Grenze International Journal of Engineering and Technology, Jan Issue

A Review on Students' Performance Prediction using Learning Analytics

Jyoti[1], Dharminder Kumar[2] and Sakshi Dhingra[3]
[1-3]Department of Computer Science and Engineering Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India
Email: jyoticherharval12@gmail.com, dr_dk_kumar_02@yahoo.com, sakshi24.dhingra@gmail.com

***Abstract*— Learning Analytics is a research area that is growing rapidly. It deals with selecting, analyzing, and reporting educational data collected from various learning environments and finding relevant patterns in students' behavior. Different online learning platforms like Massive Open Online Courses (edX, Udemy, Udacity, etc.), Learning Management Systems (Moodle, Blackboard, Classroom, etc.), and Virtual Learning Environments (DEEDS, Coursera, etc.). Learning analytics can be viewed as a multidisciplinary field that includes Machine Learning (ML) Techniques, Educational Data Mining (EDM), Statistics, Social Network Analysis, and Natural Language Processing. This paper comprises the study of Learning Analytics and its different techniques used in predicting students' performance. Further, it consists of an analysis of previous studies described in tabular form for better understanding. A basic learning analytics model is described which consists of the steps involved in creating any project.**

***Index Terms*— Learning analytics, LMS, VLE, MOOC and ML**

I. INTRODUCTION

The traditional educational landscape is changing day by day as technology is evolving rapidly. Especially after the COVID pandemic, the field of education has witnessed a paradigm shift towards the integration of technology in Virtual Learning platforms. Virtual learning offers remarkable flexibility and accessibility, enabling students to pursue education remotely and at their own pace. The amount of data generated from these learning environments is increasing exponentially. This data can be used to gain valuable insights into learning behaviors, engagement, and performance. The study of Learning Analytics has become a powerful tool for extracting valuable insights and hidden patterns from the massive amount of data collected from learning environments.

Learning analytics can be understood as the process of collecting, analyzing, and interpreting data from learning environments in order to understand and optimize the learning process as well as the environment in which it occurs [16]. By exploiting the vast amount of data generated within VLEs, teachers or educators can identify new patterns and predict students' performance. It allows them to personalize instructions, provide timely interventions, and improve educational outcomes. By Learning Analytics, researchers can develop new models and algorithms to predict students' performance by analyzing their past behaviors, engagement level, learning activities, work accomplishments, and their interactions within the VLEs (Virtual Learning Environments). Learning Analytics has become a hot topic for researchers due to many reasons. These are:

1. Increase in volume of data available about learners and learning techniques, especially in VLEs where every small interaction, each page visited, and each click can be recorded and used for future analysis [17].

Grenze ID: 01.GIJET.10.1.354

2. Prediction of students' performance at an early stage and management of learning environments in order to enhance the retention ratio and lower students' dropout from online courses [17].
3. Increase in no. of statistical tools to manage large amounts of datasets and facilitate educators [17].

Learning analytics is closely related to EDM (Educational Data Mining) and AA (Academic Analytics) [18]. EDM and LA are the two overlapping research areas as they are used interchangeably in many research [8]. The domain, datasets, methods, and objectives in EDM and LA are pretty similar yet there exist some differences.

On the basis of different previous studies, the differences between Academic Analytics, Educational Data Mining, and Learning analytics are shown in this section using the table shown below:

TABLE I. DIFFERENCE BETWEEN VARIOUS DATA ANALYTICS METHODS IN EDUCATION DOMAIN

Techniques	Description
Academic Analytics	It is defined as the use of business intelligence in education. It focuses on enrollment management and the prediction of students' academic success thus facilitating the funders and administrators. Its primary goal is to solve the political/economic challenges to improve educational opportunities and outcomes at national or international levels.
Educational Data Mining	It is defined as the use of Data Mining techniques to extract useful insights from large educational data It involves application of Machine Learning, Statistical Analysis, and Data visualization to uncover the hidden patterns in students' behaviors. It focuses more on technical challenges unlike LA which focuses on Educational Challenges.
Learning Analytics	It focuses on the use of students data collected from students' interaction with Learning environments in order to understand and evaluate the learning process. It includes Data Mining techniques along with Statistical and Visualization tools and Social Network Analysis. It uses quantitative methods and it is the first and foremost which is concerned with learning.

A. Challenges

The challenges faced by online learning environments are high drop-out rates, low performance of students, and an increase in the number of at-risk students. A study shows that the no. of uncertified students is way more than the certified students in an MOOC environment [21]. Learning analytics techniques can be used to determine students' performance at early stages thus minimizing the drop-out rates. The amount of data collected and methods like preprocessing and feature selection have a great impact on predicting students' performance [3][5]. Handling missing values and balancing the imbalanced dataset using appropriate algorithms can improve the accuracy of different predictive models. Learning Analytics would help in understanding learners' behavior and improving performance by early interventions.

II. LITERATURE REVIEW

A tabular comparison of previous work has been included in this section for a better understanding of the methods used, their work, results, and conclusions as shown in the table:

TABLE II. PREVIOUS WORK ANALYSIS

Author and year of study	Methods used in the study	Proposed Work	Results	Conclusion
Hilal Almarabeh, (2017) [1]	NB (Naïve Bayes), ID3 (Iterative Dchotomiser 3), J48(Decision Tree), Bayesian Network, NN (Neural network)	In this study: Data of 255 instances each having 10 attributes is collected from the university database. Different ML algorithms are compared based on Accuracy. Tool used for implementation is WEKA. Error measurement is done using MAE, RMSE, RAE, RRSE.	Bayesian network classifier have the highest accuracy among the classifiers. It has a minimum error value for RMSE and RRSE compared to others.	Data collected is very small. More dataset instances can be collected for better evaluation. Different DM algorithms like association and clustering could be used.

Danijel kucak et al, (2018) [2]	SLR (Systematic Literature Review) method	In this study 67 papers were collected from sources like IEEE Xplore library, Scopus database, Web of Science (WoS), ScienceDirect, Google Scholar, and others. The data can be used for: a. Giving grades to students without human biasness b. Improving students' retention rates. c. Students' performance prediction. d. Evaluating students.	ML can be used to grade students thus eliminating human biases. Identifying at-risk students early can help to maintain student retention. Using ML to predict students' performance is a hot topic for research.	ML is really helpful in the education field. New models can be created using different algorithms for different purposes.
Ali salah hashim et al, (2020) [3]	LR, DT, NB, SVM, K-NN, NN, SMO	Different ML techniques are compared on the basis of accuracy measure. The proposed methodology was divided into two stages: - 1. Data cleaning -- 499 records selected. 2. Cronbach's alpha is calculated to find internal consistency of records. 3. classification algorithms are applied using the WEKA tool.	LR performed best with accuracy of 88.8%.	It found that factors like Dataset size, domain and no. of features affects the accuracy. ROC curve and AUC helps to evaluate model's performance and accuracy of predicted values.
Sandeep Subash Madnaik, (2020) [4]	Regression model, BinaryClassification, ADA Boost classification, DT (Decision Tree), RF (Random Forest)	This study considers two different kinds of datasets and applies different ML Algorithms. It compares the current work with the previous one. It considers both the academic as well as social behavior features for predicting performance.	Random forest performed best with accuracy of 93%.	It concluded that academic features affect performance more than social features. Large amounts of data results in more accurate predictions.
Siti Dianah et al, (2021) [5]	DT(J48), SVM, NB, K-NN, LR, RF	This study proposed a multiclass prediction model SFS to reduce overfitting and misclassification results. It compared the accuracy performance of 6 ML algorithms. It uses SMOTE with Feature Selection algorithms (Wrapper and Filter based).	Proposed model integrates with Random Forest performed best with highest f-measures 99.5%	Using Feature Selection with imbalance classifications increased the accuracy. Ensemble algorithms are recommended to optimize the results for predicting students' grade.
Harikumar Pallathadka et al, (2021) [6]	SVM NB C4.5 ID3	This study compared the accuracy of different ML algorithms. It examines the UCI Machinery student performance data.	SVM performed the best among all algorithms with highest accuracy.	Predicting students' performance can help institutions to improve their instruction process.
Muhammad Adnan et al, (2021) [7]	RF (Random Forest), SVM, KNN, ANN ADA Boost classifier, Gradient Boosting Classifier DFFNN (Deep Feedforward Neural Network)	It compared performance of different algorithms using different variables. K-Fold Cross Validation and feature engineering was used to improve accuracy. Python 3.7.8 and libraries like TensorFlow, Keras, sklearn are used for implementation. Heatmaps and confusion matrices were used to check the correlation and performance	RF algorithm performed best as compared to other algorithms.	Early intervention can help in reducing students' dropout. Deep Learning models and natural language processing can be used for performance modeling using textual data.

Dalia Abdulkareem shafik et al (2022) [8]	SLR (Systematic Literature Review) method.	This paper consists of review of different research papers, journals, and conferences. It highlights the use of EDM, LA and ML algorithms in educational field.	Supervised ML algorithms are used more frequently. Most popular tool that is used is WEKA, RStudio.	Limited data is used. Clustering techniques are not used much. Combination of approaches can be used.
Guiyan Feng et al, (2022) [9]	K-means clustering algorithm CNN (Convolutional Neural Network)	This study focuses on using clustering, discrimination and convolutional neural networks in improving the accuracy of predictive models. It evaluates the performance of traditional k-means algorithm and tries to improvise it.	The proposed model has performed more accurately than existing one	Missing data handling issue. Applying EDM in learning analytics in a VLE.
Md. Hasib khan et al, (2022) [10]	LR KNN SVM XG Boost NB	This study proposed a predicational model Interpretable LIME based on AI. It first applied ML Classifiers for predicting students' performance. Used dataset from two schools. SMOTE (Synthetic Minority Oversampling Technique) is used to balance the imbalanced datasets.	It found that SVM performed better than all other classifiers with an accuracy of 96.89%	Interpretable explanations help in building students understanding in model's human friendly explanations. Recommender systems can be developed using ML and Deep learning in the education field.
Yazan A. Alsariera et al, (2022) [11]	SLR (Systematic Literature Review)	Reviewed articles from years 2015 to 2021. 39 articles selected for review. Highlights use of ML algorithms like DT, SVM, LinR (Linear Regression), Artificial Neural Network, k-Nearest Neighbor), NB.	KNN Classifier gives more accurate results than others.	New datasets and more resources can be used to find more accurate results. More algorithms can be included in future studies.
Reyhan Zeynep pek et al, (2023) [12]	NB, RF DT, KNN SVM AdaBoost Classifier, LR	A hybrid model is created using Ensemble Stacking Method. It uses stratified k-fold cross validation for preprocessing of data with k=10. Hyperparameter optimization techniques are used to improve performance. SVM is used as meta learner and all others as base learners.	The model gives accurate results with pass rate 77.48% and failure rate 22.52%.	The data is collected manually and random values are added for filling the null spaces. Can be used for university student data.
Siti Dianah et al, (2023) [13]	Reviewed articles from 2015 to 2021. 41 publications are considered.	This article highlights the problem of imbalanced classification in prediction models	It found SMOTE as less effective in case of high dimensional datasets.	Feature Engineering, Sampling method can be used to improve accuracy. Hybrid approaches can be used in resolving the multi-class problem to increase accuracy of any predictive model.
Fan Ouyang et al, (2023) [14]	Genetic Programming, SNA (Social Network Analysis), Context analysis, T-test	It used Integrated approach by combining AI (Artificial Intelligence) and LA (Learning Analytics). Quasi-Experimental research was conducted. Paradigmatic	Both AI and LA together can help in predicting performance more efficiently.	Integration of AI and LA should be imposed in future for supporting students' success

	Thematic Analysis ML Algorithms	implication and Closed Loop were implemented for AI-driven applications.		and decision making.
Blessy Paul P et al, (2023) [15]	Decision Tree, CART (Classification and Regression Tree), ANN (Artificial Neural Network)	It uses ML algorithms to predict students' performance in e-Learning environments. Performance was evaluated using confusion matrices. Behavioral features and students' interaction with the learning platform is measured.	ANN algorithm outperformed with an accuracy of 0.916. Efficiency of ANN is measured by calculating AUROC Curve.	Student interaction with learning platforms is the major element in predicting students' performance. Focused online learning without any misuse of the internet can lead to a better future.

III. LEARNING ANALYTICS MODEL

While developing any learning analytics model, a systematic approach is applied to collect and process data from a variety of sources to gain insight into student behavior, performance, and engagement. The process is divided into various steps which include data collection, preprocessing, data analysis, model selection, training and testing and at last predicting the final result. This section contains an overview of the steps required to create a learning analytics model is described below using the figure:

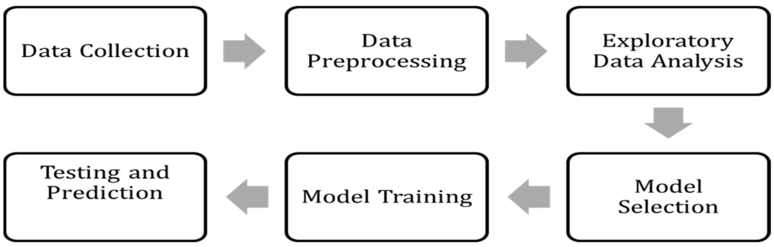

Figure 1. Learning Analytics Proposed Model

A. Data Collection

Data collection is defined as the process of gathering, storing, and analyzing data related to learners and their interaction within educational settings. While collecting data ethical and privacy concerns should be considered. This data is collected from various sources which includes:

1. Learning Management Systems

LMS include Moodle, Blackboard or Canvas which collect data in the form of student's attendance, performance scores, engagement levels etc. All this data can be used to gain valuable insights in students' performance.

2. Clickstream Data

It is defined as the sequence of clicks and interactions a student made while interacting with educational environments. It captures data including the amount of time spent in reading any article, scrolling data, and logon logoff information.

3. Social Network Analysis

Social Network is a key source of data collection. Social Network Analysis examines the interactions between learners, teachers, and learning platforms.

B. Data Preprocessing

It is the next step in which we prepare the data for analysis. When data is collected initially it is in raw form which is not suitable for analysis. Therefore, we need to perform some preprocessing tasks which include:

1. Data Cleaning

It is the process of fixing incorrect data, finding and correcting incorrect information, deleting inaccurate information, and making data clean and correct.

2. Data Transformation

It is the process of transforming data from one form to another depending upon the requirement of any model. The transformation can be normalization, standardization, logarithmic transformation, or power transformation to achieve a more symmetric distribution.

3. Data Integration

It can be understood as the process of combining data from various sources into a single platform for better analysis and visualization.

C. Exploratory Data Analysis

After preprocessing, Exploratory data analysis is performed which includes analyzing and summarizing data to understand it. It involves visualizing and exploring data using various statistical and visualization tools. Various techniques used for EDA are:

1. Correlation Analysis

It examines how different variable depend on each other thus helping in understanding the data more precisely. Correlation gives us an idea about how different variables depend on each other which helps in understanding patterns easily.

2. Missing Data Analysis

Determining missing values and finding a way to fill these missing values may affect the performance of a model in a more precise way. The missing values can be filled in different ways by using mean, median or random values.

3. Outlier Detection

Outlier detection is one of the important techniques in any model. Outliers can affect the basic computation of any model which can affect the performance of a model in a negative way and change the desired results.

4. Handling Imbalanced Datasets

Sometimes the data is Imbalanced means some of the classes have more effect on results then some other. Handling this issue is very important and can be done in various ways like oversampling, under sampling or using techniques like SMOTE. SMOTE is used in [5] [10] for handling imbalanced classes.

5. Feature Selection

Feature selection is one of the most important techniques of data analysis. It helps in finding the best possible results. Feature selection includes selecting the important features and deleting the redundant or irrelevant variables.

D. Model Selection

Selecting a model is an important task as the performance and accuracy of a system depends on the model. Selection process aims in identifying the algorithm which best fits the data and helps in obtaining the desired outcome. While selecting any model we need to consider aspects like size of data. Model complexity, domain of variables, and evaluation matrixes. Learning Analytics involves various Machine Learning, Deep Learning and statistical algorithms. Algorithms like Random Forest, Logistic Regression, Support Vector Machine, Artificial Neural Network and K Nearest neighbor are the most accurate and already applied algorithms used for predicting students' performance, finding students at risk and managing retention and dropout rates. Model Selection is an iterative process, we can compare the performance of different algorithms and then select the best one which gives the most accurate results. We can combine various algorithms into a single model using Ensemble Techniques. There are different Ensemble Methods like Stacking, voting, Bagging and Boosting by which we can combine various algorithms and improve the final accuracy of the model.

E. Model Training

After performing the data cleaning, preprocessing, and analysis steps the data is divided into two halves as training dataset and testing data set. The selected models are trained by using the training data sets. The data is fed to the models and then they learn the pattern and relationships between different modules of datasets. Hyperparameter tuning of the variables is done before training by the users. Tuning hyperparameters impacts the performance of any model to a great extent. Most Common ways to do this are grid search and random search.

F. Testing and Prediction

Once the model is trained, its validation is tested using the test data sets. The test dataset is provided to the trained model and predicted outcomes are compared with the actual sets. Evaluation of the model is done using different performance metrics like precision, accuracy, F-score, confusion matrix, recall and Receiver Operating Characteristics curve (ROC-AUC) etc. are used. Along with performance metrics various different validation techniques are also used to check the performance of the model. Cross Validation helps in estimating the performance of models across different subsets of data. Also, Holdout Validation splits the data into three parts as training, testing and validation sets where the validation sets are used for tuning the hyperparameters. Training and evaluation are also an iterative process, if performance was not found satisfactory then we may need to look back to the preprocessing steps or data splitting steps to improve the performance.

IV. CONCLUSION AND FUTURE SCOPE

Learning Analytics plays an important role in identifying students' performance in online learning platforms. This paper provides an overview of learning analytics, its applications, and methods used to perform students' performance and identify students at risk of failure or drop-out. Further, it provides a detailed description of previous studies highlighting the use of learning analytics in the educational field. A learning analytics model is also proposed which includes the basic steps in building any model.

It discusses the use of various predictive modeling techniques like regression, classification, and clustering in different studies which help in predicting students' performance in different environments. With the advantages of using learning Analytics, there are certain problems that need to be considered while making a new model. Privacy concerns and ethical considerations should be taken into account while collecting the data. Handling missing values and imbalance datasets is also an important issue that should be taken care of. Further, the integration of Learning Analytics with educational environments helps in timely intervention thus improving students' performance.

In the future, more advancement of learning analytics should be done to improve the performance score of learning environments. By early prediction of performance, the risk of dropout can be decreased. Learning analytics uses various machine learning, deep learning, and statistical methods to predict student performances. The prediction models can be embedded within the learning environments so that students and teachers both can get benefit by reviewing their performance.

REFERENCES

[1] Almarabeh, Hilal, "Analysis of Students' Performance by Using Different Data Mining Classifiers", presented at International Journal of Modern Education and Computer Science. Sep. 9-15, 2017. [Online].Available: https://www.researchgate.net/publication/318959370_Analysis_of_Students'_Performance_by_Using_Different_Data_Mining_Classifiers

[2] Kučak, Danijel & Juricic, Vedran & Đambić, Goran, "Machine Learning in Education - a Survey of Current Research Trends" 10.2507/29th.daaam.proceedings.059. [Online] Available: https://www.researchgate.net/publication/329463628_Machine_Learning_in_Education_-_a_Survey_of_Current_Research_Trends

[3] Ali Salah Hashim et al, "Student Performance Prediction Model based on Supervised Machine Learning Algorithm", IOP Conf. Ser.: Mater. Sci. Eng, 2020, 928 032019, doi:10.1088/1757-899X/928/3/032019.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Madnaik, Sandeep Subhash, "Predicting Students' Performance by Learning Analytics" Master's Projects, (2020) 941. https://scholarworks.sjsu.edu/etd_projects/941,doi: https://doi.org/10.31979/etd.6jjb-ua9w.

[5] S. D. A. Bujang et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," in IEEE Access, vol. 9, pp. 95608-95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[6] Harikumar Pallathadka, Alex Wenda, Edwin Ramirez-Asís, Maximiliano Asís-López, Judith Flores-Albornoz, Khongdet Phasinam, "Classification and prediction of student performance data using various machine learning algorithms", Materials Today: Proceedings, Volume 80, Part 3,2023,pp: 3782-3785, https://www.sciencedirect.com/science/article/pii/S221478532105241X, doi: https://doi.org/10.1016/j.matpr.2021.07.382.

[7] M. Adnan et al., "Predicting At-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," in IEEE Access, vol. 9, pp. 7519- 7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[8] D. A. Shafiq, M. Marjani, R. A. A. Habeeb and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 72480-72503, 2022, doi: 10.1109/ACCESS.2022.3188767.
[9] G. Feng, M. Fan and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," in IEEE Access, vol. 10, pp. 19558-19571, 2022, doi: 10.1109/ACCESS.2022.3151652.
[10] K. M. Hasib, F. Rahman, R. Hasnat and M. G. R. Alam, "A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0399-0405, doi: 10.1109/CCWC54503.2022.9720806.
[11] Alsariera, Yazan A., et al, "Assessment and evaluation of different machine learning algorithms for predicting student performance." Computational Intelligence and Neuroscience 2022 (2022).
[12] R. Z. Pek, S. T. Özyer, T. Elhage, T. ÖZYER and R. Alhajj, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure," in IEEE Access, vol. 11, pp. 1224- 1243, 2023, doi: 10.1109/ACCESS.2022.3232984.
[13] S. D. Abdul Bujang et al., "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review," in IEEE Access, vol. 11, pp. 1970-1989, 2023, doi: 10.1109/ACCESS.2022.3225404.
[14] Ouyang, Fan, et al. "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course.", International Journal of Educational Technology in Higher Education 20.1 (2023), pp: 1-23.
[15] B. P. P and C. Kurian, "Student performance prediction in e-learning system and evaluating effectiveness of online courses," 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS), Kochi, India, 2023, pp. 1-5, doi: 10.1109/AICAPS57044.2023.10074504.
[16] Clow, Doug, "An overview of learning analytics. Teaching in Higher Education", 18(6), (2013), pp. 683–695, http://www.tandfonline.com/doi/full/10.1080/13562517.2013.827653#.UhJALGTdLZs.doi:http://dx.doi.org/doi:10.1080/13562517.2013.827653.
[17] Ferguson, Rebecca, "Learning analytics: drivers, developments and challenges". International Journal of Technology Enhanced Learning, 4(5/6),(2013) pp. 304–317, http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijtel doi: http://dx.doi.org/doi:10.1504/IJTEL.2012.051816.
[18] Chatti, Mohamed Amine, et al, "A reference model for learning analytics." International Journal of Technology Enhanced Learning, 4.5-6 (2012) pp: 318-331.
[19] N. Iqbal, F. Jamil, S. Ahmad and D. Kim, "Toward Effective Planning and Management Using Predictive Analytics Based on Rental Book Data of Academic Libraries," in IEEE Access, vol. 8, pp. 81978-81996, 2020, doi: 10.1109/ACCESS.2020.2990765.
[20] Albreiki, Balqis, Nazar Zaki, and Hany Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," Education Sciences 11.9, (2021): 552.
[21] M. Srilekshmi, S. Sindhumol, S. Chatterjee and K. Bijlani, "Learning Analytics to Identify Students At-risk in MOOCs," 2016 IEEE Eighth International Conference on Technology for Education (T4E), Mumbai, India, 2016, pp. 194-199, doi: 10.1109/T4E.2016.048.