Hindawi
Mobile Information Systems
Volume 2022, Article ID 2670562, 9 pages
https://doi.org/10.1155/2022/2670562

Research Article

Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions

Delali Kwasi Dake and Charles Buabeng-Andoh

Department of ICT Education, University of Education, P.O. 25, Winneba, Ghana

Correspondence should be addressed to Delali Kwasi Dake; dkdake@uew.edu.gh

Received 3 September 2022; Revised 12 October 2022; Accepted 22 October 2022; Published 2 November 2022

Academic Editor: Robin Singh Bhadoria

Recently, students dropping out of school at the tertiary level without prior notice or permission has intrigued deep concern among academic authorities, instructors, and counsellors. It has therefore become necessary to understand factors that lead to high attrition rates among learners and identify at-risk students for urgent academic counselling. In providing a proactive response to learner attrition, the study deployed a machine learning algorithm with high model accuracy to predict students' drop-out rates and identify dominant attributes that affect learner attrition and retention. An attrition model was built and validated among support vector machine, decision tree, multilayer perceptron, and random forest algorithms. The machine learning algorithms were tested for accuracy, precision, recall, F-measure, and ROC using the 10-fold and the 5-fold comparative cross-validation techniques. In addition to the cross-validation technique, the chi-square feature selection mechanism was implemented to understand the algorithms' training time and accuracy. The random forest emerged as the best-performing algorithm, with an accuracy of 70.98% and 69.74% for the 10-fold and the 5-fold cross-validation implementations, respectively.

1. Introduction

Students' retention and attrition have become problematic for most higher educational institutions. Globally, the reputation of institutions is tied to enrollment management with a primary focus on reducing the attrition rate [1]. A higher drop-out rate in such institutions indicates a concern and ultimately undermines their global reputation and rankings. According to a 2009 survey by the Organisation for Economic Co-operation and Development (OECD) [2], 31% of students in 19 OECD countries fail to complete their tertiary education. In the OECD report, countries including Hungary, the United States, and New Zealand recorded more than 40% attrition rate, whiles a lower than 24% attrition rate was recorded in Belgium, Denmark, France, Germany, and Japan. The high attrition rate globally shows the inability of tertiary institutions to keep learners in school until they graduate. This global attrition phenomenon is exacerbated further by government policies, institutional culture, and unsavoury student traits [3, 4].

Following the global trend, the attrition rate of students in Africa's numerous educational institutions is significant. A recent report by the Department of Higher Education and Training (DHET) in 2022 shows that 50–60% of first-year students in South Africa drop out across tertiary institutions [5]. Nyoroge et al. [6] research on student attrition in Kenya among thirteen private universities shows a drop-out rate of 37%. A five-year survey of medical students at the Ebonyi State University in Nigeria shows a 7.8% attrition rate during preclinical classes [7]. In Uganda, almost 30% of students who enrol in tertiary institutions never finish their courses [8]. A similar study by Mwenje and Kasowe [9] in open and distance learning at Zimbabwe Open University shows an attrition rate that exceeds 50%. Even though the studies involving attrition rates in Africa are limited due to the unavailability of data, the concluding research findings indicate a more complicated educational issue.

Improving learner retention in tertiary institutions requires proactive, predictive analytics instead of a reactive solution after the problem has occurred. Modelling a

learner's behaviour and thought processes in understanding high-risk factors that lead to attrition is a priority in the solution domain. Students face varying issues from family, finances, relationships, and studies that can easily result in adverse decisions even as they strive to complete their education with good futuristic employment chances. A university graduate's skill set includes critical thinking, creativity, collaboration, information literacy, leadership, technology, and knowledge, which are part of 21st century skills [10]. These learner skills are acquired in the classroom to meet the demands of Industry 4.0 [11]. The consequences of dropping out of school are severe and include reduced employment chances, social stigma, poorer pay, increased crime, and extreme suicidal thoughts [12].

The advent of machine learning (ML) has become a catalyst for analytics and growth in varying domains. In healthcare, ML is currently used to predict the life expectancy of patients with personalised treatment based on health records and family medical history [13]. The boost in e-commerce comes with transactional fraud. The e-commerce sector has seen the application of ML in detecting suspicious transactions with real-time analytics that triggers automatic rejection when unusual patterns are discovered [14, 15]. Loan approval at the banks has seen the application of ML in detecting high-risk applicants and fraudulent paper documents [16]. In the retail industry, ML is presently employed as chatbots that perform scripted functions and leverage natural language processing for customised conversational discussions [17]. Intelligent transportation systems (ITS) have seen the integration of ML for price calculation, ridesharing, ride surge demand locations, and traffic pattern detection [18]. The ITS has seen significant traction recently and remains a key enable in tomorrow's smart cities. The social media environment has a tremendous deployment of ML in engaging billions of users. The application domain of ML in social media varies from detecting new friends, personalised news feed, and targeted adverts [19].

The educational sector has recently taken a positive trajectory in the application of ML for data-driven decision-making. Educational Data Mining (EDM) is a ML niche that identifies hidden patterns in educational data [20]. The Covid-19 pandemic has expanded the academic usage of the Internet. In addition to the traditional face-to-face classroom, massive databases of student-generated educational data have been produced. The most prevalent research in EDM involves students' academic performance prediction, learner assessment modelling, smart tutoring systems, learner attrition modelling, and behaviour modelling [20]. Even though attrition rates in Ghanaian Universities are on the rise [21, 22], implementing a machine learning model for students' drop-out detection is understudied. Based on the above research problem identified, the study is conducted.

The study's objective is to identify dominant factors that can increase the attrition thoughts among learners and predict future drop-out thoughts. In line with the objectives, the following research questions (RQs) guide the study:

RQ1. What are the dominant attributes likely to cause drop-out thoughts among students?

RQ2. Which classification algorithm has the highest accuracy in predicting learner attrition thoughts?

RQ3. To what extent has the chi-square feature selection technique improved the algorithm's accuracy and training time in research question 2?

The main contributions of the study are as follows:

(1) Comparison between the 10-fold and the 5-fold cross-validation techniques in building a classification model for students' drop-out prediction.

(2) Implementing the Chi-Square feature selection mechanism to examine the model's accuracy and training time.

The rest of the paper is organised as follows: Section 2 discusses related literature. Section 3 examines the methodological procedure, data, and algorithms. Section 4 analysis the results and findings from the classification. Section 5 discusses the findings and compares them to the literature. In Section 6, we conclude by summarising the study.

2. Review of Literature

The review section discusses research on student attrition and retention using machine learning algorithms. The specifics of the review include the data sample, machine learning algorithms, accuracy enhancement, and the findings.

The first aspect of literature relates to students' attrition modelling when engaging in online courses. The massive open online courses (MOOCs) are open and available online courses with a diverse variety of topics that expose the learner to relevant educational content. Since the classes are online, the drop-out rates are extremely high [23, 24].

Al-Shabandar et al. [25] deployed a machine learning model to detect at-risk students in danger of early withdrawal from an online course. The study focused on learner engagement levels and motivational attributes that cause students to withdraw from MOOCs. Five machine learning algorithms, including Random Forest (RF), generalised linear model (GLM), gradient boosting machine (GBM), MNET1, and MNET2, were applied to online data from Harvard University, Massachusetts University, and the Open University in building the ideal classifier. The training results show that the MNET1 algorithm has the highest accuracy of 91.57% for full and reduced set features.

Xing and Du [26] built a machine learning model using a deep learning algorithm to predict the retention probability of learners at risk in MOOCs. Data from 3,617 students under varying MOOCs activities, including access to courses, forums, quizzes, module pages, announcements, assignments, and grade books, were used as the main attributes. The drop-out week, which indicates the week learners abandoned the course, is used as the class label in building the classifier. In creating the model, the deep

learning algorithm was compared to the *k*-nearest neighbor (KNN), support vector machine (SVM), and decision tree (DT). The data set was divided 70/30 for training and testing, and the 10-fold cross-validation method was used to avoid model overfitting. The deep learning algorithm has classification accuracy with an average of 95.8% compared to 94.6% for KNN, 93.7% for SVM, and 96.7% for DT. Even though DT has the highest classification accuracy, the deep learning algorithm has a stable test data accuracy of 93.0% compared to 91.5% for DT. The results gave the deep learning algorithm more stability in building the classifier for future prediction of drop-out students.

Figueroa-Canas and Sancho-Vinuesa [27] implemented the tree-based classification models on 197 learners who have enrolled in an online course in Computer Engineering at the Universitat Oberta de Catalunya. The study aims to identify failures and drop-out-prone at-risk students halfway through the semester. The class label for the data set is based on the availability of the learner during the compulsory final exams, and the definition includes two classes, a drop-out and a completer. The conditional tree model with random undersampling is applied to the data set to eliminate bias toward the majority class and increase the classifier's accuracy. Even though only the DT classification algorithm was implemented, the F-measure performance compared with existing literature using similar attributes shows a 76.3% score after implementing the 5-fold cross-validation technique on the model.

Sun et al. [28] compared the recursive neural network with GRU units (GRU-RNN) algorithm to XGBoost, Gradient Boosting Decision Tree (GBDT), and the RF base algorithms to predict the attrition rate of learners in a MOOC course. The training data for the study consists of 10278 learners, whiles the test data has 2568 students. The study utilised the maximum input sequence feature in the RNN model and tested the max-length of 500 and 1000 RNN against the base algorithms. As increasing weekly data samples are trained with the proposed GRU-RNN algorithm to the compared base algorithms, the RNN classification accuracy increases. The 1000 GRU-RNN has a significant performance accuracy compared to the 500 GRU-RNN and the other base algorithms.

The second aspect of literature involves the traditional tertiary institution when teaching and learning occur in person. The discussion of the literature on the conventional campus involves factors that affect learner attrition in the classroom and on campus.

Solis et al. [29] analysed the accuracy of RF, NN, SVM, and Logistic Regression (LOGIC) on 80,527 records of students from the Instituto Technologico de Costa Rica (ITCR) University. The first class variable are drop-outs who have not graduated and have not enrolled in school for two years. The second class variable are active students who have graduated successfully. After implementing the 5-fold cross-validation technique across all the algorithms, the sensitivity, kappa, and true positive results were analysed. The RF emerged as the best classifier with a true positive percentage of 94% and a sensitivity of 93%. The kappa statistics of RF has a significant value of 0.85 compared to 0.84 for SVM, 0.84 for NN, and 0.84 for LOGIC.

Lee and Chung [30] trained the RF, boosted decision tree (BDT), RF with synthetic minority oversampling techniques (SMOTE), and BDT with SMOTE classifiers on 165,715 data samples obtained from the National Educational Information System (NEIS) in South Korea. The study aims to compare the sensitivity results of the classification algorithms since it represents the fraction of actual drop-out learners correctly predicted. From the 165,715 data instances, 1348 students were identified as drop-outs based on primary negative reasons, including poor academic scores, school rule violation, strict rules in school, committee requests for expulsion, and relationships with teachers and friends. A split ratio of 80% to 20% for training and testing datasets was implemented to evaluate the classifiers during preprocessing. The classification results show that the BDT algorithm with the area under the ROC Curve (AUC) value of 0.898 outperformed other algorithms and was utilised as the model for detecting early attrition among students.

Kemper et al. [31] compared the logistics regression (LR) and the DT algorithms to predict the drop-out of 3,176 data samples from the Karlsruhe Institute of Technology (KIT). The underrepresented minority class of 620 drop-outs against 2556 successful graduation created a biased classification scenario that was solved using the SMOTE technique. The 10-fold cross-validation, stopping, and pruning techniques were implemented to avoid over and underfitting of the classification model. The DT algorithm has the highest performance accuracy compared to the logistic regression algorithm.

Palacios et al. [32] separately predicted student retention among first, second, and third year students using DT, LR, RF, SVM, naïve Bayes (NB), and kNN classification algorithms on 6656 data samples from the Catholic University of Maule. The features for modelling the classifier were sought under dominant attributes, including university performance, high school performance, financial indicators, socioeconomic index, geographic origin, and demographic background. After implementing the 10-fold cross-validation and SMOTE, a predictive model for first, second, and third year students was analysed. For the first-year model results, the RF algorithm ranked superior as the highest-performing algorithm with an F-Measure score of 0.947 compared to a 0.910 score for the DT algorithm. For the second-year students, RF has an increased F-Measure value of 0.975 compared to 0.966 of KNN. RF increased again in F-Measure score to 0.984 but levelled in performance with KNN for third year students.

Perez et al. [33] compared DT, LR, and NB classification algorithms using 802 instances of data in modelling the retention rate of learners at a private university in Bogota, Columbia. The attributes for data collection included minimum demographic, expected graduation date, accessible financial aids, and official transcript records. A drop-out class type is determined by the failure to complete an undergraduate degree within six years after the start day of enrollment. The experimental results show a higher score for

DT with an AUC value of 0.94 compared to 0.92 for 0.92 LR and 0.87 for NB.

Hegde and Prageeth [34] implemented the NB classification algorithm in *R* language to predict the retention of students using 24 attributes after feature selection. The attributes were divided into demographic, psychological, academic performance, social media usage, and social integration on campus. The class label displays a survey response on whether the student will prefer to continue the course or drop-out based on personal developments in school over the 24 attributes. Even though the result of the NB classifier was not compared, the model has a significant accuracy of 72%.

Table 1 summarises the reviewed literature and highlights the limitations of previous studies in which feature selection mechanisms were not used. Furthermore, the 10-fold cross-validation was not compared to the 5-fold but was implemented in isolation.

3. Methodology

As illustrated in Figure 1, the study modified the Cross Industry Standard Process Mining (CRISP-DM) methodology [35] by replacing the business understanding with the problem definition and inculcating the classification algorithms into the diagram. The CRISP-DM is a cyclical process in a data mining project that consists of business understanding, data understanding, data preparation, modelling, evaluation, and deployment stages. The CRISP-DM approach is primarily centred around a big data engine with attributes and tuples.

3.1. Students' Data. The research data was sorted from students at the south campus of the University of Education, Winneba. Data from students in the ICT Education, Biology Education, Integrated Science Education, and Math Education departments was explicitly collected. The study utilised the convenient sampling approach, a nonprobability method for data collection. The convenient sampling method was adopted due to the respondents' immediate availability and accessibility at the university's south campus. Google form was administered to students from year one to year three from the mentioned departments, and the research objectives were clearly stated. In addition, the respondents were mandated to agree to an ethics consent form before filling out the questionnaire. Throughout the CRISP-DM process, the nondisclosure and privacy of respondents' data were adhered to strictly. In adhering to the confidentiality of data, no information in the questionnaire could be traced back to the respondents. A total of 1239 responses were received under the personal and family biodata, the senior high school (SHS) tracker, the university tracker, and the decision tracker.

3.1.1. Attribute Selection based on Student's Attrition Problem. The attribute selection for responses was linked closely to student attrition. The attributes utilised were based on factors that could influence learners' drop-out thoughts from a programme. As shown in Table 2, the attributes were grouped based on similarity in family traces and academic paradigms of the respondent.

3.2. Data Understanding. The dataset, as shown in Table 2, was collected under four major sections with twenty-three attributes. Under the personal and family biodata, 65.1% of the respondents are males, whiles 34.9% are females. Precisely 93.9% of the respondents have siblings, with 77.3% of parents being accommodating. Under the SHS tracker, 84.6% of the respondents attended mixed schools, with 70.8% boarding status. According to the university tracker, 78.8% of lecturers are accommodating and mostly encourage their students during lesson periods. About 83.9% of learners strongly suggest that most lecturers should adopt new teaching strategies and expect more advanced facilities on campus for progressive teaching and learning. The responses also show that getting accommodation on campus is difficult, and paying fees every academic year has become financially strenuous. The data reveal that 21.8% of the students have an excellent cumulative grade point average (CGPA), 71.4% have a good CGPA, and 6.8% have a poor CGPA. The data also shows that 40.4% of the respondents are in level 100, 40% in level 200, and 19.5% in level 300.

3.3. Data Preparation. In the data preparation and cleaning phase, unrelated data inconsistent with missing values are removed prior to classification. In building the classifier, a total of 1,239 responses for modelling was utilised. The data set was composed of 100% valid data, making it an optimal data set for the classification model. The decision tracker, which represents the class label, has two values, "Yes, I want to quit" and "No, I will never quit." The "Yes, I want to quit" class category are learners who have thought about dropping out of school based on difficulties. On the other hand, "No, I will never quit" class category are students who have never thought about stopping school, no matter the problem.

3.4. Classification Algorithms. The support vector machine (SVM) algorithm identifies a hyperplane that uniquely classifies the data points in an *N*-dimensional space. The hyperplane in SVM is a decision boundary that segregates the data set into classes using vectors. The SVM algorithm is one of the best-performing classification algorithms compared to other algorithms in building multiple applications [36, 37].

The Random Forest (RF) algorithm [38] is an ensemble of the decision tree (DT) algorithm and is trained using the bagging method to increase classification accuracy. In the RF algorithm, the right and wrong class for classification is determined using the margin function. In addition to classification, the RL algorithm has seen immense deployment in solving regression-related problems.

Multilayer Perceptron (MLP) [39] is a neural network with input, hidden, and output layers. MLP learns a relationship between linear and nonlinear datasets as part of the feed-forward neural network functionality. The MLP

Table 1: Summary of literature review.

Review	Mode	K-fold cross-validation	Feature selection method	Algorithms compared	Best classifier and metrics reported
[25]	Online MOOCs	Not stated	None	RF, GLM, GBM, MNET1, and MNET2	MNET1, accuracy = 91.57%
[26]	✓	10-Fold	✓	Deep learning, KNN, SVM, and DT	Deep learning, accuracy = 95.8%
[27]	✓	5-Fold	✓	DT	DT, f-measure = 76.3%
[28]	✓	Not stated	✓	GRU-RNN, XGBoost, GBDT, and RF	GRU-RNN, accuracy not stated
[29]	Traditional classroom	5-Fold	✓	RF, NN, SVM, and LOGIC	RF, accuracy = 94%
[30]	✓	Not stated	✓	RF and BDT	BDT, ROC value = 0.898
[31]	✓	10-Fold	✓	LR and DT	DT, accuracy not stated
[32]	✓	10-Fold	✓	DT, LR, RF, SVM, NB, and kNN	RF, f-measure = 0.975
[33]	✓	Not stated	✓	DT, LR, and NB	DT, ROC value = 0.94
[34]	✓	Not stated	✓	NB	NB, accuracy = 72%

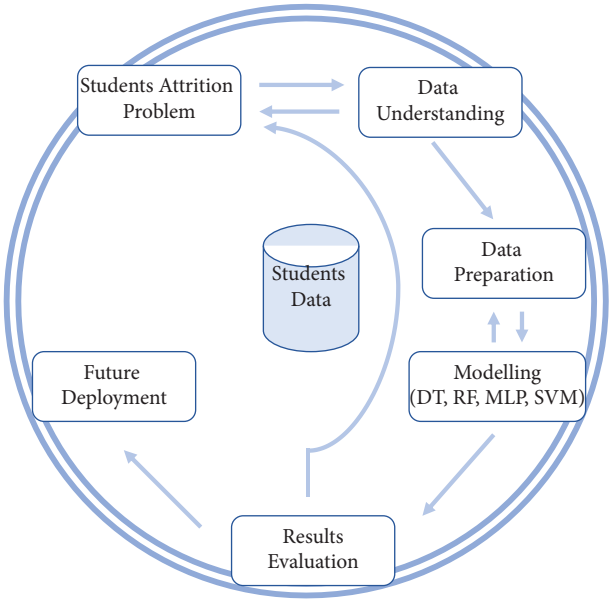

Figure 1: Modified KDD methodology.

also uses the backpropagation technique to minimise the cost function by iteratively adjusting the weights in the network.

The decision tree (DT) algorithm [40] is a tree-structured classifier with branches representing decision rules that internodes the dataset's features. In a DT algorithm, the leaf node represents the class label outcomes with no additional segregation. Using the attribute selection measure, the DT algorithm recursively generates tree nodes until a leaf node is reached.

4. Results and Analysis

In the simulation of the results, the Waikato Environment for Knowledge Analysis (WEKA) software, which offers a range of machine learning algorithms, is utilised in building a predictive model for the data set. The WEKA stable version 3.8.6 with classification and regression extensions was installed via the package manager.

4.1. Dominant Attributes. To respond to RQ1, the SMOTE instance supervised filter in WEKA was first applied to the minority class label to increase the instances by 50%, as shown in Figure 2. The SMOTE oversampling technique addresses the concern of data imbalances and prevents overfitting. Applying the SMOTE technique increased the minority class significantly and created a balanced data set for classification.

The feature selection mechanism in machine learning orders attributes based on a correlation score between the attributes and the class label. The data set instances for the study have nominal inputs with a nominal class label. The chi-square [41] feature selection methods have seen outstanding results for nominal input and output attribute instances. Table 3 depicts the top ten dominant attributes that affect learner drop-out thoughts using the chi-square attribute evaluator and the ranker search method.

The results indicate a strong influence of lecturers' encouragement, SHS counselling, the flexibility of lecturers, CGPA, accommodation difficulty, adoption of new learning strategies, fee payment difficulty, guardian education, and participation during the lesson as dominant attributes. The top three attributes, as shown in Table 3, are lecturers' encouragement during lesson time (correlation score of 33.18), SHS guidance on the program to select at the university (correlation score of 32.94), and lecturers' strictness or flexibility (correlation score of 32.82).

4.2. Classification Accuracy. In determining the best classification algorithm which answers RQ2, the SVM, RF, DT, and MLP supervised learning algorithms were used to model the dataset in WEKA. Comparatively, the 10-fold and 5-fold cross-validation techniques were implemented to ascertain the performance of the classification algorithms. The k-fold cross-validation technique divides a dataset based on the number of folds with an iterative division between the training and test data. As depicted in Table 4 and Figure 3, the 10-fold cross-validation significantly performs better than the 5-fold cross-validation technique. The Random Forest algorithm has the highest accuracy of 70.98%

Table 2: Attributes for student attrition modelling.

Section	Attributes | options
Personal and family biodata	Gender {male; female} Age {18–22; 23–25; 26 or older} My guardians/parents are {strict; accomodating} Do you have siblings {yes; no} What is the social class of your family {Upper; Middle; Lower} My guardians/parents are (educational terms) {highly educated, moderately educated; uneducated}
Senior high school (SHS) tracker	SHS school category {Single; Mixed} Residential status in SHS {Day; Boarding} In SHS, were you counselled on the programme to select at the university
University tracker	The majority of my lecturers are {Yes; No} Do you think most of your lecturers should adopt new teaching strategies for you to understand the courses in detail {Yes; No} Do you think campus facilities for students are standard enough for excellent academic work {Yes; No} During lecturers, I prefer to be {Active, answer questions; Passive, be quiet} Have you ever been counselled by the university's counselling unit before? {Yes; No} Do you have friends on campus {Yes; No} Residential status in the university *{Hostel; Hall; Home}* Accommodation status {One in a room; Two in a room; Three or more in a room} Do you find it financially difficult to pay your fees every academic year {Yes; No} Is it difficult to get accommodation every academic year? {Yes; No} So far, how will you rate your overall academic performance { Excellent (CGPA 3.5 and above); Good (CGPA 2.5 to 3.4); Poor (CGPA 2.4 and below)} What is your current level {100; 200; 300}
Decision tracker	Have you ever considered/thought of stopping your programme of study at the university? {Yes, I want to quit; No, I will never quit}

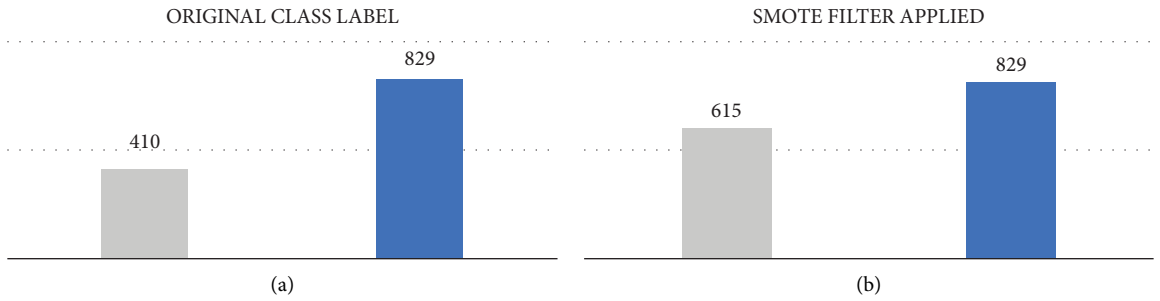

Figure 2: Original class label and SMOTE filter applied. (a) Original class label. (b) SMOTE filter applied.

compared to the decision tree accuracy of 65.03%, the MLP of 63.92%, and the SVM of 62.60%.

In analysing the confusion matrix to validate the performance of the classification algorithms, the precision, recall, f-measure, and receiver operating characteristics (ROC) curve results of the supervised learning algorithm were analysed. As depicted in Table 5, the RF algorithm maintained dominance with a precision of 0.708 and a ROC value of 0.771 compared to the DT algorithm, which has a precision value of 0.649 and 0.653 ROC value. The ROC curve indicates the true positive (TP) and false positive (FP) measure of the classification model based on the actual and the predicted class of the confusion matrix. A ROC curve value closer to 1 shows a good measure of separability between the positive and negative classes. As shown in Table 5, the RF ROC value of 0.771 is the highest and indicates that the model built with the RL algorithm has a 77.1% chance of correctly distinguishing between "Yes, I want to quit" and "No, I will never quit" class label among learners.

Table 3: Attributes ranking using chi-square.

Attributes	Rank
Lecturers encouragement	33.18
SHS counselling	32.94
Majority of lecturers	32.82
CGPA	24.88
Accommodation difficulty	23.03
New learning strategies	21.23
Fees at university difficulty	19.77
Social class	14.10
Guardian education	11.65
Class hours	11.47

4.3. Classification Accuracy with Chi-Square Feature Selection. In response to RQ3, the feature selection mechanism is primarily implemented to remove weaker attributes and maintain more vital features to improve classification accuracy. As already depicted in Table 3, the chi-square feature mechanism is adopted, and the top ten relevant attributes are

Table 4: Accuracy of the classifiers.

K-fold cross validation	SVM	RF	DT	MLP
10-Fold	62.60	70.98	65.03	63.92
5-Fold	62.61	69.74	63.23	62.40

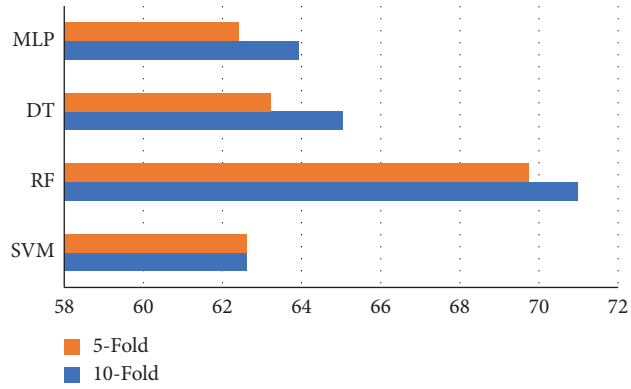

Figure 3: Classifier accuracy.

Table 5: 10-fold cross-validation.

Classifier	Precision	Recall	F-measure	ROC
SVM	0.623	0.626	0.596	0.587
RF	0.708	0.710	0.707	0.771
DT	0.649	0.650	0.649	0.653
MLP	0.642	0.639	0.640	0.662

Table 6: 10-fold cross-validation with chi-square feature selection.

K-fold cross validation	SVM	RF	DT	MLP
10-Fold	60.73	63.09	61.98	61.43

Table 7: Model training time.

K-fold cross validation	SVM (s)	RF (s)	DT (s)	MLP (s)
10-Fold–Training time without feature selection	0.2	0.23	0.01	6.27
10-Fold–Training time with feature selection	0.1	0.16	0.01	1.64

maintained for classification. After implementing the 10-fold cross-validation on the chi-square data, the classification algorithms' accuracy decreased across all the algorithms, as indicated in Table 6 but slightly improved the model's training time, as shown in Table 7. The training time for SVM improved from 0.2 s to 0.1 s, that of RF from 0.23 s to 0.16 s, and MLP from 6.27 s to 1.64 s. DT algorithm has no change in its training time.

5. Discussion of Findings

Student retention modelling in higher educational institutions using a machine learning approach is primarily determined by the attributes for prediction and the algorithms implemented. The study's findings using the chi-square feature selection mechanism listed lecturers' encouragement, SHS counselling, and lecturers' accessibility and flexibility as the most relevant attributes. The feature selection mechanism help identify attributes that have a high correlation to learner drop-out thoughts. The features that rank high after applying the appropriate feature selection techniques affect the class label significantly. Feature selection application on data attributes is suitable for the academic counselling of students since counsellors will identify significant features likely to cause attrition or retention among learners. Developing a highly accurate predictive model for student attrition is based on the supervised learning algorithms deployed. The performance of classification algorithms is linked closely to the type of dataset. The literature reviewed by Solis et al. [29] and Palacios et al. [32] compared the RF method to various classification algorithms, and the classification accuracy results show RF as the best-performing algorithm. The RF in this study also emerged as the best-performing algorithm for the dataset with a 70.98% accuracy using the 10-fold cross-validation technique. The DT algorithm ranked as the second best-performing algorithm with an accuracy of 65.03%, while the SVM's 62.60% was the worst-performing algorithm. The

results from research question 3 also indicate that feature selection implementation decreased the classification algorithms' accuracy but with a better model training time. In this research, the feature selection mechanism restricted to the top ten performing attributes did not increase the algorithm's accuracy. For the dataset, classification accuracy increased with larger sample sizes and attributes.

6. Conclusion

This research focused on three aspects: (i) the use of feature mechanism to list significant attributes for learner attrition modelling; (ii) building a learner attrition predictive model using a classification algorithm with the highest accuracy; and (iii) understanding the impact of the feature selection method chi-square on the accuracy of the algorithm. Since learner attrition has become problematic to academic authorities, implementing the model for future prediction will help identify learners with attrition thoughts for immediate academic counselling. The model primarily forms the bases for future prediction of learner attrition thought among students at the University of Education, Winneba, Ghana. Given a test data, the RF-based model has a 77.1% chance of separability between the two class labels, "Yes, I want to quit" and "No, I will never quit."

In building the model, the RL, SVM, MLP, and DT algorithms were compared using the 10-fold and the 5-fold cross-validation techniques. Before comparing the algorithms, the SMOTE technique was utilised during the data preprocessing stage to increase the minority class by 50% for a more balanced dataset. The chi-square feature selection mechanism was also utilised to sort relevance attributes with a high correlation value to the class label. Among the supervised algorithms compared, the RF algorithm performed best with an accuracy of 70.98% and 69.74% for the 10-fold and the 5-fold cross-validation implementations, respectively. The precision, recall, F-measure, and ROC results also indicate the RF algorithm's dominance compared to other tested algorithms.

Data Availability

Data available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010.

[2] UNESCO, "School drop-out: patterns, causes, changes and policies," 2010, https://unesdoc.unesco.org/ark:/48223/pf0000190771.

[3] D. Whitehead, "Do we give them a fair chance? Attrition among first-year tertiary students," *Journal of Further and Higher Education*, vol. 36, no. 3, pp. 383–402, 2012.

[4] D. Whitehead, "Do we give them a fair chance? Attrition among first-year tertiary students," *Journal of Further and Higher Education*, vol. 36, no. 3, pp. 383–402, 2012.

[5] B. Dyomfana, "Half of University Students Drop Out in First Year," 2022, https://www.careersportal.co.za/news/half-of-university-students-drop-out-in-first-year.

[6] M. M. Njoroge, T. Wang eri, and C. Gichure, "Examination repeats, semester deferments and dropping out as contributors of attrition rates in private universities in nairobi county Kenya," *International Journal of Educational Research*, vol. 4, no. 3, pp. 225–240, 2016.

[7] O. A. Anyanwu and G. E. Anyanwu, "Five-year survey of medical student attrition in a medical school in Nigeria: a pilot study," *Advances in Medical Education and Practice*, vol. 1, pp. 53–57, 2010.

[8] C. Businge, "Universities: Student Drop-Out Rates Alarming," 2019, https://www.newvision.co.ug/new_vision/news/1502587/universities-student-drop-rates-alarming.

[9] S. Mwenje and R. Kasowe, "Student involvement in enhancing student, retention, persistence and success in open and distance learning at Zimbabwe open university," *African Educ. Res. J*, vol. 1, no. 1, pp. 46–50, 2013.

[10] L. I. González-Pérez and M. S. Ramírez-Montoya, "Components of education 4.0 in 21st century skills frameworks: systematic review," *Sustainability*, vol. 14, no. 3, pp. 1–31, 2022.

[11] M. Ghobakhloo, "Industry 4.0, digitization, and opportunities for sustainability," *Journal of Cleaner Production*, vol. 252, Article ID 119869, 2020.

[12] C. Campbell, "The socioeconomic consequences of dropping out of high school: evidence from an analysis of siblings," *Social Science Research*, vol. 51, pp. 108–118, 2015.

[13] K. Aggarwal, M. M. Mijwil, S. Alomari, M. Gök, A. M. Z. Alaabdin, and S. H. Abdulrhman, "Has the future started? The current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal for Computer Science and Mathematics*, pp. 115–123, 2022.

[14] V. De, J. De, B. denAkker, J. Smith, O. Thuong, and L. Bernardi, "Machine learning for fraud detection in E-commerce: a research agenda," *Deployable Machine Learning for Security Defense*, CCIS, vol. 1482, pp. 30–54, 2021.

[15] C. Tejasri, C. H. Sai, U. Aryan, D. Deekshith, A. Chintu, and T. S. Reddy, "Fraud detection in E-commerce using machine learning," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 3, pp. 2206–2211, 2021.

[16] W. Fang, X. Li, P. Zhou, J. Yan, D. Jiang, and T. Zhou, "Deep learning anti-fraud model for internet loan: where we are going," *IEEE Access*, vol. 9, pp. 9777–9784, 2021.

[17] M. R. Kumar, J. Venkatesh, and A. M. J. M. Z. Rahman, "Data mining and machine learning in retail business: developing efficiencies for better customer retention," *Journal of Ambient Intelligence and Humanized Computing*, Article ID 0123456789, 2021.

[18] T. Yuan, W. Rocha Neto, C. E. Rothenberg, K. Obraczka, C. Barakat, and T. Turletti, "Machine learning for next-generation intelligent transportation systems: a survey," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 4, pp. 1–35, 2022.

[19] T k. Balaji, C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: a survey," *Computer Science Review*, vol. 40, Article ID 100395, 2021.

[20] S. Ventura, "Educational data mining and learning analytics: an updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, pp. 1–21, 2020.

[21] "Reducing distance learners' attrition rate at the university of cape coast: tutors'/students' perception," *International Journal of Learning and Development*, vol. 3, no. 3, p. 214, 2013.

[22] G. Modupe, "50% Attrition Rates in Africa's Tertiary Institutions - Ghana Education Minister," 2019, https://tribuneonlineng.com/50-attrition-rates-in-africas-tertiary-institutions-ghana-education-minister/.

[23] L. N. M. Bezerra and M. T. da Silva, "A review of literature on the reasons that cause the high drop-out rates in the MOOCS," *Espacios*, vol. 38, no. 5, p. 11, 2017.

[24] X. Lu, S. Wang, J. Huang, W. Chen, and Z. Yan, "What decides the dropout in MOOCs?" *Database Systems for Advanced Applications*, LNCS, vol. 10179, pp. 316–327, 2017.

[25] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, "Detecting at-risk students with early interventions using machine learning techniques," *IEEE Access*, vol. 7, pp. 149464–149478, 2019.

[26] W. Du and D. Du, "Dropout prediction in MOOCs: using deep learning for personalized intervention," *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547–570, 2019.

[27] J. Figueroa-Canas and T. Sancho-Vinuesa, "Early prediction of dropout and final exam performance in an online statistics course," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 15, no. 2, pp. 86–94, 2020.

[28] D. Sun, Y. Mao, J. Du, P. Xu, Q. Zheng, and H. Sun, "Deep learning for dropout prediction in MOOCs," in *Proceedings of the 2019 Eighth International Conference on Educational Innovation through Technology (EITT)*, pp. 87–90, Biloxi, MS, USA, October 2019.

[29] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez, "Perspectives to predict dropout in university students with machine learning," in *Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, San Carlos, Costa Rica, July 2018.

[30] J. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Applied Sciences*, vol. 9, no. 15, Article ID 3093, 2019.

[31] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: a machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, 2020.

[32] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "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.

[33] B. Perez, C. Castellanos, and D. Correal, "Applying data mining techniques to predict student dropout: a case study," in *Proceedings of the 2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, Medellin, Colombial, May 2018.

[34] P. Hegde and P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," in *Proceedings of the 2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 694–699, Icisc, Coimbatore, India, January 2018.

[35] R. Wirth, J. Hipp, and D. M. Crisp, "Towards a standard process model for data mining," in *Proceedings of the fourth international conference on the practical application of knowledge discovery and data mining, 29-39*, Article ID 24959, pp. 29–39, Oldenburg, Germany, Feburary 2000.

[36] Y. Zhang, "Support vector machine classification algorithm and its application," *Communications in Computer and Information Science*, CCIS, vol. 308, pp. 179–186, 2012.

[37] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.

[38] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: random forest," *Information Computing and Applications*, LNCS, vol. 7473, pp. 246–252, 2012.

[39] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: architecture optimization and training," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, pp. 26–30, 2016.

[40] B. R. Patel and K. K. Rana, "A survey on decision tree algorithm for classification," *Ijedr*, vol. 2, no. 1, pp. 1–5, 2014.

[41] M. Zaffar, M. A. Ahmed, K. S. Savita, and S. S. H. Sajjad, "A study of feature selection algorithms for predicting students academic performance," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 541–549, 2018.