year: 2023.

[20.] Abu Tair, M.M. and El-Halees, A.M. (2012). Mining educational data to improve students' performance: a case study. *International Journal of Information and Communication Technology Research.*, *2*(2):140-146.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

[21.] Adekitan, A.I. and Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon.*, *5*(2): e01250.