ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 8 (Issue 1) (2024) Pages 29-33

Leveraging Artificial Intelligence - Driven Automata for Improved Dropout Prediction in Schools

Jackson Chansa
Computer Science Department,
Copperbelt University
Kitwe, Zambia
jacksonchansa@yahoo.com

Jameson Mbale
Computer Science Department,
Copperbelt University
Lusaka, Zambia
jameson.mbale@gmail.com

***Abstract*— This work investigates the use of Artificial Intelligence - driven automata to predict student dropout risks, emphasizing how artificial intelligence can enhance educational sustainability. By integrating automata theory with cutting-edge machine learning techniques, the study develops a predictive framework to identify students at high risk of dropping out. The research includes a thorough review of automata theory, Artificial Intelligence applications in education, and current dropout prevention strategies. Using automata-based algorithms, we analyze educational datasets to model student behavior and risk factors. Preliminary findings indicate that these models can accurately forecast dropout probabilities, facilitating timely interventions to boost student retention and promote a more equitable educational system. Additionally, this work aligns with broader sustainability goals by supporting improved educational outcomes, which are crucial for advancing green economies amidst climate change. The results highlight the potential of automata-based AI in dropout prevention and offer insights for future research on integrating AI solutions in educational settings to address climate-related challenges.**

Keywords—Automata-based Models, Dropout Prediction, Artificial Intelligence, Educational Data Mining, Student Retention, Computational Intelligence in Education

I. INTRODUCTION

High school dropout rates present a significant challenge to educational systems worldwide. Predicting which students are at risk of dropping out allows for timely interventions which can lead to improved retention rates. The customary models of prediction are based on linear, static or rule based methodologies which can fail to capture the dynamic patterns of student behavior. Automata-based models provide a solution lying in the utilization of formal methodologies for modeling and analysis of behaviors.

Basing on [1], automata theory is fundamental in studying computation and dynamic systems modeling which also plays a crucial role in evaluating student behaviors. Therefore, methods such as Finite State Machines (FSM), Hidden Markov Models (HMM), Probabilistic Finite Automata (PFA), Cellular Automata and Regular Expression provide multiple ways to capture the temporal and probabilistic aspects of student data where the temporal aspect is important for predicting dropout risks.

Moreover, the sourcing of Artificial Intelligence (AI) especially in machine learning and data mining, has boosted the capacity of these models to deal with a huge amount of educational system. [2] also pointed out the importance of AI in the development of learning systems that can learn and improve over time while contributing to the development of optimal approaches to applying predictive modeling in education. Data mining preprocessing as discussed by [3], helped extract relevant patterns from the education data which in turn helped in the training of automata-based models for a better prediction of dropouts.

Furthermore, the concept of learning analytics described by [4] provides a theoretical perspective that is complementary to the goals of automata-based approaches by emphasizing the dynamics of students' monitoring. Such analytics, accompanied by AI-driven automata, can give insights on students' activity and results in real-time and thus help to identify students in danger zones.

The others include the application of computational intelligence in education which is also highlighted by [5] and enhances the aspect of AI- driven automata in altering educational practices. This makes it possible to provide tailored care, thus enhancing students' retention and success rates for which these models will be useful from the results of analyzing intricate datasets.

On the basis of such automata-based algorithms as Finite State Machines, Hidden Markov Models, Probabilistic Finite Automata, Cellular Automata, and Regular Expressions, this paper explores how AI can be used in improving the prediction of student dropping out risk with the help of educational data. The purpose of this research is therefore to propose AI driven automata as a potential solution to one of the biggest problems facing education through a blend of theoretical analysis and implementation..

A. Problem Statement

High dropout rates in educational institutions continue to be a major challenge preventing the achievement of high student completion and retention rates. the current application of traditional predictive models that rely on static and linear or rule based approaches have often proven to be inadequate in fully capturing the behavior of students and their socioeconomically dynamics. Therefore, they are restricted in their capacity to predict dropout risks and to offer timely

solutions. This calls for more complex and adaptive models that will determine and or predict dropout risks by utilizing complex calculations and algorithms.

B. *Research Objectives*

1. To evaluate the effectiveness of automata-based AI models in predicting students at risk of dropping out or missing school hours.
2. To investigate how automata-based AI models can be applied to large-scale educational datasets to identify students at risk of dropout or missing school hours.
3. To establish the impact of automata-based models on improving completion and retention rates for a sustainable education system.
4. To determine if automata-based models for dropout prediction can be improved in terms of predictive accuracy through the application of data mining techniques.
5. To examine at the real-world effects of using automata-based AI models in schools for long-term workforce development.

C. *Research Questions*

1. To evaluate the effectiveness of automata-based AI models in predicting students at risk of dropping out or missing school hours.
2. To investigate how automata-based AI models can be applied to large-scale educational datasets to identify students at risk of dropout or missing school hours.
3. To establish the impact of automata-based models on improving completion and retention rates for a sustainable education system.
4. To determine if automata-based models for dropout prediction can be improved in terms of predictive accuracy through the application of data mining techniques.
5. To examine at the real-world effects of using automata-based AI models in schools for long-term workforce development.

II. Literature Review

The challenge of accurately predicting student dropout has been widely acknowledged in educational research. Traditional models for dropout prediction often rely on static and linear approaches, which can overlook the complexity of student behavior and the multifaceted nature of educational environments [2]. These limitations suggest the need for models that are more responsive to the temporal and probabilistic nature of the student data.

Computational Complexity has traditionally offered a firm basis for the analysis of automated computations at large [6]. There are other automata which include Finite State Machines (FSMs), and Hidden Markov Models (HMMs) which have been useful in modeling the temporal behavior of such systems [1]. These models are especially used when dealing with dropout prediction since they will enable the modeling of all the events that led to the decision that the student comes up with and makes to drop out of school.

In education, the use of artificial intelligence has been used to create even more complex predictions. [7] note that machine learning approaches in teaching procedures will create significant change since they offer better and more timely responses. These models proved to be useful in conjunction with the automata-based approaches to provide a more detailed insight into the processes of engagement and learning in students.

In addition, Data mining methods are used to capture interesting patterns from large databases and these patterns serve as a basis for developing predictive models [3]. By applying the analysis techniques on educational data it becomes possible to reveal such trends and behavior which are tightly connected to dropout risks. When these findings are incorporated into the automata-based models, the latter would be more accurate and reliable.

Learning Analytics has also proven to be a valuable component when it comes to the student retention issue. [4] defined learning analytics as the real-time tracking and analysis of learners' data for learning interventions. The combination of automata-based models and learning analytics can supersize a tool that will assist in the identification and reduction of dropout rates. As the discussion about Education as the field of Computational Intelligence is shown in the article by [5], it is crucial to understand that the educational field has to use such sophisticated methods as advanced algorithms and models to solve various kinds of difficulties. By using the automata-based approach to AI models, educators can find out even more about students' behaviors and potential ways to increase retention rates [8].

Recent advancements in Artificial Intelligence (AI), particularly in machine learning as researched by [10] explored the application of deep learning models for dropout prediction, significant improvements in accuracy were demonstrated compared to traditional methods. further, reinforcement learning techniques have been employed to develop adaptive learning environments that respond dynamically to student behaviors, as discussed by [11]. These newer methods highlight the potential for AI to transform dropout prediction by enabling more sophisticated modeling of student interactions and engagement

Ccombining ensemble methods with deep learning approaches resulted in superior performance in predicting student outcomes as [12] demonstrated. These findings underscore the necessity of incorporating recent advancements in AI and machine learning into dropout prediction frameworks to improve their effectiveness.

III. Methodology

A. *Data Collection*

The dataset used for this research includes a comprehensive range of educational data:

1) Student Academic Records: Grades, course enrollments, and attendance records.

2) Survey Data: Results from student satisfaction surveys

3) Demographic Data: Information on socioeconomic background, parental education levels, and geographical location.

4) Historical Dropout Data: Past dropout rates and retention trends at various education levels.

5) The dataset containing a total of 395 records was downloaded from the [9]

Dataset	Total Records	Features
UCI Student Dataset	395	33

B. Data Cleaning

Data preprocessing included imputation of missing data, handling outliers, and normalizing student features. Given the imbalanced nature of dropout vs. non-dropout data, we applied techniques such as Synthetic Minority Over-Sampling Technique (SMOTE) to balance the classes.

C. Feature Engineering

For feature engineering the three performance scores were used, if any of the scores G1, G2 or G3 were less than 10 then the student was labelled as a dropout(dropout=1), otherwise they were labelled as Non dropout(dropout=0).

D. Training and Testing

To train and test the models the 395 record dataset was split into 316(80% training) and 79 (20% testing/evaluation)

E. Implementation

The models were implemented using a Python libraries TensorFlow. Models were trained on historical data and evaluated using performance metrics accuracy, precision, recall, and F1-score.

IV. RESULTS

We evaluated the performance of :the Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Neural network models using key metrics which included accuracy, precision, recall, and F1-score, as these provide a balanced view of model performance in predicting school dropouts, considering the class imbalance the overall effectiveness of the model is measured by its Accuracy, while deeper insight into how well the model performs in predicting school dropouts is given by the precision and recall, Given the potential class imbalance (more non-dropouts than dropouts), the F1-score provides a balanced metric that considers both precision and recall.

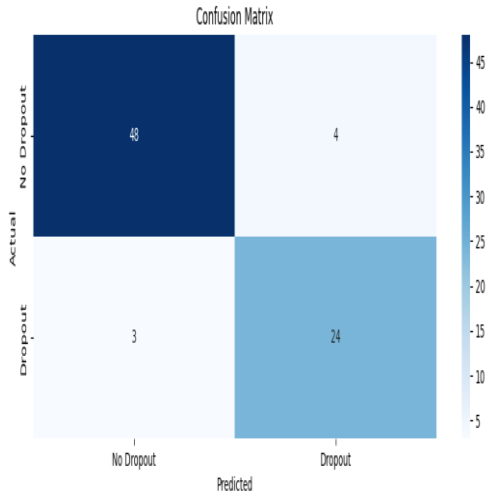

FIG 1: LOGISTIC REGRESSION CONFUSION MATRIX

Confusion Matrix Overview:

- **True Positives (TP)**: 24 cases where the model correctly predicted that students would drop out.
- **True Negatives (TN)**: 48 cases where the model correctly predicted that students would not drop out.
- **False Positives (FP)**: 4 cases where the model incorrectly predicted that a student would drop out when they did not.
- **False Negatives (FN)**: 3 cases where the model incorrectly predicted that a student would not drop out when they actually did.

Table: **Models Performance in Predicting Student Dropout**

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**0.91**	**0.86**	**0.89**	**0.87**
Decision Tree	**0.9**	**0.83**	**0.89**	**0.86**
Random Forest	**0.9**	**0.81**	**0.93**	**0.86**
K Nearest Neighbor(KNN)	**0.91**	**0.86**	**0.89**	**0.87**
Support Vector Machine(SVM)	**0.87**	**0.86**	**0.86**	**0.87**
Neural Network	**0.59**	**0.60**	**0.41**	**0.48**

- High Accuracy (91.1%): The Logistic Regression model performed very well in predicting the correct class (dropout or no dropout) for the majority of the cases.

- High Recall (88.9%): The model was very effective in identifying students who are at risk of dropping out, missing only 3 out of 27 actual dropouts.
- Good Precision (85.7%): There were some false positives (4 students were predicted to drop out but didn't), but the precision was still high, meaning most of the dropout predictions were accurate.
- Balanced F1-Score (87.3%): The F1-score shows that the model maintains a good balance between precision and recall, making it suitable for both identifying at-risk students and minimizing incorrect predictions

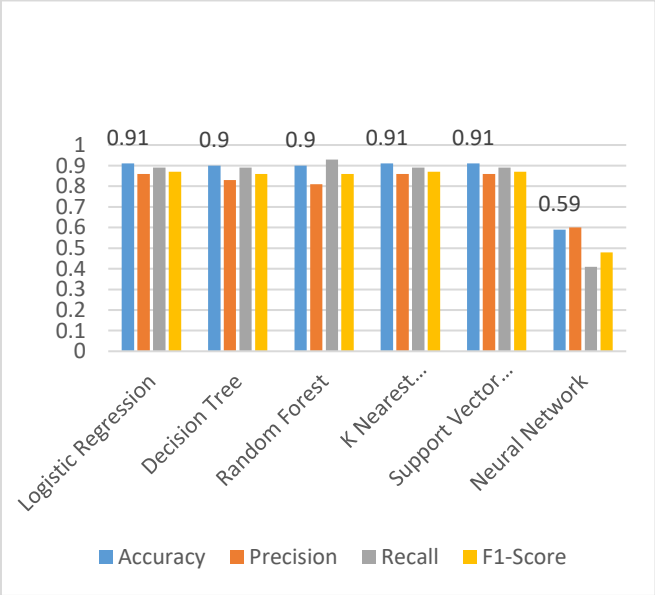

Fig 2: Graphical Comparison of Models Performance

V. Discussion

The logistic regression model with an accuracy of 91 percent is used as a good benchmark for dropout prediction. although very simple, it remains interpretable and yields actionable information on students' dropout risks, making it suitable for real-world educational interventions, the Random Forest model had nearly performed as efficient as the logistic regression model, an indication that an ensemble model can effectively predict dropouts. Although Decision Trees, while performing slightly lower, highlight the need for model optimization when dealing with large datasets. The neural network model significantly underperformed likely due to many factors such as inadequate training data , poor parameter tuning, overfitting, model complexity, Models with a high F1-Score had a good balance between precision and recall making them suitable for identifying at risks students and minimizing incorrect predictions, which can have an unintended effect on both the school and/or student affected Future research should focus on adequately training neural networks and ensemble model to better capture non-linear relationships between student data features and dropout risks.

VI. PRACTICAL IMPLICATIONS

Integrating AI-driven models into existing school systems necessitates careful planning. The ethical implications and barriers to adoption must be addressed proactively to ensure successful implementation. Educational institutions should prioritize transparency and data security while equipping educators with the necessary tools and training to utilize these models effectively. These models can be seamlessly incorporated into existing school systems to monitor student performance and engagement continuously. By providing educators with real-time insights into students at risk of dropping out, timely interventions can be implemented to support at-risk students effectively.

To facilitate the practical implementation and scalability of these AI-driven models, several considerations should be addressed. Schools must invest in the necessary infrastructure to support data collection and analysis, including robust Learning Management Systems (LMS) that can integrate AI tools. Training educators to understand and utilize these models will also be crucial. Additionally, collaboration between educational institutions and technology providers can help tailor AI solutions to meet the unique needs of different educational environments.

Moreover, developing user-friendly interfaces for educators can enhance the usability of these models, allowing teachers to access insights and recommendations easily. The scalability of these AI-driven solutions is essential to ensure they can be deployed across various educational contexts, from small rural schools to large urban districts. By addressing these practical considerations, educational stakeholders can harness the power of AI to improve student retention and foster sustainable educational outcomes.

VII. Limitations

The datasets that were collected for our study had several limitations that impacted model performance. One major issue was inadequate representation of student characteristics such as behavior, disciplinary status, and interaction with Learning management systems(LMS) which led to limited training data for the models, hindering the model's ability to learn effectively. Additionally, the presence of missing values and wrong data formats further reduced the number of usable records, compounding the challenge of developing robust predictive models. The imbalance between the dropout and non-dropout instances also posed a problem with significantly more non-dropout cases than dropouts, the models were prone to biased predictions favoring the majority class. Moreover, the assumption of linear relationships in some modelling approaches may not capture the complexities of student behaviors and dropout risks.

VIII. Conclusion and Future Work

The results demonstrate the potential of automata-based AI models in predicting student dropout with high accuracy. Future works should try to address the various limitations presented in this study such as use of advanced and ensemble AI models which can address the relatively low precision all

employed models gave, which can significantly reduce the number of false positives, However, advanced models like neural networks and reinforcement learning techniques can further refine prediction accuracy. Future works around actual deploying or integrating AI student dropout predictor in school learning management system. Future work should address the limitations presented here by collecting more diverse and comprehensive data, ensuring proper data formatting and implementing methods that can better handle class imbalances and complex relationships among variables.

REFERENCES

[1] M. Sipser, Introduction to theory of computation(2nd ed.), Thomson course technology, 2006.

[2] S. &. N. P. Russell, Artificial intelligence:A modern approach(4th ed,), Pearson, 2021.

[3] J. P. Han, Data mining:Concepts and techniques(3rd ed)., Morgan Kaufmann, 2011.

[4] S. F. Buckingham Shum, Learning analytics:Theoretical framework and practices., Springer., 2012.

[5] S. &. J. L. Brahman, Computational intelligence in education., Springer, 2011.

[6] D. Kozen, Automata theory and computability., Springer, 1997.

[7] W. &. F. C. Holmes, Artificial intelligence in education:Promises and implications., Routledge, 2019.

[8] M. V. M. J. M. a. L. B. Realinho, Predict Students' Dropout and Academic Success, 2021.

[9] M. Lichman, ""UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences.," [Online]. Available: https://archive.ics.uci.edu/ml. [Accessed 10 09 2024].

[10] Zheng, Y, Reinforcement learning for adaptive learning environments: Enhancing student engagement and retention. Journal of Educational Technology.2023

[11] Doleck, T., Lemay, D. J., & Brinton, C. Predictive analytics in education: A comparison of deep learning frameworks. Education and Information Technologies, 2021