SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V12I7P105

Volume 12 Issue 7, 43-49, July 2025
© 2025 Seventh Sense Research Group®

Original Article

Mining Educational Data to Predict Influential Factors Patterns for Student Dropout in Iraq

Ali Abdul Kadhim Aakool[1], Dahair Abbas Redha[2], Hayfaa Abdulzahra Atee[3]

[1,2]*Department of Information Technology, Technical College of Management, Middle Technical University (MTU), Baghdad, Iraq.*
[3]*Department Computer Systems, Institute of Administration Rusafa, Middle Technical University (MTU), Baghdad, Iraq.*

[1]*Corresponding Author : dac2006@mtu.edu.iq*

Received: 17 May 2025 Revised: 29 June 2025 Accepted: 17 July 2025 Published: 30 July 2025

Abstract *- This study investigates the application of data mining techniques to identify and predict influential patterns leading to student dropout in Iraq. Given the growing concern over high dropout rates and their socio-economic implications, a dataset was collected from 541 students across 43 elementary and secondary schools in Kut city. The study applied five supervised machine learning models: Decision Tree, Random Forest, Support Vector Machine, XGBoost, and Logistic Regression. After preprocessing and evaluation, the models were compared using accuracy, recall, precision, and ROC curve metrics. Results revealed that Logistic Regression and Random Forest models performed best, achieving 96% accuracy. The most significant factors contributing to dropout included academic performance (GPA, participation, failure history), institutional support, and socioeconomic factors. This research highlights the potential of educational data mining to proactively identify at-risk students and inform targeted interventions in the Iraqi education system.*

Keywords *- Educational Data Mining, Student Dropout, Iraq.*

1. Introduction

Education is a crucial component that significantly contributes to the socioeconomic advancement of individuals, communities, and nations. Consequently, countries must prioritize establishing and enhancing educational frameworks as instruments for development and capacity enhancement. In this regard, numerous scholars have examined the challenges confronting this domain and have endeavored to identify effective solutions to augment education's efficacy [1].

One of the significant problems facing the educational systems is student dropout, which is considered one of the most critical problems. It constitutes one of the factors contributing to “Educational Wastage”. The first studies in this field appeared in the United States of America during the 1930s of the last century. John McNeely was one of the first researchers in this field. He conducted an examination of the demographic and social characteristics of students of 60 educational institutions [2]. The advent of data mining techniques has generated considerable interest within this domain of knowledge, as evidenced by the numerous studies that have been published. This environment is characterized by abundant data, various complex problems, and diverse objectives. As a result, it is regarded as a fertile ground replete with challenges [3].

The evolution of educational data mining is inextricably linked to the progression of machine learning and statistics. Significant advancements in numerous techniques utilized in contemporary educational data mining applications emerged during the 1990s and early 2000s. The emergence of educational data mining as a distinct research area can be pinpointed to a workshop held at the University of Melbourne, Australia, in August 2004. Since this significant event, the field has undergone considerable growth, evidenced by the annual publication of numerous research papers, the organization of an international conference in Montreal, Canada, in 2008, and the creation of a specialized scientific journal focused on educational data mining, with its first issue published in October 2009. A wide range of research papers tackles various challenges within the education sector and illustrates effective solutions achieved through the application of data mining methodologies [4].

Researchers differed in defining the phenomenon of dropout due to the difference in scientific fields that covered the topic, and the wide difference in viewpoints between researchers in the same field due to the multiplicity of objectives of conducting these studies, which resulted in many definitions of this concept. However, in this study, the official definition adopted by the United Nations Educational, Scientific and Cultural Organization

(UNESCO) was addressed, where it defined dropout as “every person who does not complete his studies and leaves the teaching benches before completing the years of study” [5].

The societal and educational problems stemming from dropouts are very harmful, including the waste of resources and the ineffectiveness of education as a whole. Moreover, this continues to increase the rates of child labor, homelessness, and uneducated workers entering the workforce. This also increases the chances of many ignorant problems like drug use, terrorism, crime, and immorality [6]. Also, there are economic repercussions of student dropout; the financial burdens incurred by dropout rates and failures within the educational system in Iraq were assessed utilizing data from 2014-2015. This assessment is detailed in the report published by UNICEF in 2016, which estimates these costs at 1.5 trillion Iraqi dinars. This amount represents 18.8% of the investment budget specifically allocated to the education sector [7].

The problem of this study lies in the early detection of the most common influential factors patterns that lead students to drop out, which, according to the researchers' knowledge, is considered one of the first studies to apply data mining techniques to real data in the Iraqi educational system.

The gap is the lack of applied research that employs a large number of social, economic, institutional and academic characteristics for the purpose of classifying students at risk of dropping out. Despite the difficulty of collecting data from schools due to the lack of a central database in the Iraqi education system, observations were obtained.

This study aims to explore the patterns of reasons that lead to students dropping out of school in Iraq by applying statistical methods and data mining techniques to predict students who are at risk of dropping out.

2. Related Works and Research Gap

There are many significant studies covering student dropout in higher education institutions, but there are just a few of them focusing on elementary and high school, so this field conceder as an emerging research area. Barros et al. employed three data mining techniques on a dataset obtained from the Unified Public Administration System (SUAP) in Brazil, 2018, which consists of 7718 observations and 25 features. Their findings indicated that using Balanced demonstrated superior performance and enhanced accuracy of the used techniques [8]. Orooji and Chen found that Decision Tree precede other techniques used, like Artificial Neural Network (ANN) and Multi-Layer Perceptron, in imbalanced data, despite using learning approaches to deal with such data; accuracy achieved around 91%. The data comprises administrative records from the Louisiana Department of Education, comprising 366806 observations and 18 features [9].

Hassan and Mizra employed several data mining techniques, such as Decision Tree, Random Forests, Logistic Regression, Support Vector Machine and Naïve Bayes to predict student retention in developing countries using socioeconomic and school factors as exploratory variables. Results show that Logistic Regression and Random Forest are the most effective algorithms for predicting school dropout in the used dataset [10]. Mnyawami et al. developed an AutoML technique to improve the accuracy of predictions using algorithms like Decision Tree, K-Nearest Neighborhood (KNN), Multi-Layer Perceptron and Bayesian Classifier. The study found that students’ grades, number of siblings, distance from school and age are the most statistically significant factors compared to other factors [11].

Selim and Rezk, developed a Logistic Regression Classifier to forecast at risk students using a dataset from the Survey of Young People in Egypt (SYPE) where several experiments containing undersampling and oversampling methods are conducted to enhance the performance of the classifier, chronic diseases, co-educational settings, parental illiteracy, educational performance, and teacher care are the most significant factors influencing students dropout [12]. Krüger et.al, propose an approach to build a generalizable dataset which has been applied to predict student dropout cases, the dataset were collected from 19 schools in Brazil, the results showed that there is a dynamic results differences between countries due to their regulation and many different features used in studies, however XG Boost outperformed other techniques used in study [13].

Bulut et al., developed a method to collaborate the human and machine visions to enhance the efficiency of decision making and increase the accuracy of predicting factors of dropping out, the study used an empirical data from High School Longitudinal Study of (HSLS:09) in United States, results indicated that Random Forest preformed butter than deep learning in this case [14]. Sariman et al. employed three machine learning algorithms to forecast student attrition from public educational institutions in Selangor, Malaysia. These algorithms included Naïve Bayes, Decision Tree and Random Forest. The dataset from students and schools consisted of 2482 observations and 22 attributes. The findings indicated that the Random Forest algorithm demonstrated superior predictive accuracy in comparison to the other algorithms examined in this study [15]. Awedh and Mueen combined two machine learning algorithms, Logistic Regression and K-Nearest Neighborhood (LR-KNN), for predicting at-risk students. They utilized a dataset from King Abdulaziz University, their findings show that the proposed model enhanced the accuracy, precision, recall and F1-score [16].

3. Methodology and Data Preparation

3.1. Dataset Description

In this study, a questionnaire form consisting of 48 demographic, socioeconomic, academic and institutional attributes was collected from elementary and secondary administration schools and students in Kut city center of Wasit governorate. The sample size was equal to 541 observations distributed among 43 schools in different urban and rural districts. The total number of dropouts in this study was 134 students, equivalent to 24.7% of the total number of dropouts; 43 were males and 91 were females. Many procedures were applied to this dataset to get it ready to use by machine learning algorithms.

Prior to initiating the application of data mining techniques, it is imperative to possess a meticulously arranged and prepared database. Consequently, data preprocessing is crucial for the successful execution of the prediction process. Several steps must be adhered to, including:

3.2. Data Preprocessing

3.2.1. Missing Value Handling

Missing data is a relatively common problem in almost all research; it can greatly impact conclusions drawn from data. Missing values (or missing data) are defined as data values that are not stored for a variable. Therefore, some studies have focused on dealing with missing data and the problems it causes. Several methods for dealing with this problem include ignoring the entire constraint, filling a general constant in the missing data or using the most likely value using Bayesian theory. The most common of these methods is to use one of the measures of central tendency (the arithmetic mean, median or mode), depending on the type of data [17].

This study used means for numeric variables and modes for categorical variables. The largest number of missing values was found in the Grade Point Average (GPA) of the last year variable(X_{22}), which was 44 values. The second one was the income variable (X_5), which contained 17 values. The other variables contain scattered missing values. The total number was 119 missing values (Figure 1).

3.2.2. Encoding

Programming algorithms do not deal with categorical variables, so one of the coding techniques must be used to transform these variables into numerical representations. There are several encoding techniques for processing categorical variables, such as One Hot Encoding, which is used for variables containing only two categories and converted into a binary matrix. Label Encoding, used for variables containing more than two categories, gives each category a unique numerical value. In this paper, Label Encoding was used because most of the study variables were categorical and had more than two categories.

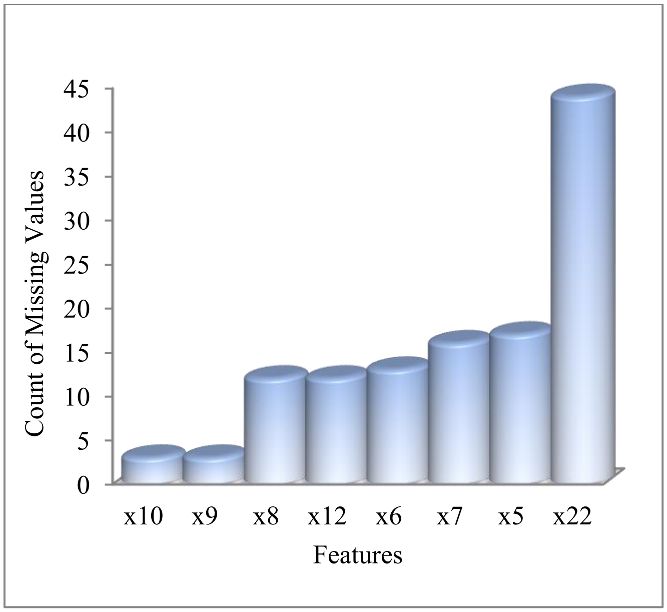


Fig. 1 Missing values in the dataset

3.2.3. Data Splitting and Validation

The dataset was divided into two sets, one for training the model and the other for testing it, at a rate of 70% and 30%, respectively. So the training set was equivalent to 378 observations, and the remaining 163 were the testing set. To guarantee randomness and avoid bias in the selection of the two sets, the 5-Fold Cross Validation technique was used, which divides the data into five subsets, each subset being used as test data in a specific repetition.

3.3. Machine Learning Techniques

Five supervised machine learning techniques were implemented to predict student dropout influential patterns: Decision Tree, Random Forest, Support Vector Machine, Extreme Gradient Boosting (XGBoost), and Logistic Regression (LR).

4. Results and Discussion

Five techniques were applied to predict student dropout patterns. The elementary results were unsatisfactory and had some technical problems, such as overfitting, so some enhancements were applied by tuning the hyperparameters of these algorithms. Finally, optimal performance was reached and measured by employing a confusion matrix and mean squared error. The tables below show the results of using models for the testing set.

4.1. Decision Tree

The number of correct and incorrect predictions by the Decision Tree model was (140 and 23), respectively. Regarding the aim of the study, which is to identify the dropout cases, the model's performance was weak, as the correct predictions for this class label were (28) predictions out of (39) dropout students (Table 1).

Table 1. Decision tree results

	Predicted Dropouts	Predicted Non-dropouts
Actual dropouts	28	11
Actual non-dropouts	12	112
Accuracy	86%	
MSE	0.141	
Recall	0.72	
Precision	0.70	

4.2. Random Forest

The number of correct predictions by the Random Forest model was (156) of (163), which is a good ratio. Regarding the predictions important to the study's objective, there were (93) predictions out of (95) positive true, while the number of incorrect predictions was only (2) predictions, the correct predictions were (37) predictions out of (39) actual dropouts, which is considered an outstanding performance for this model (Table 2).

Table 2. Random Forest results

	Predicted Dropouts	Predicted Non-dropouts
Actual dropouts	37	2
Actual non-dropouts	5	119
Accuracy	96%	
MSE	0.042	
Recall	0.95	
Precision	0.88	

4.3. Support Vector Machine

The SVM algorithm was outperforming in predicting the correct number of dropouts, as it was (36) out of (37). However, the prediction of continuing students was not at the required level compared to the random forest algorithm, as the number of wrong predictions was (10) out of (124), which weakens the performance of the model (Table 3).

Table 3. SVM results

	Predicted Dropouts	Predicted Non-dropouts
Actual dropouts	36	3
Actual non-dropouts	10	114
Accuracy	92%	
MSE	0.079	
Recall	0.92	
Precision	0.78	

4.4. XG Boost

The performance of the XG Boost model was comparable to that of Random Forests on the training set; however, the results on the test set were unsatisfactory, with only 150 correct predictions out of 163 observations. Explaining this divergence in performance is challenging, as the Random Forest algorithm depends on randomness in the trees-building process and functions as a black box, rendering its outcomes difficult to interpret (Table 4).

Table 4. XG Boost results

	Predicted Dropouts	Predicted Non-dropouts
Actual dropouts	34	5
Actual non-dropouts	8	116
Accuracy	92%	
MSE	0.079	
Recall	0.87	
Precision	0.81	

4.5. Logistic Regression

The logistic regression model was developed with an accuracy of 96%, as there was only one erroneous prediction concerning dropouts and merely six erroneous predictions regarding students who persisted in the study. This demonstrates the robustness of the model in terms of predictive capability (Table 5).

Table 5. Logistic Regression results

	Predicted Dropouts	Predicted Non-dropouts
Actual dropouts	38	1
Actual non-dropouts	6	118
Accuracy	96%	
MSE	0.042	
Recall	0.97	
Precision	0.86	

4.6. Discussion

Random Forest and Logistic Regression algorithms outperform the other algorithms, with an accuracy of 96%. However, there was a slight superiority of the Recall metric, which measures the ratio of the model's correct positive predictions to the number of true positive cases, for the Logistic Regression algorithm over the Random Forest algorithm. The situation is reversed with the Precision metric, which measures the ratio of the model's false positive predictions to the number of true positive cases, where Random Forest outperformed Logistic Regression. This indicates that the predictive statistical method, Logistic Regression, outperforms the other algorithms used in this study, because it correctly predicted the number of dropouts (38) slightly more than the number of correct predictions for the Random Forest algorithm (37), which reflects the goal of the study, which is to predict the number of dropouts. The performance of the algorithms on the ROC showed the superiority of the random forest algorithm with a value of 0.99. In contrast, the rest of the algorithms were equal in this measure with a value of (0.98), except for the decision tree algorithm with a value of (0.87), as shown in the Figure 2.

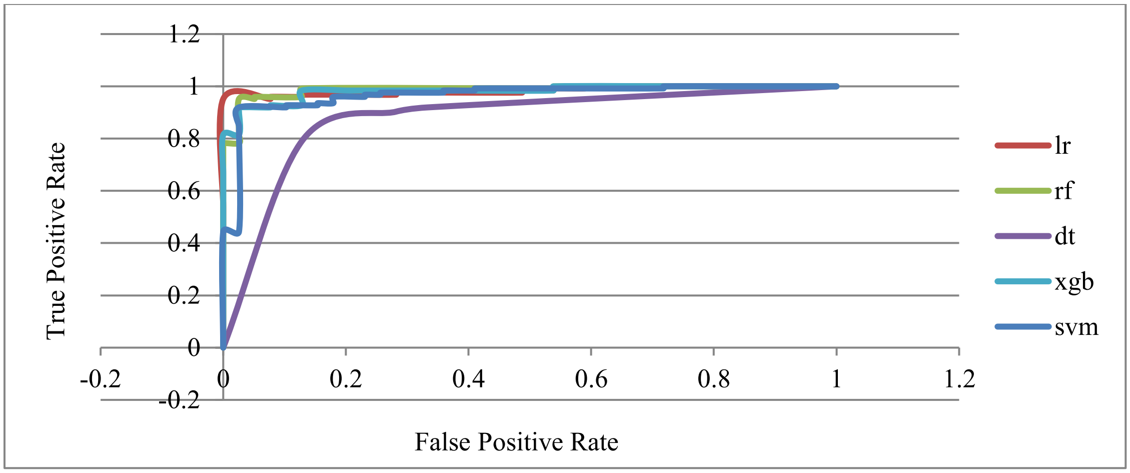
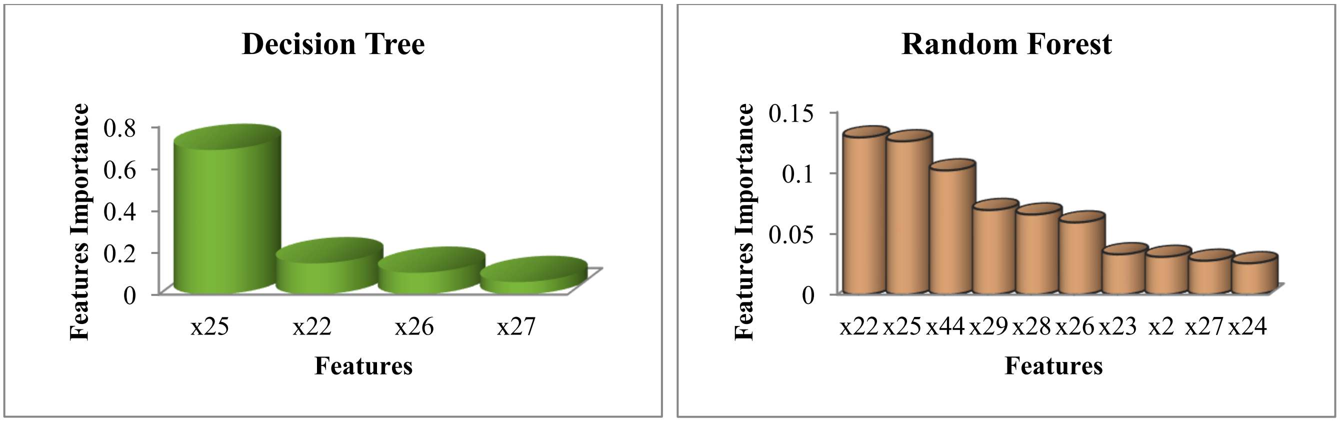


Fig. 2 ROC curve

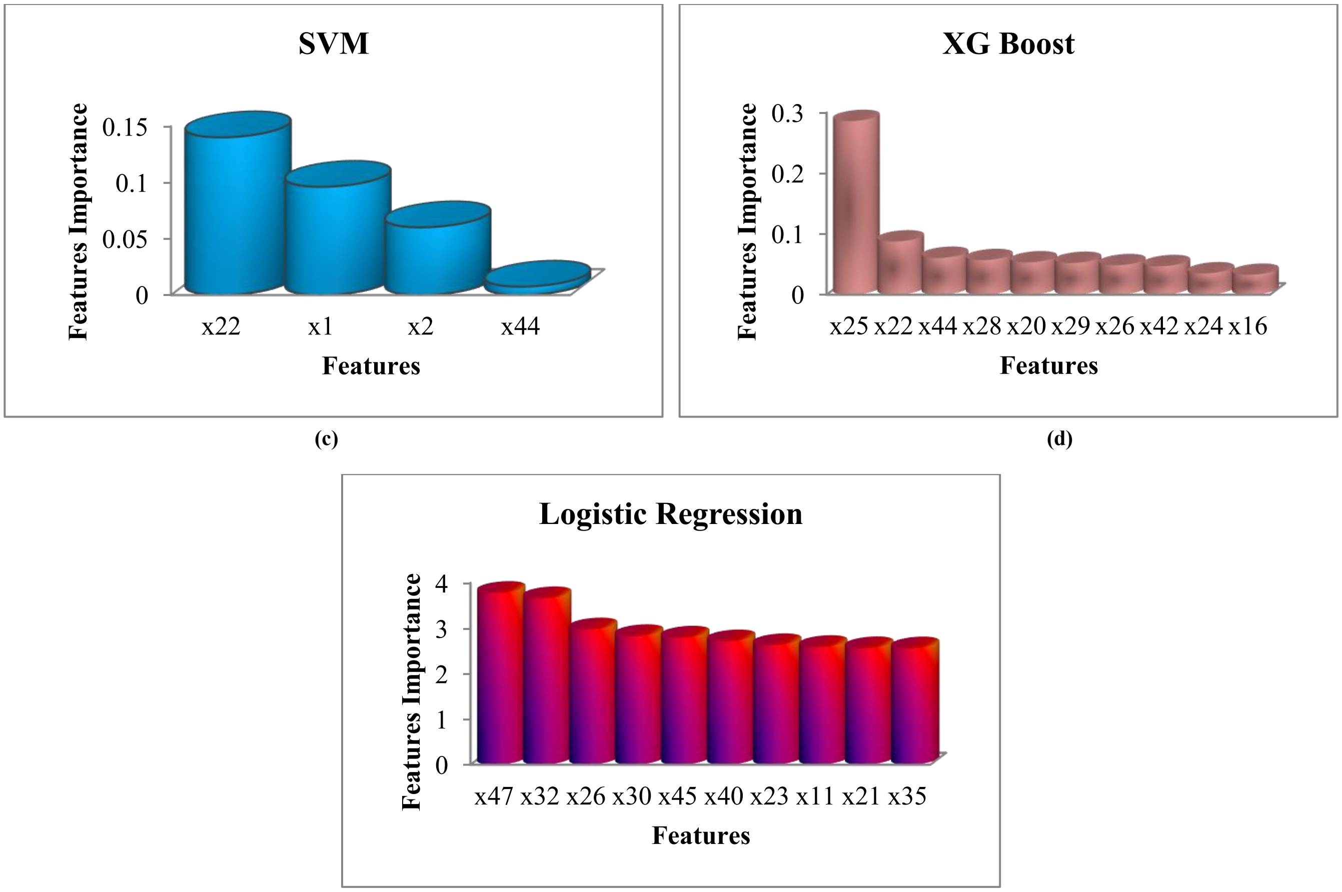
5. Feature Importance Patterns

Upon conducting a thorough analysis of the attributes utilized by each algorithm in the classification of the data set, main patterns of factors were identified that contribute to student attrition, which are: Academic Factors Pattern: Features related to the academic performance have a prominent role in influencing students to make decision of ending their study life, as Random Forest Figure 3(b) and SVM Figure 3(c) showed that the first most important feature is (X_{22}), which represents the previous year's GPA. While the results of the Decision Tree Figure 3(a) and XG Boost Figure 3(d) algorithms showed this factor in second place in terms of the order of the most important influential features. It is also noted that the feature (X_{23}), which represents the number of failure years, ranked as the sixth scale for Logistic Regression Figure 3(e) and seventh for Random Forest. Finally, the difficulty of the curriculum, represented by the feature (X_{20}), ranked as the fifth order for XG Boost. This is consistent with previous studies. Likewise, the other feature related to the academic factors is (X_{25}), which represents the student's participation in classroom activities. The decision tree and XG boost algorithm showed that it is the first of the most important features, while the random forest algorithm ranked second. Institutional Factors Pattern: (X_{44}) factor, which represents the use of teaching aids by the teaching staff, has a significant impact. It came in third place in the Decision Tree and XG Boost results. While the (X_{47}) factor, which represents the performance of the educational counselor in the school, came in first place in the importance list according to the logistic regression algorithm, (X_{30}) and (X_{32}), which represent the structural condition of the school building and restrooms, have an effect on the dependent variable. Socioeconomic Factors Pattern: This pattern has a significant influence on students' decision to discontinue their academic life. Factors (X_{26}), (X_{27}), (X_{28}), and (X_{29}) belong to social factors, which represent the participation in extracurricular, sport activities, student' relation with teaching staff and with their classmates. Similarly, the economic factors exemplified (X_8), which denote the occupation of the student's father, have been identified as one of the most important factors.



(a)

(b)



(c)

(d)

(e)

Fig. 3 Models' Feature Importance

6. Conclusion

Education is an essential component of social and economic development, wherefore countries must reinforce their education systems to achieve success. To enhance the educational sector, problems facing it must be detected and handled. One of the substantial problems is student dropout; therefore, schools and educational institutions must identify students who are at risk of dropping out. This study applied five educational data mining techniques, and the results show that Logistic Regression and Random Forest outperformed other techniques with an accuracy of 96%.

Conflict of Interest

The authors declare no conflicts of interest.

Funding Information

This study was conducted without any external financial support.

References

[1] Dilnavoz Shavkidinova, Feruza Suyunova, and Jasmina Kholdarova, "Education is an Important Factor in Human and Country Development," *Current Research Journal of Pedagogics*, vol. 4, no. 1, pp. 27-34, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Alan Seidman, *College Student Retention: Formula for Student Success*, Rowman & Littlefield, 2024. [Google Scholar] [Publisher Link]

[3] Carlos A. Palacios et al., "Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile," *Entropy*, vol. 23, no. 4, pp. 1-23, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Dorina Kabakchieva et al., "*Research Phases of University Data Mining Project Development*," *Second International Conference S3T*, 2010. [Google Scholar] [Publisher Link]

[5] Jinan Hatem Issa, and Hazri Jamil, "Overview of the Education System in Contemporary Iraq," *European Journal of Social Sciences*, vol. 14, no. 3, pp. 360-386, 2010. [Google Scholar]

[6] Senik Tahir Mahmood, Hakim Qadir Taha, and Mahmood Muhammad Hamza, "Causes of School Dropouts in the Kurdistan Region of Iraq in 2021," *Journal of University of Raparin*, vol. 12, no. 1, pp. 694-712, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] UNICEF, The Cost and Benefits of Education in Iraq: An Analysis of the Education Sector and Strategies to Maximize the Benefits of Education, 2017. [Online]. Available: https://reliefweb.int/report/iraq/cost-and-benefits-education-iraq-analysis-education-sector-and-strategies-maximize

[8] Thiago M. Barros et al., "Predictive Models for Imbalanced Data: A School Dropout Perspective," *Education Sciences*, vol. 9, no. 4, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Marmar Orooji, and Jianhua Chen, "Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques," *18th IEEE International Conference on Machine Learning and Applications*, Boca Raton, FL, USA, pp. 456-461, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hassan M., and T. Mirza, "Prediction of School Drop Outs with the Help of Machine Learning Algorithms," *SIG Science Journal*, vol. 7, no. 7, pp. 253-263, 2020. [Google Scholar] [Publisher Link]

[11] Yuda N. Mnyawami, Hellen H. Maziku, and Joseph C. Mushi, "Comparative Study of AutoML Approach, Conventional Ensemble Learning Method, and KNearest Oracle-AutoML Model for Predicting Student Dropouts in Sub-Saharan African Countries," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1-26, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Kamal Samy Selim, and Sahar Saeed Rezk, "On Predicting School Dropouts in Egypt: A Machine Learning Approach," *Education and Information Technologies*, vol. 28, no. 7, pp. 9235-9266, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] João Gabriel Corrêa Krüger, Alceu de Souza Britto Jr., and Jean Paul Barddal, "An Explainable Machine Learning Approach for Student Dropout Prediction," *Expert Systems with Applications*, vol. 233, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Okan Bulut et al., "Enhancing High-School Dropout Identification: A Collaborative Approach Integrating Human and Machine Insights," *Discover Education*, vol. 3, no. 1, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Siti Rafidah Sariman, Habibah Ab Jalil, and Erzam Marlisah, "Prediction Model of School Drop Out Factors Using Classification Techniques in Selangor," *Malaysian Journal of Social Sciences Humanities*, vol. 9, no. 6, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] M. Awedh, and A. Mueen, "Early Identification of Vulnerable Students with Machine Learning Algorithms," *WSEAS Transactions on Information Scince and Applications*, vol. 22, pp. 166-188, 2025. [Google Scholar]

[17] Jiawei Han, Micheline Kamber, and Jian Pei, *Data* Mining, Third Edition, Morgan Kaufmann: Boston, 2012. [Google Scholar]