International Journal of Applied Engineering & Technology

DATA DRIVEN APPROACH FOR ANALYSING LEARNER BEHAVIOUR IN ONLINE COURSES USING MACHINE LEARNING TECHNIQUES

Pradnya Mhatre[1] and Dr. Kelapati[2]
[1]Research Scholar and [2]Assistant Professor, Department of Computer Science, Shri JJT University, Rajasthan India

ABSTRACT

In this context, leveraging machine learning techniques for learner behaviour analysis emerges as a promising approach. This paper presents a thorough review of the data-driven methodologies used in analysing learner behaviour in online courses using machine learning techniques. The introduction of online courses has revolutionized the educational landscape by enabling students all throughout the world to access excellent content from the comfort of their homes. Nevertheless, ensuring effective Learning objectives in virtual settings requires a deep understanding of learner behaviour and engagement patterns. In order to extract meaningful insights from learner data, the paper first examines the various dimensions of learner behaviour, such as engagement, interaction patterns, performance, and learning styles. Based on these insights, the paper then applies machine learning algorithms, such as classification, clustering, and sequence mining, to enable the development of personalized learning experiences, adaptive interventions, and targeted support mechanisms that are suited to the needs of individual learners. The study also addresses the difficulties in interpreting machine learning models, scalability, and data privacy when studying learner behaviour in online courses. In addition, it emphasizes current developments as well as potential paths forward in this area, such as the incorporation of multimodal data sources and the application of deep learning methods for more in-depth analysis. Overall, this research highlights the potential of methods for machine learning to improve the effectiveness and efficiency of online education and highlights the significance of a data-driven strategy for understanding learner behaviour in online courses.

Keywords: *classification, clustering, data driven approaches, data privacy, machine learning, online education*

1. INTRODUCTION

A new age of education has begun with the rise of online education, which provides students all around the world with previously unheard-of accessibility and flexibility. However, maintaining the efficacy of online courses necessitates a sophisticated comprehension of learner engagement and behaviour patterns. The intricacy and dynamics of online learning environments are frequently beyond the scope of traditional approaches of learner behaviour analysis. As a result, incorporating machine learning techniques offers a viable way to mine the enormous amounts of information generated by online courses for insightful information.

A branch of artificial intelligence called machine learning provides effective methods for spotting hidden links and patterns in massive datasets. Machine learning makes it possible to automatically analyse learner behaviour through the application of an algorithm derived from data. This results in more individualized and flexible learning environments. By using machine learning approaches, educators and instructional designers can improve the efficacy of online courses by gaining deeper insights into learner preferences, performance trajectories, and engagement determinants.

In this paper, we investigate the use of ML techniques to analyse learner behaviour in online courses through a data-driven approach. We explore all the facets of learner behaviour, from engagement levels and interaction patterns to performance indicators and learning styles. We seek to clarify the possible uses, difficulties, and opportunities associated with using machine learning to analyse student behaviour in online education by synthesizing existing research and approaches.

By this investigation, we hope to contribute to the developing field of research on the subject of online learning and machine learning, highlighting the revolutionary potential of data-driven methods for reshaping the educational landscape. We can create more efficient, individualized, and interesting online learning experiences by utilizing machine learning to extract useful insights from learner data.

Online education is becoming increasingly important due to a number of factors:

A. **Accessibility:** Distance learning breaks down geographical boundaries and gives students from all backgrounds and places access to excellent educational materials. For those who might encounter obstacles in their pursuit of traditional forms of education, such as time, distance, or mobility issues, this accessibility is very helpful.

B. **Flexibility:** Learners can customize their online courses to suit their schedules and tastes, since they provide an unmatched level of flexibility. Online learning gives students the freedom to interact with course materials at their own step and convenience, whether they are juggling job, family obligations, or other obligations.

C. **Cost-effectiveness:** In contrast to conventional physical schools, online learning frequently removes the need for costly infrastructure and resources. Because of its affordability, education is now more accessible and inexpensive for a wider range of people, including those who might not otherwise be able to attend traditional universities.

D. **Technological developments:** The quality and interactivity of online learning experiences have been considerably improved by the rapid advances in technology, notably in the areas of digital communication, multimedia material delivery, and online collaboration tools. With the aid of these technology advancements, teachers can design immersive, captivating learning environments that are on par with traditional classroom settings.

E. **Lifelong learning:** Upskilling and constant learning are essential in the fast-paced, dynamic world of today. Lifelong learning opportunities are provided by online education, allowing people to update their knowledge, pick up new skills, and remain employable in a constantly changing labour market. [1][5][10]

2. Related Work

Research and interest in the ground of machine learning based learner behavior analysis in online courses are growing. Predicting student outcomes—like grades or completion rates—based on behavioral data is a common subject of research. Neural networks, decision trees, and logistic regression are few of the methods. Muhammad Zine in 2023 [1] surveyed Tlemcen University's Economics faculty members and students in order to collect data based on the five components of the ADKAR model: awareness, desire, knowledge, ability, and reinforcement. They carried out a correlation study, which showed a strong correlation between each dimension. The readiness and awareness, desire, knowledge, ability, and reinforcement pairwise correlation coefficients are 0.5233, 0.5983, 0.6374, 0.6645, and 0.3693. In order to decide which ADKAR criteria were most crucial in determining e-learning readiness, two machine learning methods were utilized: Random Forest (RF) and Decision Tree (DT). Ability and knowledge consistently found to be the most important components in the outcomes. Two sets of empirical experiments were conducted by Raghad Al-Shabandar et al. (2018) [8] to analyze learner performance using statistics and ML techniques. Successful and failing students were compared using a descriptive statistical test; the findings showed a substantial difference in the two groups' levels of involvement. Early in the course, machine learning methods like decision trees, neural networks, and regularized discriminant analysis are utilized to automatically identify students who lack motivation. This gives instructors information about student withdrawal. Farrukh Saleem and colleagues (2021) [3] Recognize and forecast student performance using characteristics taken from electronic learning management systems. Five conventional machine learning algorithms make up the suggested model, which is further improved by using four ensemble techniques: bagging, boosting, stacking, and voting. A total of DT (0.675), RF (0.777), GBT (0.714), NB (0.654), and KNN (0.664) are the individual models' F1 scores. When employing ensemble techniques, the model's performance has significantly improved. The behavior classification-based e-learning performance prediction framework is proposed by Feiyue Qiu et al. (2022) [2]. This framework chooses the characteristics of e-learning behaviors, combines feature fusion with behavior data in accordance with the behavior classification model to obtain the category feature values of each type of behavior, and then constructs a machine learning-based learning performance predictor. A fresh approach has been taken in creating an e-learning performance predictor, which offers a fresh way to assess e-learning classification techniques.An overview of the use of machine learning methods in education science research is given in this article, as mentioned by Nguyen Thi Kim Son et al. (2021) [4]. The research methodology shows that the use of technology in learning and teaching, information gathering, analysis, and processing to create highly accurate answers or guidance in resolving educational difficulties, is the trend and strength in education science research.

3. The Requirement for Efficient Techniques To Examine Student Behaviour

While there are many advantages to online learning, comprehending and evaluating learner behaviour is crucial to guaranteeing the program's efficacy. When educators and instructional designers analyse learner behaviour well, they can do the following:

Determine engagement patterns: Teachers can understand how students interact with the course materials, take part in conversations, and participate in learning activities by examining the behaviour of the learners. By being aware of these trends, teachers can modify their approaches and interventions to improve the motivation and involvement of their students.[12]

3.1Customize Learning Experiences:

Learner behavior analysis makes it easier to create individualized learning programs that are appropriate for each student's preferences, learning preferences, and degree of proficiency. Teachers can modify the content, pacing, and teaching strategies of their courses to accommodate the varied requirements of their students by utilizing the insights obtained from learner data.

3.2 Forecast the Performance of Students and Identify Students Who May be at Risk of Academic Failure or Disengagement:

Learner data combined with machine learning techniques can be utilized to estimate student achievement and identify students who may be at risk of these outcomes. In order to improve student outcomes, educators can intervene proactively and offer tailored help to students who are identified as being at-risk early on.

4. Challenges Associated with Analysing Learner Behaviour in Online Courses

There are several obstacles to overcome when analysing learner behaviour in online courses, from the technological difficulties of managing heterogeneous data to the moral dilemmas associated with privacy. Let's examine each of these difficulties in turn:

4.1 Heterogeneity of Data:

Diversity of Data Sources: A variety of data types are produced by online courses, such as clickstream data, assessment outcomes, forum activities, and more. It can be challenging to incorporate combining these various data sources into a logical investigation.

4.2 Data Formats: Information can be found in a diversity of forms, including unstructured text, semi-structured logs, and organized databases. Different preprocessing and analysis approaches are needed for each format.

4.3 Scalability: Huge Data Volumes: An exponential rise in the number of participants in online courses results in an exponential growth in the amount of data created. In order to deliver timely insights, scalable algorithms and infrastructure are needed for the analysis of such big datasets.

4.4 Real-Time Analysis: In order to respond to student behaviour in a timely manner, educators occasionally need real-time feedback. Complexity is increased when real-time data stream processing is introduced into systems.[4][6][11]

5. Machine Learning Algorithms Commonly Used in Analysing Learner Behaviour

5.1 Classification:

Defining the category or class label of fresh observations based on data from the past is the aim of the supervised learning task of classification.

a) Utilization in the Analysis of Learner Conduct:

Classification algorithms can be used to predict a range of learner behaviour-related outcomes, including the identification of students who are at-risk, the classification of engagement levels, and the classification of learning styles. Decision trees, random forests, support vector machines (SVM), logistic regression, and naive bayes are a few examples of algorithms.

b) Advantages:

Predictive Accuracy: Classification models are useful for tasks like predicting student outcomes or identifying at-risk pupils because they can achieve high predictive accuracy when trained on well-structured and labelled data.

Flexibility: A wide range of applications in the analysis of learner behaviour are made possible by the ability of classification algorithms to handle binary and multiclass classification problems.

c) Constraints:

Unbalanced Data: Unbalanced datasets, in which one class is noticeably more common than the others, might result in models that are skewed in favour of the dominant class. To solve this problem, methods like resampling or changing class weights can be required.

Overfitting: Intricate classification models may overfit to anomalies or noise in the data, which impairs their ability to generalize to new data. To reduce overfitting, regularization strategies and cautious model selection are crucial. [2][7]

5.2 Clustering:

A definition of clustering is the process of grouping comparable data points together according to their attributes through unsupervised learning.

a) Utilization in the Analysis of Learner Conduct:

Algorithms for clustering can be used to find trends and subpopulations of learners. Clustering, for instance, can be used to find student groups with comparable behavioural patterns or learning preferences. Gaussian Mixture Models (GMM), DBSCAN, K-Means, and Hierarchical Clustering are a few examples of algorithms.

b) Advantages:

Pattern Discovery: By grouping students based on shared learning preferences or behavioral tendencies, clustering algorithms can reveal latent patterns and structures within the student body.

Unsupervised Learning: Clustering can be used for investigative research and extracting insights from unlabeled datasets since it doesn't require labeled data.

Scalability: A lot of clustering methods can handle big datasets, making it possible to analyze student behavior in massively parallel online courses in an effective manner.

c) Constraints:

Subjectivity: The number of clusters, initialization settings, and distance metric selected can all have an impact on the clustering findings, allowing for varying interpretations.

Interpreting clusters: Determining the significance of a cluster can be difficult, especially in high-dimensional environments where visual examination is not practical. Interpreting clustering results often requires domain expertise.

Cluster Validity: Since there is no one-size-fits-all way to quantify cluster validity, assessing the caliber of clustering results can be challenging. For the same dataset, different clustering techniques may yield different outcomes. [13][14]

5.3 Regression:

Predicting a continuous numerical value from input features is the purpose of this supervised learning activity.

a) **Utilization in the Analysis of Learner Conduct:**
Regression algorithms are useful for forecasting a range of continuous outcomes that are associated with the behavior of learners. These outcomes include time spent on activities, completion rates, and performance scores. Algorithm examples include Gradient Boosted Regression Trees (GBRT), Support Vector Regression (SVR), Lasso Regression, Ridge Regression, and Linear Regression.

b) **Advantages:**

Quantitative Prediction: By providing quantitative forecasts of continuous outcomes, regression models enable educators to project a range of learner behavior-related indicators, including completion rates and performance scores.

Variable Importance: By determining the proportionate weight of input characteristics in forecasting the target variable, regression models can shed light on the variables influencing student behavior.

Interpretability of the Model: Teachers can more easily comprehend the links between input factors and results by using simple regression models, such linear regression, which provide clear interpretations of the coefficients.

c) **Constraints:**

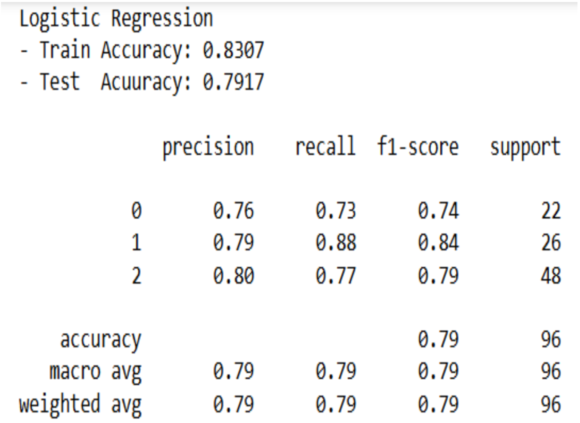
Supposition Infraction: The linearity, independence, and normalcy of residuals assumptions are fundamental to parametric regression models, including linear regression. When these presumptions are broken, prejudiced or untrustworthy outcomes may follow.

Overfitting: When there are more characteristics than there are samples in a complex regression model, like polynomial regression or ensemble approaches, the model may overfit to the training set.

Limited Flexibility: Feature engineering or the usage of more complicated models are necessary when using certain regression models since they are not as flexible in capturing nonlinear relationships or interactions between variables.[15]

6. Different Classifier Models

In order to determine the accuracy of different machine learning methods, Student Academic Performance data is considered as an input to conduct the experimentation. Training accuracy and Test accuracy are evaluated for different classifier models. For experimental evaluation, Student Academic Performance Dataset from the UCI Machine Learning Repository is considered as input. The data set URL is given as https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data/data. The data set comprises the student information such as Gender, Nationality, GradeID, Semester, raisedhands etc. By using above dataset there is creation of model and analysis is performed on different classifier model.

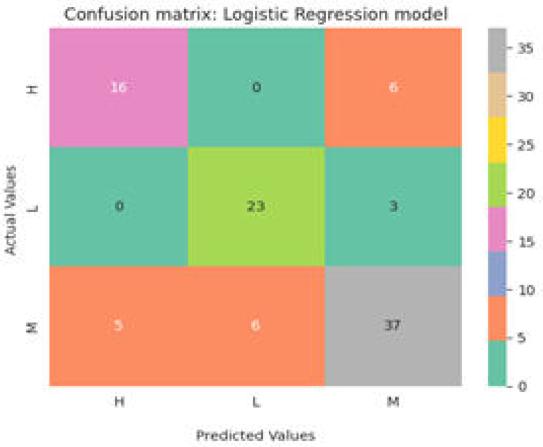


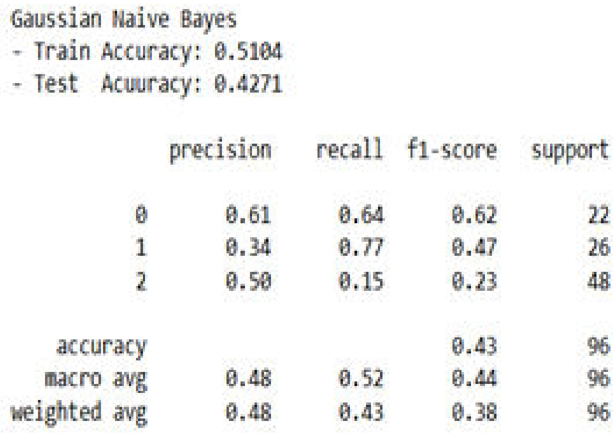
```
Logistic Regression
- Train Accuracy: 0.8307
- Test  Acuuracy: 0.7917

              precision    recall  f1-score   support

           0       0.76      0.73      0.74        22
           1       0.79      0.88      0.84        26
           2       0.80      0.77      0.79        48

    accuracy                           0.79        96
   macro avg       0.79      0.79      0.79        96
weighted avg       0.79      0.79      0.79        96
```



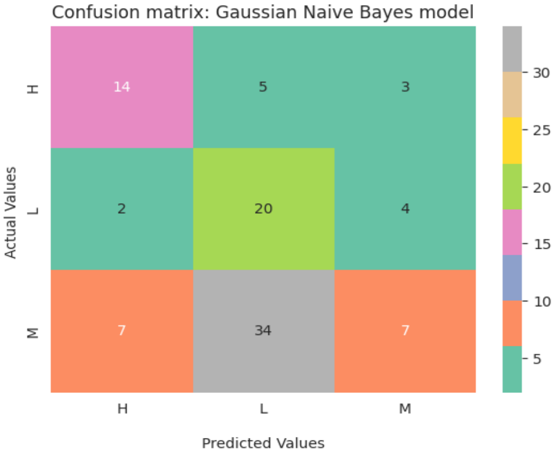


```
Gaussian Naive Bayes
- Train Accuracy: 0.5104
- Test  Acuuracy: 0.4271

              precision    recall  f1-score   support

           0       0.61      0.64      0.62        22
           1       0.34      0.77      0.47        26
           2       0.50      0.15      0.23        48

    accuracy                           0.43        96
   macro avg       0.48      0.52      0.44        96
weighted avg       0.48      0.43      0.38        96
```



The quantitative analysis is compared with different classifier model. As a result of applying the different classifier, we obtain following table,

Table1: Tabulation for analysis of different classifier model

Classifier Model	Train Accuracy	Test Accuracy
Logistic Regression	0.8307	0.7917
Gaussian Naïve Bayes	0.5104	0.4271
SVM	0.8307	0.7917
Random Forest	1.0000	0.8750
KNN	0.7630	0.6458
XGBoost	1.0000	0.8438

7. Identification of challenges associated with employing machine learning techniques for analysing learner behaviour

There are a number of difficulties when using ML algorithms to analyze learner behavior, such as:

7.1 Data Quality:

Due to a number of issues, including missing data, noise, and inconsistency, it might be difficult to guarantee the accuracy and comprehensiveness of data gathered from online learning platforms.

7.2 Data Preprocessing:

To preserve data quality, meticulous attention must be paid to cleaning and preprocessing the data to eliminate outliers, handle missing values, and standardize formats.

Data bias: Unfair results for some learner groups and distorted analytic results can occur from biases in the data, such as sampling or demographic biases.

7.3 Ethical Considerations:

1) **Privacy:** Handling sensitive data, like learning records and personal information, when analysing learner behaviour raises privacy issues. To protect learner privacy, it's critical to put strong data protection procedures in place and adhere to pertinentrules.[9]

2) **Fairness:** Certain learner groups may be treated unfairly by machine learning models because these algorithms may unintentionally encode biases found in the data. Important ethical considerations are minimizing algorithmic bias and guaranteeing fairness in model predictions.

3) **Scalability:** Large Datasets is a significant obstacle when analysing learner behaviour in online courses, which frequently involve processing huge amount of data generated by multiple students interacting with different learning resources.[3]

8. Emerging Trends and Future Directions

8.1 Deep Learning Techniques:

Convolutional neural networks and recurrent neural networks are two examples of deep learning approaches that are growing in popularity in educational data mining. These methods work well with structured and sequential data, which makes them appropriate for examining temporal trends in learner interactions. Future studies could investigate how deep learning models can be applied to tasks like learner engagement dynamics prediction and subtle behavioral cue detection that indicate learning progress.

8.2 Explainable AI in Education:

Ensuring the interpretability and openness of ML models is of utmost importance as they grow more complicated, especially in educational contexts where stakeholders need actionable information. The aim of current research is to create explainable AI methods that offer clear justifications for model predictions and suggestions. Teachers can better comprehend the reasoning behind algorithmic decisions and provide knowledgeable interventions to benefit students by improving the interpretability of ML models.[3]

9. CONCLUSION:

To sum up, data-driven strategies that make use of machine learning techniques have a great deal of promise to improve our comprehension of how learners behave in online courses. By examining a diversity of data sources, such as clickstream data, assessment outcomes, and forum exchanges, researchers have been able to forecast learner performance, identify engagement patterns, and tailor learning experiences. But issues like protecting data privacy, reducing algorithmic bias, and improving model interpretability still need to be taken into account.

In the future, there will be plenty of opportunity to explore new trends like the creation of explainable AI frameworks, the use of deep learning techniques, and the integration of multimodal data sources. Realizing the full potential of data-driven approaches in online education will also require a sustained focus on ethical and responsible AI practices, longitudinal studies, and adaptive learning environments.

In the end, researchers, educators, and legislators may leverage the potential of data-driven approaches to improve the caliber, inclusivity, and accessibility of online learning experiences for students throughout the world by embracing these trends and working together to address difficulties.

REFERENCES

[1] Zine, M., Harrou, F., Terbeche, M., Bellahcene, M., Dairi, A., Sun, Y., (2023), E-Learning Readiness Assessment Using Machine Learning Methods. Sustainability 2023, 15, 8924.

[2] Feiyue Qiu, et al., (2022), Predicting students' performance in e-learning using learning process and behaviour data, Scientific Reports 12, Article No.453.

[3] Saleem, F., Ullah, Z., Fakieh, B., Kateb, F., (2021), Intelligent Decision Support System for Predicting Student's E-Learning Performance Using Ensemble Machine Learning, Mathematics 2021, vol 9, issue 17

[4] Nguyen Thi Kim Son, et al., (2021), The Applications of Machine Learning in Education Science Research, VNU Journal of Science: Education Research, Vol. 37, No. 4 (2021) 19-26

[5] The meaning of e-Education in India, https://talentedge.com/articles/meaning-e-education/ Accessed on: 06 February, 2020.

[6] Tara Rawat, Dr. Vineeta Khemchandani, (2019), Feature Engineering (FE) Tools and Techniques for Better Classification Performance, International Journal of Innovations in Engineering and Technology (IJIET),82.024.

[7] Abdallah Moubayed, et al., (2018), E- Learning:Challenges and Research Opportunities Using Machine Learning & Data Analytics, IEEE Access, pp 99:1-1

[8] Raghad al-Shabandar, et al., (2018), Analyzing Learners Behavior in MOOCs: An Examination of Performance and Motivation Using a Data-Driven Approach, IEEE Access, pp 99:1-1

[9] Vasan, Nirmal, et al. "An overview of e-learning in higher education using machine learning." International Journal of Information Management 52 (2020): 102034.

[10] Fareeha Rasheed, Abdul Wahid, (2021), Learning Style detection in E-learning system using machine learning techniques, Expert Systems with Applications, Vol 174

[11] S. Agarwal, G. Pandey, and M. Tiwari, (2012), Data Mining in Education: Data Classification and Decision Tree Approach, In International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2.

[12] Minaei-Bidgoli, Behrouz, et al. "Machine learning techniques in e-learning: principles and applications." World Applied Sciences Journal 9.7 (2010): 823-828.

[13] Aggarwal, Charu C., et al. "A survey of privacy protection techniques in e-learning systems." Journal of Computer Science and Technology 26.5 (2011): 787-798.

[14] Papanikolaou, Kostas A., et al. "Educational data mining: A review of the state of the art." IEEE Transactions on Learning Technologies 11.1 (2017):4-19.

[15] A. Ahadi and R. Lister, Arto Vihavainen (2016), On the Number of Attempts Students Made on Some Online Programming Exercises During the Semester and their Subsequent Performance on Final Exam Questions, In ITiCSE proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science