Big Data Analytics in Online Education: Learning Behaviour and Dropout Prediction

Tushar Sinha
SVKM'S NMIMS
School of Technology Management and Engineering
Navi Mumbai, India
sinhatushar12@gmail.com

Veer Javadia
SVKM's NMIMS
School of Technology Management and Engineering
Navi Mumbai, India
Veerjavadia@gmail.com

Satyam Shukla
SVKM'S NMIMS
School of Technology Management and Engineering
Navi Mumbai, India
satyamshukla9791@gmail.com

Anindita Khade
SVKM's NMIMS
School of Technology Management and Engineering
Navi Mumbai, Maharashtra
anindita.khade@nmims.edu

***Abstract*—The rapid growth of online education, especially through Massive Open Online Courses (MOOCs) and virtual learning platforms, has led to an enormous amount of learner data. Big Data Analytics (BDA) has become a key method for making use of this information to improve teaching, personalize learning experiences, and tackle the ongoing problem of high dropout rates. This review looks at how BDA is used to analyze learning behavior and predict student dropout, based on recent developments in Educational Data Mining (EDM) and Learning Analytics (LA). We discuss various analytical methods, including statistical techniques such as correlation and regression. We also cover machine learning models such as Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Decision Trees (GBDT). Additionally, we address deep learning approaches like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks for modeling sequential behavior. These methods have shown strong predictive ability, with models like GBDT reaching up to 88% accuracy in predicting dropouts. Key applications include profiling student engagement, identifying students at risk, and enabling timely teaching interventions. However, there are still challenges related to data quality, understanding models, scaling, and ethical issues concerning privacy and bias. Future efforts should focus on combining different types of data, creating adaptive real-time intervention systems, and applying privacy frameworks to ensure that online education analytics are scalable, clear, and fair.**

***Index Terms*—Big Data Analytics, Learning Analytics, Educational Data Mining, MOOCs, Dropout Prediction, Machine Learning, Deep Learning**

I. Introduction

During the past decade, online education has grown rapidly due to improvements in internet access, mobile technology, and digital learning platforms. Massive Open Online Courses (MOOCs) from providers like Coursera, EdX, Udacity, and FutureLearn have changed the way people access higher education. They offer free or low-cost courses to a global audience. The popularity of MOOCs is evident in the numbers. Millions of learners from various backgrounds enroll each year, looking for flexible, self-paced, and often worthwhile learning opportunities [1].

However, online education faces a major issue: low course completion rates. The enrollment numbers are high, but the percentage of learners who finish their courses is much lower, often below 10 to 15% in large-scale studies. Several reasons contribute to this problem, such as lack of motivation, insufficient interaction with instructors and peers, poor feedback systems, and conflicting personal or work obligations. It is crucial to identify and address these dropout risks to ensure that online education meets its goal of providing fair and high-quality learning opportunities [2]–[4].

Big Data Analytics (BDA) presents a strong solution to this problem. Modern online learning platforms create large amounts of structured and unstructured data. This data includes clickstream logs, video viewing habits, forum interactions, quiz results, and student demographics. Analyzing these data streams with advanced computational methods allows researchers and educators to find patterns in student engagement, understand learning behaviors, and predict dropout risks. Predictive models, whether they use machine learning algorithms such as support vector machines (SVM), random forests (RF), and gradient-boosting decision trees (GBDT), or deep learning techniques like long- and short-term memory networks (LSTM), have shown high precision in identifying at-risk students [3], [5], [6].

II. Related Work

The application of big data analytics and machine learning in education has gained a lot of attention in recent years.

Researchers have concentrated on predicting student performance, early warning signs of dropout, and supporting teachers with data-driven decisions. Several methods have been designed that combine student background, demographic data, and interaction data into predictive models to offer informative recommendations for teachers. [7]

A significant area of research has been devoted to designing dashboards and visualization interfaces that enable teachers to track student progress and respond in a timely manner. These interfaces are likely to integrate structured educational records and unstructured information, such as discussion boards and text content to offer a richer perspective of student activity. Distributed computing technologies such as Apache Spark and natural language processing have been used to analyze large data sets efficiently. Random Forest and ensemble learning techniques have been shown to improve accuracy in the prediction of dropout risk.

Another key contribution has been creating blended learning models that are customized for regional education systems. For example, some models blend face-to-face instruction with online materials, using big data analysis to guide teaching practices and encourage student participation. Such models illustrate how predictive analytics can be incorporated into the learning process itself, even in low-resource environments.

Benchmark data sets have also been key in spurring research in the field. The Open University Learning Analytics Dataset (OULAD) has facilitated fine-grained traces of learner activity and assessment data that have allowed researchers to experiment with predictive models under practical conditions. Similarly, other open platforms have aggregated data sets utilized to compare machine learning classifiers such as logistic regression, decision trees, and gradient boosting models to predict dropout. [8]

Deep learning methods have increased their importance in recent years. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models have been used to recognize student behavior patterns over time, leading to greater prediction accuracy for dropouts at an early point. Scalable learning pipelines utilizing cloud analytics coupled with ensemble techniques have also displayed very high prediction accuracy, which reflects the strength of using classical machine learning together with modern big data platforms. [3]

In general, current research stresses three aspects: the integration of structured and unstructured data sources, the use of scalable frameworks for big data, and the employing of understandable and complex sequential models. The outcomes lay a sound foundation for further study and also point toward challenges in ensuring data quality, scalability, and ethical use of predictive learning systems. [9]

III. Big Data Analytics in Education

A. Definition and Characteristics

Big data education is used to describe vast and intricate sets of data generated by online learning environments such as MOOCs, Learning Management Systems (LMSs), and other online platforms. Its key characteristics are generally characterized by the 3Vs model: Volume, which refers to enormous levels of data from millions of students; Velocity, which describes rapid creation and processing of data in close-to-real time; Variety, which encompasses varied forms of data ranging from numerical scores to text forum postings and multimedia exchanges. Researchers have broadened this over time to the 5Vs with the addition of Veracity, which means data quality and reliability, and Value, which means the actionable information one can derive. Some frameworks extend to 7Vs or more, such as Variability, which emphasizes that meaning changes over time, and Visualization, which means methods of interpreting complicated patterns. [7]

Educational data can be broadly classified as structured and unstructured. [10]

Structured data comprise log files, quiz scores, attendance records, and demographic data in relational databases.

Unstructured data contain discussion forum posts, video transcripts, assignment comments, audio-visual learning objects, and even eye-tracking or biometric data from immersive environments. The capacity to integrate and analyze both is required to provide a comprehensive picture of learner modeling. [11]

Cloud-based analytics emerged as the focus for scalability in the processing of educational big data. Cloud providers like AWS, Azure, and Google Cloud offer distributed storage solutions in the form of Amazon S3 and Hadoop Distributed File System, as well as parallel processing frameworks like Spark and MapReduce, which efficiently process terabytes of information. Cloud architectures support computational scalability as well as storage. They also enable flexible allocation of resources, making it economical to process data when there is a high demand for it, such as at the beginning or end of academic semesters. Furthermore, cloud analytics improve collaboration by allowing shared dashboards and predictive tools to be accessed in real time by various stakeholders. [12] [5]

B. Educational Data Mining and Learning Analytics

Educational Data Mining (EDM) is an area of research that specializes in discovering patterns and acquiring valuable knowledge from educational data. It uses computational techniques such as classification, clustering, association rule mining, and sequential pattern analysis to learn about student learning behavior, predict performance, and inform course design [6]. EDM tends to focus on improving algorithms to make precise predictions, for example, detecting students at risk of dropping out or detecting knowledge gaps in adaptive learning environments. [1]

Learning Analytics (LA), though parallel to EDM, is more concerned with quantifying, gathering, analyzing, and reporting student data to improve learning experiences and outcomes. It is not merely about algorithms; it encompasses teaching theory, visual analytics, and intervention design. LA frequently employs real-time monitoring dashboards for teachers, facilitating timely interventions like the sending of personalized reminders or extra materials [1].

The difference is nuanced but significant:

- **EDM** is primarily data-informed and tends to be grounded in computer science and artificial intelligence studies, with the focus of high prediction accuracy.
- **LA** is instructional-driven and tends to focus more on the ways in which the findings of the data can inform teaching activities and the support of the learner.

The stakeholders for both EDM and LA are the following:

- **Students** — are treated to targeted learning paths, adaptive feedback, and proactive warnings of impending performance problems.
- **Educators/Instructors** — leverage analytics to identify students at risk, enhance their own teaching techniques, and enhance student engagement.
- **Educational Managers/Administrators** — utilize insights to inform resource utilization, curriculum design, and tracking institutional performance.
- **Policymakers** — employ aggregated analytics to inform educational policy reforms, national skill development strategies, and online program quality assurance [1], [5].

C. Initial Data Exploration

To develop a foundational understanding of the dataset, we start by visualizing the target variable alongside the most pertinent features. Examining these distributions helps us uncover underlying patterns and relationships that might otherwise remain hidden. This initial exploration is essential for informing subsequent analysis and modeling decisions.

The dataset used for this analysis, `JEE_Dropout_After_Class_12.csv`, contains 5000 entries and 15 columns. These columns have a mix of numerical (5 float64, 2 int64) and categorical (8 object) data types. Below is an initial overview of the dataset structure:

- **Numerical Features:** `jee_main_score`, `jee_advanced_score`, `mock_test_score_avg`, `class_12_percent`, `daily_study_hours`, `attempt_count`, `dropout`.
- **Categorical Features:** `school_board`, `coaching_institute`, `family_income`, `parent_education`, `location_type`, `peer_pressure_level`, `mental_health_issues`, `admission_taken`.

It was observed that the `coaching_institute` column had some missing values, with 3791 non-null entries out of 5000. All other columns were complete. Data were divided into training and testing sets in an 80/20 ratio. This resulted in 4000 training samples and 1000 testing samples. Stratified sampling was used to keep the dropout ratio consistent between both two sets, with about 21% of the students in each set classified as dropouts.

Furthermore, preprocessing steps containing handling categorical variables through label encoding and normalizing numerical characteristics to confirm uniform scale. Outliers were carefully inspected and analyzed to preclude model bias. The missing values in the coaching column _institute were ascribed using mode substitution. This confirmed that the data set was clean, balanced, and ready for model training.

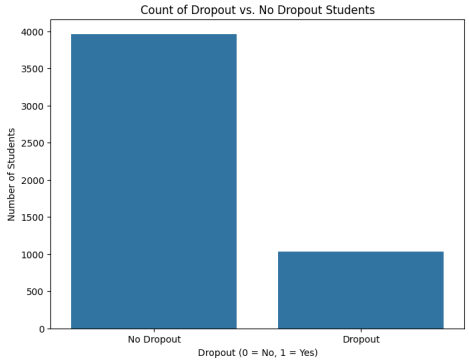

Fig. 1: Count of Dropout vs. No Dropout Students.

A glance at Figure 1 reveals a pronounced disparity in group sizes—most students remain enrolled, while only a small subset ends up dropping out. This imbalance sharply underscores a major challenge in dropout prediction: it's much harder, yet absolutely crucial, to accurately identify students who are at risk of leaving. If a predictive model primarily "gets it right" by just guessing most students will stay, its educational value drops off quickly. The real test lies in distinguishing true potential dropouts from the far larger pool of continuing students. In essence, a model's ability to isolate those at risk isn't just an academic exercise—it's vital for enabling worthwhile interventions.

IV. Learning Behavior Analysis

Learning behavior analysis in online education uses a variety of data sources to track how learners engage with digital platforms. Key categories include:

A. Data Sources

Clickstream Logs. These in depth reports include all of the user interactions which they have had with the system from page views to resource downloads, navigation actions and time spent in each activity. Clickstream data also enables us to do temporal analysis which in turn may reveal times of inactivity or what appear to be atypical learning behaviors that may in fact be signs of disengagement. [8]

Forum Interactions. This includes study of discussion boards, peer reply networks, and question answer platforms. We look at the size and timing of posts, sentiment of contributions, and social network metrics which we use to determine levels of collaboration. [13]

Video Engagement. Video analysis which reports on what time students begin and end videos, skip rate, replay frequency and watch time also is performed. We use these metrics to determine if students go through content in order or which hard topics they skip. [10]

Quiz and Assessment Results. Scores, timing of submissions, and pattern of attempts report which in turn give

us insight into knowledge acquisition and exam preparation behavior. [6]

Demographic Information. This includes what age, gender, location, educational background, and professional status are. [14]

Psychographic Information. This includes topics like what motivates students, what which self regulated learning styles they are at, what prior knowledge they bring to the course and what factors which drive their engagement often we collect this info via pre course surveys.

B. Feature Engineering

Feature extraction takes raw data and transforms it for use in predictive models. In learning behavior analysis, we also see:.

Enrollment Features. Information provided during registration—such as educational background, language proficiency, selected courses, and time zone—is collected.

User Activity Features. Metrics such as login frequency, session duration, activity on discussion forums, and the proportion of completed video content are often analyzed using log data to assess user engagement and behavior.

Course-level Features. Factors pertaining to the course itself include its level of difficulty, whether it's self-paced or led by an instructor, the organization of the content, and any expectations for student collaboration or peer interaction.

C. Pre-processing Challenges

- **Noise Filtering:** Clickstream logs often contain automated bot activity or accidental clicks.
- **Data Imbalance:** Dropout cases are often much higher than completion cases, leading to skewed datasets.
- **Missing Data:** Learners may not always provide demographic or psychographic details, creating gaps in feature sets.
- **Heterogeneity:** Data originates from a variety of sources—such as video logs, assessments, and forum discussions—necessitating a coherent process for integration. Consolidating these disparate formats into a unified structure remains a significant and ongoing challenge in academic research.

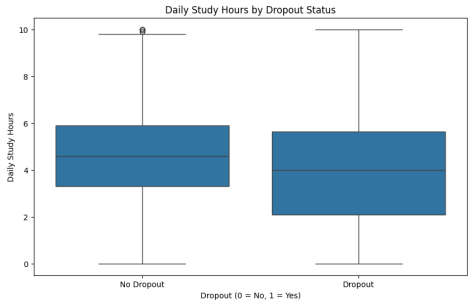

Fig. 2: Daily Study Hours by Dropout Status.

Figure 2 is a boxplot of the study hours of dropout and non-dropout students. It depicts that dropout students generally have lower study hours and more variability in their study patterns than non-dropout students. It thus indicates that less and uneven study time is a strong predictor of dropout risk.

Visualization highlights the significance of continually study habits in academic success. Students with consistent and longer study durations exhibit greater stability and lower chances of dropping out. On the other hand, irregular patterns mean a lack of discipline and academic engagement. This reinforces the role of time management as an important factor in predicting student retention.

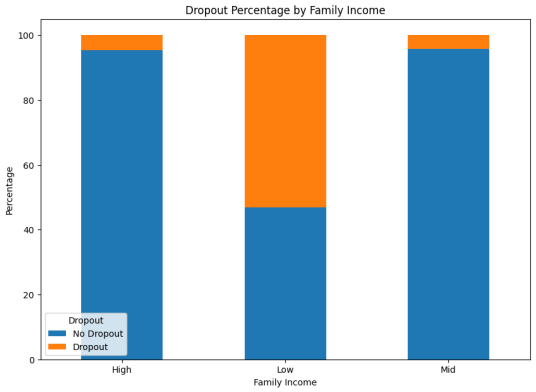

Fig. 3: Dropout Percentage by Family Income.

Figure 3 displays the proportion of families with lower income groups who drop out. Shows that students with lower incomes are more prone to dropping out than those who have higher and middle income levels. The effects on persistence in and dropout models, as well as socioeconomic status, are impacted by this.

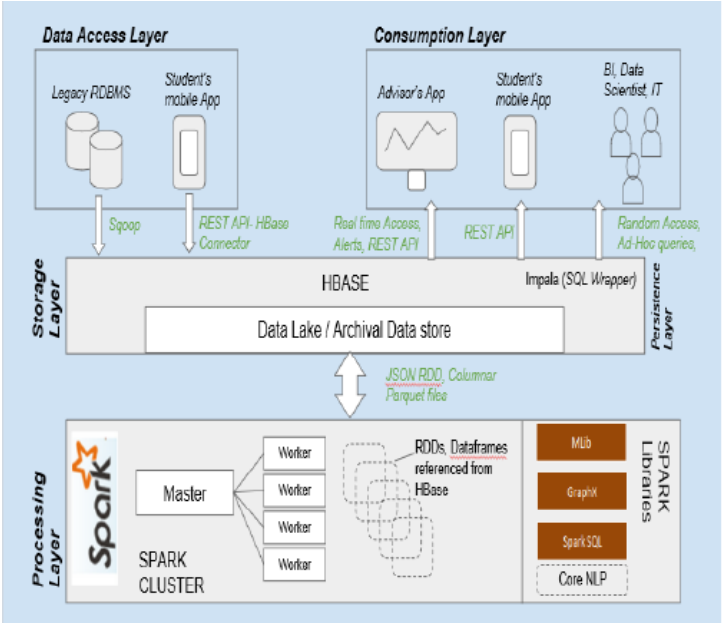

Fig. 4: Proposed Architecture in Alblawi and Alhamed (ICBDA 2017).

Alblawi and Alhamed's proposed big data and NLP architecture is illustrated in Figure 4 of ICBDA 2017. Data access, storage, and consumption are among the layers that make up the architecture. A Spark cluster is utilized for massive processing. In order to improve learning analytics and predict dropouts, the framework highlights how structured data sets like student interactions and text-based feedback can be merged with unstructured learning data. [7]

D. Analytical Approach

Statistical Analysis

Early learning behavior research utilized descriptive statistics (distribution, frequency counts) and inferential methods like correlation and regression analysis to investigate the links between variables. Despite their interpretability, they are usually limited in their ability to convey non-linear and high-dimensional connections.

Machine Learning Classifiers

Support Vector Machines (SVM) —Provides effective solutions for binary classification tasks, such as dropping out, when the feature space is well-defined and separated. Hence,

Logistic Regression (LR) — A basic model used for its simplicity and easy interpretation to predict the probability of dropout.

Random Forest (RF) — A method that utilizes multiple decision trees to improve prediction accuracy and reduce overfitting in an ensemble.

Gradient Boosting Decision Trees (GBDT) — When dealing with noisy and unbalanced educational data, boosting algorithms such as XGBoost or LightGBM often perform better than simpler classifiers.

Deep Learning Models Long Short-Term Memory (LSTM) networks are well-adapted to model the temporal learning pattern using sequences. By learning new dependencies over time (such as weekly usage patterns), they can make use of their skills to predict early dropouts. Improvements in handling unpredictable time spacing and altered sequence lengths have been achieved by extensions such as bidirectional LSTMs and attention mechanisms [3].

V. Dropout Prediction in Online Learning

A. Definition and Importance

In the realm of distance learning, dropout is defined as the act of discontinuing a course before its completion. Various studies have yielded inconsistent definitions: [3]

Platform-specific definitions are employed in certain cases, where a student is considered to be inactive for an assigned timeframe, such as without clicks, submissions or logins. This occurs typically after 10 consecutive days of completion. [2]

Others utilize academic performance metrics, where students who do not submit required assignments or quizzes are classified as dropouts. [4] [5]

A finer-grained strategy accounts for module-level dropouts, which includes students who drop out of a certain unit but remain enrolled in the course.

Significance: Dropout is a vital issue in MOOCs and online courses, where completion rates are consistently lower than 15%. High dropouts not only lower the educational impact but also lead to [1] [6]

- Misspent instructional resources — hosting, grading, and support expenses are incurred for non-completing learners.
- Reputation risks to the institution — ongoing high dropout rates might lower perceived course quality.
- Opportunity costs — learners' time and motivation are lost, a disincentive to further online education participation.

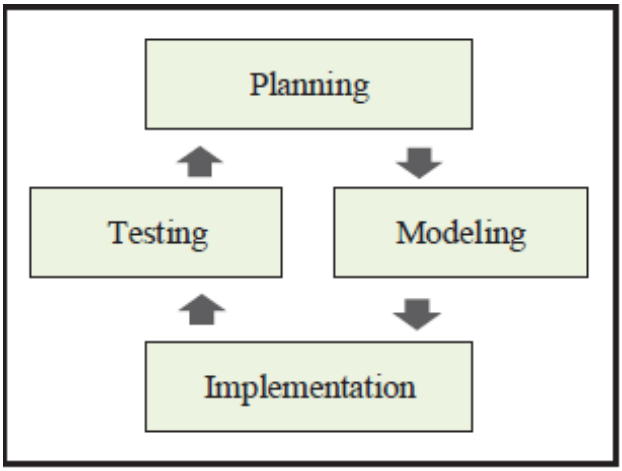

Fig. 5: Block diagram representation of the proposed 4SM scheme.

Figure 5 displays a block diagram of the proposed 4SM concept. It is divided into four stages: Planning, Modeling, Implementation and Testing.). A loop of continuous improvement is established in educational analytics through the descriptive flow where each stage feeds into the next. This model is a definitive guide to developing and improving big data-based solutions for online learning. Predictive models are assessed and systematically deployed. [9]

VI. Model Performance and Evaluation

This section presents the performance metrics and confusion matrices for the Logistic Regression, Random Forest, and Gradient Boosting models. [10] [11]

TABLE I: Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9810	0.9845	0.9227	0.9526
Random Forest	0.9940	1.0000	0.9710	0.9853
Gradient Boosting	0.9980	0.9952	0.9952	0.9952

As shown in Table I, Student dropout was accurately predicted by all three models. With the highest accuracy, precision, recall, and F1-score achieved through Gradient Boosting. Both dropout and non-dropout students can be accurately identified by this. Also, Random Forest was a hit with its

perfect precision score of 1.0000; this is in addition to the lack of false positives. Figures 6, 7, and 8 The confusion matrix offers an in-depth explanation of the classification results for each model. They present the quantities of true positive integers and false negative ones.

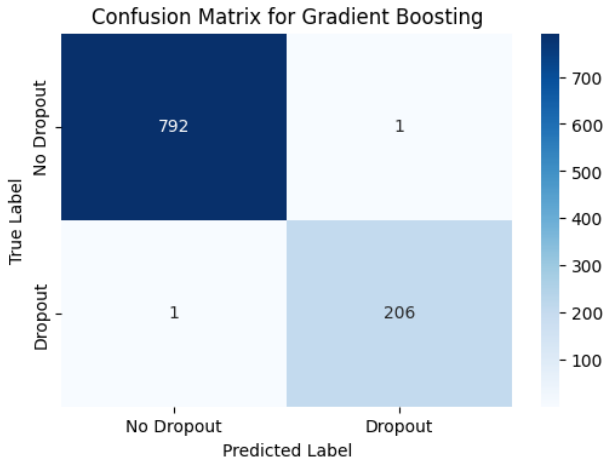

Fig. 6: Confusion Matrix for Gradient Boosting Model.

Referencing Figure 6, the confusion matrix highlights just how well the Gradient Boosting model performed. It achieved notably high accuracy, correctly identifying almost all instances of both dropouts and non-dropouts. Only a handful of misclassifications appeared. This really points to the model's solid handling of class imbalance and ability to capture intricate patterns within the dataset. In short: it delivered impressive, reliable results. [12]

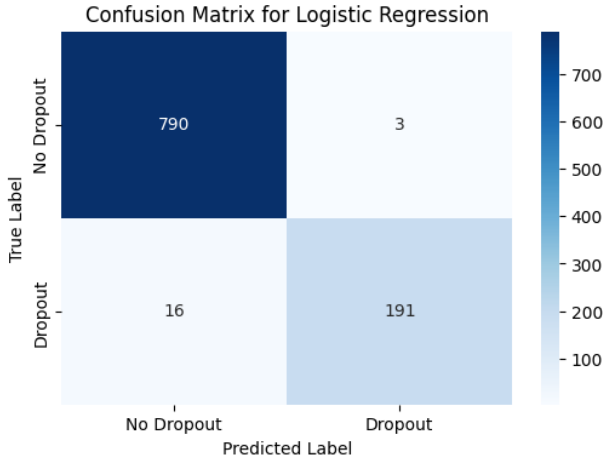

Fig. 7: Confusion Matrix for Logistic Regression Model.

The confusion matrix for the Logistic Regression model is shown in Figure 7. Most students were classified correctly using this model, but it had a weaker ability to identify dropouts, especially when contrasted with the Gradient Supporting approach used in the previous study. Although widely available, Logistic Regression is limited in its ability to handle nonlinear patterns in student data. Thus, it is not as effective in predicting dropout rates among students.

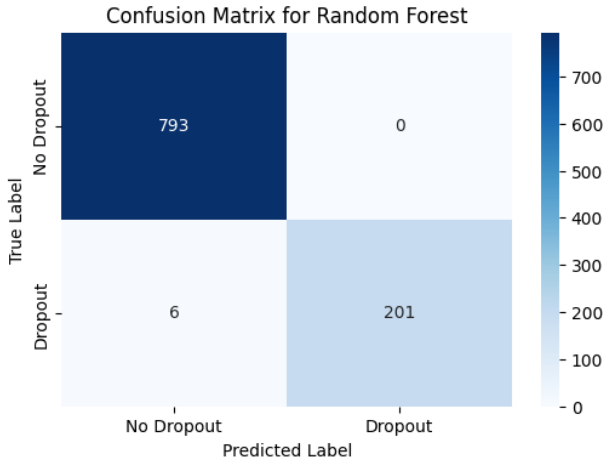

Fig. 8: Confusion Matrix for Random Forest Model.

In Figure 8, the random forest model exhibits significant predictive power as evidenced by the confusion matrix shown below. Even though the model does occasionally misclassify cases, it is still able to distinguish between dropout and non-dropout cases with accuracy. Interestingly, Random Forest strikes a good balance between accuracy and sensitivity, accurately representing both categories with fidelity.? The model's inflexible approach to both categories is a significant feature, even though it may not be flawless.

A. *Feature Importance Analysis*

This section looks at how important different features are according to the Random Forest and Gradient Boosting models. It offers insights into which factors have the biggest impact on student dropout.

VII. Dataset Used

TABLE II: Top 10 Feature Importances for Random Forest

Feature	Importance
family_income_Low	0.371818
admission_taken_Yes	0.287042
daily_study_hours	0.094602
family_income_Mid	0.042999
jee_main_score	0.033691
mock_test_score_avg	0.028153
jee_advanced_score	0.027839
class_12_percent	0.026923
peer_pressure_level_Low	0.025014
peer_pressure_level_Medium	0.019177

The table presents the top ten factors produced by the Random Forest model, ranked according to importance. The strongest predictor is socioeconomic status indicated by "Family Income Low" (0.372) which is exceedingly more important than any of the other factors. The next strongest predictor is "Admission Taken Yes" (0.287), which reaffirms

TABLE III: Comparative summary of the reviewed papers

Paper / Focus	Data / Features	Methods	Key Outcome
Qureshi et al.	Admin logs, LMS, external data	Association, clustering, dashboards	Framework, impact visualization
Alblawi	Structured marks, text	Spark, CoreNLP, RF	Sentiment improves prediction, pipeline arch.
Khan et al.	Blended learning data	4SM process	Practical model, incremental adoption
Alternate copy	Clickstreams, assessments	Data mining, visualization	Adaptive learning, KPIs
Kuzilek et al.	Large-scale MOOC dataset (student activity, demographics)	Statistical analysis, learning analytics	Open benchmark dataset for dropout research
Mostipak	Student demographics, activity logs	ML baselines (LR, RF, GBDT)	Public dataset enabling model benchmarking
Deep Learning Dropout Study	Sequential logs, video engagement	RNN, LSTM models	Captures temporal behavior, improves early prediction
Kaushik et al.	Student comments, online learning activities, academic records	Statistical, computational, and analytical techniques (descriptive, predictive, prescriptive)	Use of data analytics to forecast student performance and offer individualized instruction
YU et al.	Not specified beyond "big data"	Big data analysis method, MongoDB	A method for using big data to reconstruct online education for individualized development
Fuentes et al.	Student dropout risk factors	Fuzzy logic-based algorithm, prescriptive analytics	An algorithm to identify and suggest appropriate intervention programs for students at risk of dropping out
Laveti et al.	MOOC data from edX platform (over 2 lakh students from 39 courses)	Learning analytics, Apache Spark, ML algorithms (Random Forest, Gradient Boost, Logistic Regression), stacked ensemble model	A drop-out prediction model with 91.2% accuracy
Santur et al.	Online education data (age, gender, occupation, education level, location)	Machine learning and big data approach	A proposed approach to develop student-specific content and learning activities based on data analysis
Zan	Structured and unstructured data (not specified beyond "big data")	Big data analytics, sociotechnical approach	An examination of how big data can improve education quality and reduce dropout rates

the importance of enrollment status on outcomes. Variables related to academic performance demonstrated significantly lower contributions: Daily Study Hours (0.095) to multiple exam scores of "JEE Main Score": 0.034 and "JEE Advanced Score" 0.028. "Class 12 Percent" (0.023) and "Average Mock Test Scores" (0.022) also add evidence that study habits and academic preparation are relevant predictors. "Family Income Mid" (0.049) illustrates the nuances that require attention where the socio-economic status dichotomy becomes more hazy. There were indications of psychosocial variables with "Peer Pressure Level Low" (0.025) and "Peer Pressure Level Mid" (0.019), suggesting there are environmental factors at play, although minor relative to the other factors of interest. Overall this data has measured the multiple complexities of dropout risk and future trajectory factors that include social-economic status, academic performance, and psychosocial context, more complexity than simply one more factor. [8] [13].

TABLE IV: Top 10 Feature Importances for Gradient Boosting

Feature	Importance
admission_taken_Yes	0.479400
family_income_Low	0.328508
daily_study_hours	0.078028
peer_pressure_level_Low	0.048911
peer_pressure_level_Medium	0.048425
jee_advanced_score	0.009555
jee_main_score	0.006659
class_12_percent	0.000181
mental_health_issues_Yes	0.000180
mock_test_score_avg	0.000124

The table shows the key features that predict student dropout using the Gradient Boosting model. The results indicate that `admission_taken_Yes` and `family_income_Low` are the main predictors, contributing over 80% to the model's decisions. Other features, such as `daily_study_hours` and peer _pressure _`level`, also matter, but to a lesser extent.

Both the random forest and gradient boost models identified `family_income_Low` and `admission_taken_Yes` as highly significant predictors of dropout. This suggests that socioeconomic factors and whether a student has secured admission elsewhere play a crucial role. `daily_study_hours` also consistently showed up as an important feature in both models, highlighting the impact of engagement levels on the likelihood of dropping out. Other notable features include `peer_pressure_level` and `jee_main_score`/`jee_advanced_score`.

VIII. Comparative Insights from Literature

This summary highlights shared themes and distinct contributions from the four papers reviewed.

Synthesis: In all studies the recurring recommendations are:

- Combine structured and unstructured (text) features; NLP adds value.
- Use distributed processing (Spark/HBase/HDFS) for scale.
- Prefer ensemble tree models and/or sequential models (LSTM) depending on feature type.
- Design pipelines that feed analytics into instructor dashboards for interventions. [7]

IX. Challenges and Limitations

Adopting big-data-driven dropout prediction and learning analytics presents several challenges:

- **Data Quality & Missingness:** incomplete profiles, noisy clickstreams, and inconsistent logging formats. [13] [5]

- **Privacy and Ethics:** need for consent, de-identification, and careful handling of sensitive attributes, such as socioeconomic or health data. [4] [15]
- **Scalability Infrastructure:** processing at scale requires distributed stores and computing, like HBase, HDFS, and Spark, along with proper configuration. [12] [5]
- **Model Interpretability:** tree ensembles and deep models can be opaque; it is recommended to use XAI tools, such as SHAP and LIME, for actionable explanations. [3] [15]
- **Generalizability:** models trained on one institution or dataset may not transfer well; careful feature normalization and re-calibration are necessary. [8] [11]

X. Conclusion

Big Data Analytics presents tremendous potential for enhancing online learning through the delivery of customized learning experiences, early warning signs for struggling students, and institutional policy choices. The studies used in this review indicate that the integration of organized educational data sets with unstructured feedback, such as text-based remarks, and the examination of both using distributed systems, such as Spark or HBase, results in improved predictive modeling. It also helps to create effective intervention pipelines. The principal challenges are privacy and interpretability, and the feasibility of incremental deployment. The future efforts should be to combine different types of data and privacy-protecting analytics and to develop transparent explainability frameworks so that analytics are effective and ethically help student welfare. [7] [6] [5]

XI. Future Directions

From the literature surveyed and evidence gathered in this study, some promising directions can enhance big data analytics in education:

- **Multimodal Integration:** Future studies can extend beyond one-source data by integrating categories of data such as video lectures, clickstream activity, discussion boards, physiological responses, and sensor data. Integrating these will produce comprehensive learner profiles that incorporate the cognitive, emotional, and behavioral dimensions of learning. This integrated strategy enhances prediction accuracy and personalization. [11] [3]
- **Real-time Intervention Engines:** Probably the most extreme near-term application will be the development of rapid analytics pipelines to identify disengagement as it happens, in real-time. These pipelines can provide micro-interventions like personalized nudges, adaptive content, or reminders. In effect, prediction models become proactive learners, and support retention and student success. [1] [14] [15]
- **Federated Learning and Privacy-Preserving ML:** Since student data is sensitive, federated learning enables collaborative model training by institutions without exposing raw data. Minimize privacy threats while enabling scalability. Methods such as differential privacy and secure multiparty computation provide additional security for learner data. [15]

References

[1] H. Qureshi, A. K. Sagar, R. Astya, and G. Shrivastava, "Big data analytics for smart education," in *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, 2021.

[2] T. Yu, D. Jo, and S. Kim, "Student dropout prediction in online courses using temporal activity patterns," in *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, 2017.

[3] A. Fuentes, H. Huang, R. Khor, and D. Scott, "Deep learning-based early prediction of student dropout in massive open online courses," *IEEE Access*, vol. 8, pp. 190 474–190 484, 2020.

[4] H. Zan and B. Yildiz, "A comparative study on machine learning algorithms for predicting student dropout in online learning," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 22, pp. 36–49, 2019.

[5] R. Laveti, K. R. Moparthi, and A. S. Reddy, "Student dropout prediction in moocs using big data analytics," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 3, pp. 3755–3761, 2020.

[6] S. Kaushik and P. Saini, "Predicting student performance using machine learning algorithms," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, 2019.

[7] A. S. Alblawi and A. A. Alhamed, "Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, nlp and analytics," in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, 2017.

[8] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, no. 170171, 2017.

[9] A. Khan, M. K. Vakil, and M. S. B. Arif, "Proposed 4sm model for indian education system implementing big data analytics and blended learning methods," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, 2018.

[10] S. Yu, D. Yang, and X. Feng, "A big data analysis method for online education," in *2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2017.

[11] Y. Santur, M. Karaköse, and E. Akın, "Improving of personal educational content using big data approach for mooc in higher education," in *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2016.

[12] R. N. Laveti, S. Kuppili, J. Ch, S. N. Pal, and N. S. C. Babu, "Implementation of learning analytics framework for moocs using state-of-the-art in-memory computing," in *2017 5th National Conference on E-Learning and E-Learning Technologies (ELELTECH)*, 2017.

[13] J. Mostipak, "Educational dropout prediction dataset," https://www.kaggle.com/datasets/jessemostipak/educational-dropout-prediction, 2019, accessed: 2025-09-15.

[14] P. Kaushik, M. Kakkar, M. Yadav, C. Jegadheesan, P. N. Sripada, and K. Chahal, "Data analytics in education: Enhancing student learning outcomes," in *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, 2024.

[15] N. B. Fuentes, L. S. Feliscuzo, and C. L. C. S. Romana, "Enhancing student retention in higher education: A fuzzy logic approach to prescriptive analytics," in *2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI)*, 2024.