International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Survey on Student Data Mining Indicators and Techniques for Grade Predictions

[1]Rishi Kumar PhD Scholar, [2]Associate Professor Dr. Pritaj Yadav, [3]Professor Dr. Alok Katiyar

[1,2]Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, MP, India
[3]School of Computer Science and Engineering, Galgotias University, Greater Noida, UP

Abstract- The rise of digital platforms has significantly enhanced the management and analysis of educational data. In this context, data mining techniques play a crucial role in extracting meaningful patterns from raw student data, which can be utilized for predicting academic performance and improving learning outcomes. This paper presents a comprehensive survey of recent research focused on student data mining, with particular emphasis on indicators and techniques used for grade prediction. Various data mining methods such as classification, clustering, and regression are discussed in detail, highlighting their applications in academic settings. Additionally, the study examines key indicators that influence student performance and explores privacy-preserving approaches to ensure the ethical use of student data. Evaluation parameters for comparing the effectiveness and issues aspects of different techniques are also analyzed to guide future research in this domain.

Keywords- Data mining, Information Extraction, Association Rule,

I. INTRODUCTION

With rapid technological advancements and the widespread adoption of computing devices, the volume and variety of data have grown exponentially. In the era of big data, much of the collected information includes not only core content but also valuable auxiliary insights and hidden patterns. These developments in computing technologies have significantly impacted various sectors—including healthcare, transportation, industry, education, commerce, and social communication—driving transformative changes and unlocking new economic opportunities.

One emerging application of these advancements is the prediction of student academic performance, which has gained considerable attention in recent years. Accurate grade prediction can benefit both students and educators. For students, it offers guidance in selecting courses more strategically based on their predicted outcomes, potentially enhancing their overall academic success. For instructors, such predictions enable better planning and adjustment of teaching strategies to support diverse learner needs and improve course effectiveness.

Neural networks have proven to be effective in addressing numerous problems within educational data mining. For instance, Sharma et al. [1] introduced a composite deep neural network to determine the liveliness of educational videos. Their model utilized a convolutional neural network (CNN) to extract visual features from videos and a deep recurrent neural network (RNN) to analyze human motion, ultimately classifying the video's liveliness.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

In another study, Elbadrawy et al. [2] developed a semi-supervised classification framework leveraging deep variational autoencoders to identify students with developmental dyscalculia. This model demonstrated how deep learning could uncover latent cognitive issues in educational contexts.

Similarly, Piech et al. [3] proposed the Deep Knowledge Tracing (DKT) model, which employed recurrent neural networks to track and predict student learning behavior. Their experiments revealed that DKT could effectively model the hidden relationships among various assessments, thus providing insights into student knowledge progression.

Building on this line of research, current models aim to improve grade prediction for future academic terms using students' past academic records and course enrollment patterns. A key innovation in such models is the explicit inclusion of co-enrolled courses within a matrix factorization (MF) framework, thereby enhancing the model's predictive capabilities by capturing interactions among simultaneously taken courses.

II. INDICATORS OF GRADE PREDICTION

Engagement: In recent years, there has been a growing interest in learning analytics focused on student engagement, significantly broadening the scope of educational research. Higher education institutions are increasingly adopting analytics tools to better understand and enhance student engagement. These tools can serve as a bridge for improved communication between students and instructors, contributing to more effective learning experiences, heightened awareness, and more responsive handling of academic challenges [4]. Engaged students tend to perform better academically and often find greater satisfaction in the learning process. Studies have demonstrated that engagement is a key factor in continuous learning, long-term academic success, and overall student well-being.

In [5], the authors provided an overview of various methods used to detect student engagement in online learning environments, outlining associated challenges. These methods were grouped into three categories: fully automated, semi-automated, and manual. Automated methods typically collect behavioral data from digital platforms. Among these, log files have proven particularly valuable in capturing learner activity patterns in virtual settings. For instance, the study in [6] utilized log data from an online tool called HTML-tutor, extracting 30 different features such as the number of tests taken, correct answers submitted, pages visited, and more [5].

Demographics: Demographic information is frequently used to build predictive models aimed at identifying students who are at academic risk—those likely to fail courses or drop out of educational programs [7]. Demographic variables typically include attributes such as age, gender, ethnicity, income level, education background, and religious affiliation. Despite their usefulness, these features often raise ethical and privacy concerns, particularly around data access and sharing [8]. Since demographic data can potentially reveal the identities of individuals, strict privacy measures are required to ensure data protection. One common approach is pseudonymization, such as applying k-anonymity techniques [9]. However, achieving meaningful anonymization while retaining the predictive utility of the data remains a major challenge and often diminishes the effectiveness of these demographic features in model training.

III. LITERATURE SURVEY

Hussain et al. [10] conducted an evaluation of student academic performance using four classification algorithms: J48, PART, BayesNet, and Random Forest. The study was based on 12 features reflecting

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

both academic and personal attributes. Among the evaluated techniques, Random Forest was identified as the most accurate classifier with the least classification error. Additionally, the Apriori algorithm was implemented using WEKA to extract the most significant association rules from the dataset.

In a similar effort, Hasan et al. [11] explored predictive modeling of students' semester-end academic success. They utilized data gathered from various sources including student information systems, learning management systems, and mobile applications. The dataset was processed using eight different classification algorithms. To enhance model performance and reduce feature complexity, preprocessing methods such as data transformation, genetic search, and principal component analysis were employed. Tools like CN2 Rule Inducer and multivariate projections were used to help faculty gain deeper insights into student behavior. Their results indicated that Random Forest achieved the highest prediction accuracy of 88.3% using equal width discretization and information gain ratio for feature evaluation.

Zhang et al. [12] conducted a comprehensive review of student performance prediction from both data mining and machine learning perspectives. They organized the predictive process into five key stages: data collection, problem definition, model development, performance prediction, and real-world application. Two datasets were used—one consisting of records from 1,325 students across 882 courses at a Chinese university, and another public dataset. Techniques such as naive Bayes, decision trees, support vector machines, bagging, and random forest were applied using the WEKA 3.8 tool. The input attributes were grouped into two categories: demographic/background features and grades from prerequisite courses. Their analysis showed that Random Forest outperformed other algorithms in accuracy and that prerequisite course grades had the strongest influence on predictions, while demographic data had minimal effect.

Niyogisubizo et al. [13] proposed a stacked ensemble model combining Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Artificial Neural Networks (ANN) to identify students at risk of dropping out from individual courses. The dataset, comprising 261 student records and 12 features collected from 2016 to 2020 at Constantine the Philosopher University in Nitra, included data points such as test scores, project evaluations, final grades, and academic year information.

Sahlaoui et al. [14] compared several machine learning models, including RF, ANN, Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, Bagging, and XGBoost. The research utilized a dataset from a public university in Jordan containing 480 student records and 16 independent variables across different academic levels and subjects. The models aimed to classify students into three categories—high, average, and low performers—based on the likelihood of course dropout.

In an innovative study, Ghazvini et al. [15] introduced a novel loss function called MSECosine, which integrates Mean Squared Error (MSE) with LogCosh to reduce error amplification. This custom loss function was used to train four deep learning time-series models: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a hybrid CNN-LSTM. Two datasets, comprising student assessment and LMS data, were used for experimentation. Among these models, LSTM emerged as the most effective. Enhancing the LSTM with the proposed MSECosine loss function resulted in a superior model named eLSTM, which demonstrated improved prediction performance.

Sun et al. [16] proposed a new predictive model leveraging multi-feature fusion and attention mechanisms to evaluate student academic performance at the university level. This approach analyzed students' historical grade records across multiple dimensions to capture intricate relationships between courses, between students, and across student-course interactions. The incorporation of an attention

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

mechanism allowed the model to focus on the most relevant features. By using a dataset of related course triplets and actual student grades, the study validated the effectiveness of multi-dimensional features and demonstrated how course interrelationships influence performance predictions.

Alhazmi et al. [17] explored both clustering and classification approaches to assess early-stage academic performance and its effect on GPA. For clustering, they used T-SNE for dimensionality reduction, applying it to early indicators like admission scores, entry-level course grades, and standardized test results such as AAT and GAT. For classification, various machine learning models were applied using course grades and admission test scores. The findings confirmed that analyzing early academic indicators can help educational institutions identify and address the risk of academic failure early in a student's academic journey.

Table 1. Comparison of existing models.

Models	Technique	Advantage	Limitations
A. Ghazvini et. al. 2024	LSTM	Predict the student grade by overcoming the gradient issues of learning.	Need some pre-processing techniques for improving learning.
D. Sun et al. 2023	Multi feature Fusion	University student grade prediction.	Prediction accuracy is low.
Y. Liu et. al. 2023	Deep Learning and attention mechanism (MCAG)	Student Achievement Prediction.	Need to learn about behavior as well.
J. Niyogisubizo et. al. 2022	Ensemble Models	Student dropout prediction.	Large number of models increase complexity.
H. Sahlaoui et. al. 2021	synthetic minority oversampling technique (SMOTE)	School Student Dropout prediction.	Need to generalize for grade prediction,.

IV. DATA MINING TECHNIQUES

Decision Tree: Decision tree classification involves constructing a model in the form of a hierarchical structure where internal nodes represent attribute tests, branches denote the outcomes of these tests, and leaf nodes specify class labels. This tree-like representation enables straightforward classification and interpretation. Decision trees stand out due to their simplicity, interpretability, and flexibility in handling both categorical and numerical data. Moreover, they require minimal data preprocessing and offer robustness even when model assumptions are not strictly satisfied. Due to their efficiency, they are capable of handling large datasets with fast processing times on conventional computing systems, aiding timely decision-making processes [10].

K-Nearest Neighbor (K-NN) Classifier

The K-NN algorithm is a fundamental method used for both classification and regression. It operates on the principle that instances with similar features are likely to belong to the same class. The algorithm identifies the 'k' closest data points in the training set to a new instance and assigns a class based on the majority class among these neighbors. Typically, Euclidean distance is used to measure similarity in numerical datasets, while alternative metrics like Hamming distance are used for categorical or textual data. The model does not learn explicitly during training, which categorizes it as a lazy learning

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

approach. K-NN's simplicity and effectiveness make it a popular baseline method in predictive modeling studies, including those dealing with educational data mining [11].

Artificial Neural Network

Inspired by the functioning of the human brain, Artificial Neural Networks (ANNs) are widely used in machine learning for modeling complex nonlinear relationships. An ANN consists of interconnected nodes (neurons) arranged in layers, and it learns patterns from data by adjusting the connection weights during training. The learning process enables the network to predict outcomes for new, unseen inputs. ANNs have been used extensively in educational settings. For instance, Niyogisubizo et al. [13] incorporated ANNs in an ensemble model to predict student dropout risks, demonstrating the capability of neural models to capture intricate patterns related to academic performance and behavioral attributes.

Support Vector Machine (SVM)

Support Vector Machines are powerful supervised learning models used for classification and regression tasks. By mapping input data into high-dimensional space, SVMs identify the optimal separating hyperplane that maximizes the margin between two classes. The margin is defined by the distance between this hyperplane and the nearest data points from each class—referred to as support vectors. A wider margin typically results in better generalization performance. SVMs have been employed in numerous educational prediction tasks, including the work of Zhang et al. [12], where they were used alongside other models to analyze student academic data. Their study found SVMs to be competitive in performance, particularly in complex, high-dimensional datasets.

Association Rule Mining

Association rule mining is a data analysis method designed to uncover interesting relationships, patterns, or correlations among sets of items in large datasets. This technique is particularly useful in domains like market basket analysis, where understanding co-occurring items can inform cross-selling strategies. In the context of education, rule mining helps reveal behavioral patterns that influence learning outcomes. Hussain et al. [10], for example, used the Apriori algorithm in the WEKA tool to extract association rules from student data, identifying significant patterns linked to academic performance. While this method can generate a vast number of rules, not all of them are meaningful, and thus it is essential to filter rules using metrics like support and confidence.

V. IDENTIFIED ISSUES & EVALUATION PARAMETERS

Education professional focus on the lecture and training of young scholars. So other technical issues that directly related to scholars security is depend on university or higher governing bodies. Many of researcher work on different segment for data security by network, hardware and software solutions. Each of solution takes to the transformation of information from one form to other. But some situations arise when data sharing is important but sharing may increase the chance to extract hidden information. So privacy preservation come in existence. Scholar data is highly sensitive for individual, community, nation, etc. Many of researcher work in this field of education data with privacy preservation to find pattern of learning, performance, course impact analysis, etc.

It was found that scholar data need feature analysis first with privacy preserving mining approach. Further it was found that most of research work was done in numeric data type only, so generalize model should be develop as scholar data has all type of content. Some of researcher apply encryption algorithm for privacy but that is security not privacy.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Parameters

In order to compared proposed model with other existing models following set of parameters were evaluated [75]. Where TP is true positive, FP is false positive, similarly TN is true negative and FN is false negative. All are counter having 0 initial value. Let us understand for Average Grade. TP increments if a student predicted grade class is average and actual grade class is also average. Similarly TN increments if a student predicted grade class is average and actual grade class is also other than average. In case of FP increments done if a student predicted grade class is other than average and actual grade class is average Similarly FN increments if a student predicted grade class is other then average and actual grade class is also other than average.

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

F_Measure=(2*Prcision*Recall)/(Precision+Recall)

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Execution Time: In order to compared the proposed algorithm execution time was also compared with existing models. Time was estimated for the prediction of grades.

VI.CONCLUSION

Data mining plays a crucial role in uncovering patterns, making predictions, and extracting meaningful insights across various sectors, including business, education, and healthcare. Techniques such as classification and clustering support decision-making by identifying trends that can drive growth and innovation. This paper has provided a comprehensive overview of key data mining methods, with a particular focus on approaches aimed at safeguarding sensitive and private information. Despite the progress in this field, existing algorithms still have limitations in fully ensuring data security. Therefore, there is a need for the development of more advanced and secure data mining techniques that can offer stronger protection while maintaining analytical effectiveness.

REFERENCES

1. [Arjun Sharma, Arijit Biswas, Ankit Gandhi, Sonal Patil, and Om Deshmukh. Livelinet: A multimodal deep recurrent neural network to predict liveliness in educational videos. In EDM, pages 215–222, 2016.
2. Asmaa Elbadrawy, Agoritsa Polyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, and Huzefa Rangwala. Predicting student performance using personalized analytics. Computer, 49(4):61–69, 2016
3. Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In Advances in Neural Information Processing Systems, pages 505–513, 2015.
4. Silvola, A., Näykki, P., Kaveri, A., & Muukkonen, H. (2021). Expectations for supporting student engagement with learning analytics: An academic path perspective. Computers & Education, 168, 104192.
5. Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. Smart Learning Environments, 6(1).

6. Cocea, M., & Weibelzahl, S. (2011). Disengagement detection in online learning: Validation studies and perspectives. IEEE Transactions on Learning Technologies, [online] 4(2), pp.114–124.
7. S. Alturki, I. Hulpu, and H. Stuckenschmidt. Predicting academic outcomes: A survey from 2007 till 2018. Technology, Knowledge and Learning, pages 1–33, 2020.
8. G. Fenu, R. Galici, and M. Marras. Experts' view on challenges and needs for fairness in artificial intelligence for education. In International Conference on Artificial Intelligence in Education, pages 243–255. Springer, 2022.
9. L. Sweeney. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems, 10(05):557–570, 2002.
10. S. Hussain, N.A. Dahan, F.M. Ba-Alwib, N. Ribata Educational data mining and analysis of students' academic performance using WEKA Indones. J. Electr. Eng. Comput. Sci., 9 (2) (2018), pp. 447-459 Feb 2.
11. [30] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K.U. Sarker, M.U. Sattar Predicting student performance in higher educational institutions using video learning analytics and data mining techniques Appl. Sci., 10 (11) (2020), p. 3894 Jun 4
12. Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, X. Shang Educational data mining techniques for student performance prediction: method review and comparison analysis Front. Psychol., 12 (2021).
13. J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization," Computers & Education: Artificial Intelligence, vol. 3, Article ID 100066, 2022.
14. H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and shapley additive explanations," IEEE Access, vol. 9, pp. 152688–152703, 2021.
15. A. Ghazvini, N. Mohd Sharef and F. B. Sidi, "Prediction of Course Grades in Computer Science Higher Education Program via a Combination of Loss Functions in LSTM Model," in IEEE Access, vol. 12, pp. 30220-30241, 2024.
16. D. Sun et al., "A University Student Performance Prediction Model and Experiment Based on Multi-Feature Fusion and Attention Mechanism," in IEEE Access, vol. 11, pp. 112307-112319, 2023.
17. E. Alhazmi and A. Sheneamer, "Early Predicting of Students Performance in Higher Education," in IEEE Access, vol. 11, pp. 27579-27589, 2023.
18. Y. Liu, Y. Hui, D. Hou and X. Liu, "A Novel Student Achievement Prediction Method Based on Deep Learning and Attention Mechanism," in IEEE Access, vol. 11, pp. 87245-87255, 2023.
19. Kingsley Okoye, Julius T. Nganji, Jose Escamilla, Samira Hosseini. "Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education", Computers and Education: Artificial Intelligence, Volume 6, 2024.
20. Ran Song, Fei Pang, Hongyun Jiang, Hancan Zhu. "A machine learning based method for constructing group profiles of university students", Heliyon, Volume 10, Issue 7, 2024.