Global challenges of students dropout: A prediction model development using machine learning algorithms on higher education datasets

Meseret Yihun Amare [1,*], and *Stanislava* Simonova[2]

[1]University of Pardubice, Faculty of Economics and Administration, Institute of System Engineering and Informatics, Studentská 84, 532 10 Pardubice, Czech Republic
[2]University of Pardubice, Faculty of Economics and Administration, Institute of System Engineering and Informatics, Studentská 84, 532 10 Pardubice, Czech Republic

Abstract

Research background: In this era of globalization, data growth in research and educational communities have shown an increase in analysis accuracy, benefits dropout detection, academic status prediction, and trend analysis. However, the analysis accuracy is low when the quality of educational data is incomplete. Moreover, the current approaches on dropout prediction cannot utilize available sources.
Purpose of the article: This article aims to develop a prediction model for students' dropout prediction using machine learning techniques.
Methods: The study used machine learning methods to identify early dropouts of students during their study. The performance of different machine learning methods was evaluated using accuracy, precision, support, and f-score methods. The algorithm that best suits the datasets for these performance measurements was used to create the best prediction model.
Findings & value added: This study contributes to tackling the current global challenges of student dropouts from their study. The developed prediction model allows higher education institutions to target students who are likely to dropout and intervene timely to improve retention rates and quality of education. It can also help the institutions to plan resources in advance for the coming academic semester and allocate it appropriately. Generally, the learning analytics prediction model would allow higher education institutions to target students who are likely to dropout and intervene timely to improve retention rates and quality of education.

Keywords: *Globalization; Prediction; Machine Learning; Learning Analytics; Global challenges.*

JEL Classification: *C60; C80; D80; I23; M15; O30*

* Corresponding author: yihunm@gmail.com

1 Introduction

The growth of data in research and education communities has shown an increase in analysis accuracy, benefits dropout detection, academic status prediction, and trend analysis. However, the analysis accuracy is low when the quality of educational data is incomplete. (Chen et al., 2017)

Higher education institutions face the challenge of low student retention rates and an increased number of dropouts. (Zhang and Rangwala, 2018) Therefore, higher education institutions need to develop learning analytics systems to find students at-risk of failing at earliest possible time and provide timely intervention. The main principle of learning analytics is identifying at-risk students and given timely intervention based on the results of student behavior investigation. (Huang et al., 2020) Early prediction of students' academic status helps to intervene early and act accordingly to improve learning outcomes. It helps increase graduation rates by appropriately helping students, helping higher education policymakers, monitoring the efficiency and effectiveness of teaching-learning activities, giving critical feedback to students and teachers, and modifying learning activities. (Ofori et al., 2020)

A practical prediction algorithm results in a high prediction accuracy of the students' achievement; identify the low-performing students at the beginning of the learning process. However, to achieve these objectives, a large volume of student data must be analyzed and predicted using various machine learning models. Online learning environments such as Moodle systems and Student Information System (SIS) assist the learning analytics paradigm by providing datasets for further analysis and reporting. Using the available data from online learning systems, it can be possible to support decision-making in students' learning process and use it to timely intervene students who are likely to drop out to improve their respective performances.

The work by Zhang and Rangwala (2018) has discovered key features using the traditional statistical methods to identify at risk of dropping out students from their study. Machine learning offers an advantage over traditional forms of statistical analysis, emphasizing predictive performance over provable theoretical properties. Machine learning methods are used to develop prediction models and plot patterns using available data, which is helpful in decision-making (Hussain et al., 2018). Machine learning is a software modeling technique of self-learning systems that makes meaningful inferences from data or experiences with mathematical and statistical operations (Alpaydin, 2020). An effective prediction of students' dropout during the early stages can provide course instructors with timely intervention. This helps to reduce the underlying problem by implementing rapid and consistent intervention mechanisms.

The first section of this paper provides introductory information about the prediction model in global higher education institutions. The second section discusses related studies on students' performance prediction models and associated techniques. Comparison of previous works has been discussed in selected studies. The third section describes the methodology used for this study. In addition, the experimental procedures, data preprocessing methods and the steps involved in developing proposed predictive model were discussed in this section. The fourth section provided the study results and discussed the performance measurements to identify the winning prediction model.

2 Literature Review

This study aims to predict student academic status using Random Forest, Naïve Bayes Classifier, Logistic Regression, and Decision Tree machine learning methods. The recent research by Ofori et al. (2020) suggested that several Machine learning models could be adapted to analyze the data, such as clustering, classification, and association rules mining based on the suitability of collected data and aims of the data analytics process.

Students' dropout is one of the most complicated and challenging problems worldwide that students and global institutions face. Therefore, effectively predicting students' dropout could help alleviate social and economic costs. (Chacha et al., 2019)

Recent studies have confirmed that Machine Learning methods are used to predict students at risk of failing and dropout rates to improve their performance during their studies. (Albreiki et al., 2021)

Several Machine learning models have been created to predict student dropout based on algorithms such as decision trees, neural networks, random forests, vector support machine, logistic regression. Machine learning takes advantage of traditional statistical analysis, stressing predictive performance over verifiable theories. (Ofori et al., 2020) Recently, Machine learning methods are used commonly to predict students' achievement in their academics (Emirtekin et al., 2020). However, these models' effectiveness varies mainly due to the type and size of datasets used in the model, feature selection strategies, performance measurement criteria, and experimental procedures. In addition, different kinds of literature used other techniques and selected predictive variables. The following table summarizes recent studies related to students' academic status achievement, dropout rate prediction, identification of at-risk students, and evaluation criteria that measure the effectiveness of the prediction model.

Table 1. Comparison of related literatures.

Ref.	Description and methodology used	Performance	Limitation
Gray and Perkins (2019)	Student's dropout prediction at the 3rd week of the semester using Nearest Neighbor.	97% accuracy	Difficult to apply to other universities.
Imran et al. (2019)	Predicting & explaining students dropout using Feed-forward deep neural networks.	>90% accuracy	No possibility for intervention beforehand.
Adnan et al. (2021)	Developed prediction model for students at-risk of dropout at a different percentage of course length using Random Forest, feature engineering techniques.	92% precision, 91% Recall, 91%, F-score, & 91% accuracy.	Difficult to implement as effectiveness of study could vary across the semester
Kabathova and Drlik (2021)	Compared the performance of several machine learning classifiers	Accuracy ranging from 77% to 93%	Decisive parts of assessment such as projects were missed.
Alyahyan and Dustegor (2020)	Guidelines for determining students' success using data mining process.	NA	Difficult to measure accuracy
Alipio (2021)	Developed a prediction model based on psychological factors expectancy-value beliefs, & academic performance using conceptual model and descriptive statistics	Developed conceptual model for psychological factors & expectancy value beliefs	Datasets were collected only from freshmen students
Waheed et al. (2020)	Developed a deep Artificial Neural Network (ANN) on selected features using clickstream data extracted from the virtual learning environments to predict at-risk students	84%–93% accuracy	Actual datasets & their analysis methods can vary from University to University
Peach et al. (2019)	Identified students at-risk of low performance in online engagement using time-series clustering and graph partitioning algorithms.	NA	Limited datasets and accuracy were not clearly stated
Bujang et al. (2021)	Developed a predictive analytics model based on historical academic performance of studies using Support Vector Machines, Random Forest, Decision Tree & Logistic Regression.	99.6% accuracy	Experiment was conducted only on a single course
Zohair (2019)	Used visualization & clustering algorithms to explain the possibility of creating a prediction model with small data sets of selected features (50 records).	79% accuracy using LDA & 79% SVM	Limited possibility for intervention beforehand.

3 Methodology

The following are the main methods and used in this study to develop students' early dropout prediction model:

- Data collection and integration from the University's SIS system.
- Preprocessing - Data cleansing using Power BI and Excel and feature selection.
- Model building – Training and testing datasets using selected machine learning techniques.
- Prediction and performance evaluation - Applying prediction using new datasets on proposed models. The model performance was compared and evaluated using accuracy, precision, F1score, and Recall performance metrics.

3.1 Data collection

For this study, historical data are collected from Hawassa University Student Information Systems Portal (HUSIS) URL: https://sis.hu.edu.et/ with relevant credentials assigned to the School of Informatics registrar. The datasets contain academic information of B.Sc. students at the Institute of Technology, Hawassa University, Ethiopia. The historical data spans from September 2017 to July 2020. The dataset contains data of 472 students enrolled under the Faculty of Informatics Students Information System (SIS) for the course "Advanced Database Systems." The HUSIS provides student-teacher interaction through its online learning portals, and it is available for students and teachers assigned to a particular Faculty. This study has carried out experiments to evaluate the performance and usefulness of different classification algorithms for predicting the students' academic status.

Table 2. Characteristics of collected datasets.

Fieldname	Description	Domain	Datatype
StudentID	Unique identifier of student	IoT00001/2017 -IoT9999/2020	Char
Year	Entry year	2017-2020	Numeric
Gender	Gender identifier	Male, Female	Char
HoursSpent	Total number of hours a student spent on Student Information System (SIS)	0 – 9999	Numeric
Quizes	Surprise tests each of which weights 5 points.	0 - 5	Numeric
Tests	Students took three scheduled exams at 5^{th},9^{th} & 12^{th} weeks.	0 -10	Numeric
Participation	Students' active engagement during the lecture and lab works – Weights 5 points	0 - 5	Numeric
Final Exam	Written exam at the end of semester	0 - 30	Numeric
Marks	Numeric value of sum of grade points obtained	0 - 100	Numeric
Grade	Letter grade based on fixed scale	Letter grades	Char
Attempts	Number of times a student repeats a course	1, 2, 3	Numeric
Tutorial	Additional lecture and lab work hours	Yes, No	Char
Status	Pass or Fail	0 (failed), 1(passed)	Binary

In addition, data related to students' assignment submission rate, the number of hours spent in an SIS system, and their active engagement on discussion forums have been collected. But due to the complexity of these datasets, we couldn't incorporate them for prediction purposes during the model development process.

3.2 Data preprocessing

Preprocessing is applied to clean the collected data and prepare it as input to the machine learning algorithms. Entries with incomplete datasets (null values) and duplicated values were eliminated using Microsoft Power BI Desktop version. To improve the performance efficiency of the predictive models, entries with the missing values and nulls were removed. In addition, the results of assessments that were missing in the evaluations were removed.

Attributes that are irrelevant for prediction are removed based on their correlation to the final mark. Features that highly correlate to students' final points were selected. Accordingly, unrelated student data for prediction has been removed. Selected attributes that were input to the model are total engagement time in study resources (Spent time), Quizzes, Tests, projects, and Participation (attendance). These are the final sets of predictors selected in the final dataset. Results of quizzes, tests, participation are were grouped to get a clear correlation to the final points. The number of attempts a particular student sits for the course didn't correlate with the Student's achievement and dropped in the experiment. As the final exam is usually taken at the end of the semester and intervention needs to be made before the end of the semester, we opted to incorporate it during prediction modeling.

The information used as predictors for academic status can be identified using the correlation of independent variables to the total marks (Rovira et al., 2017). In this study, the total number of hours a student spent in the learning environment didn't correlate to the final points. Hence, it was less likely to predict dropouts. The datasets were classified for training and testing the model. The datasets used for this study passed a fundamental preprocessing stage to get cleaned datasets for the model intake. The data sets are anonymized to preserve the privacy policy of the University. A total of 360 cleaned datasets has been extracted out of 472 records available from academic years (2017 - 2020). These processed datasets were used as input for each prediction model.

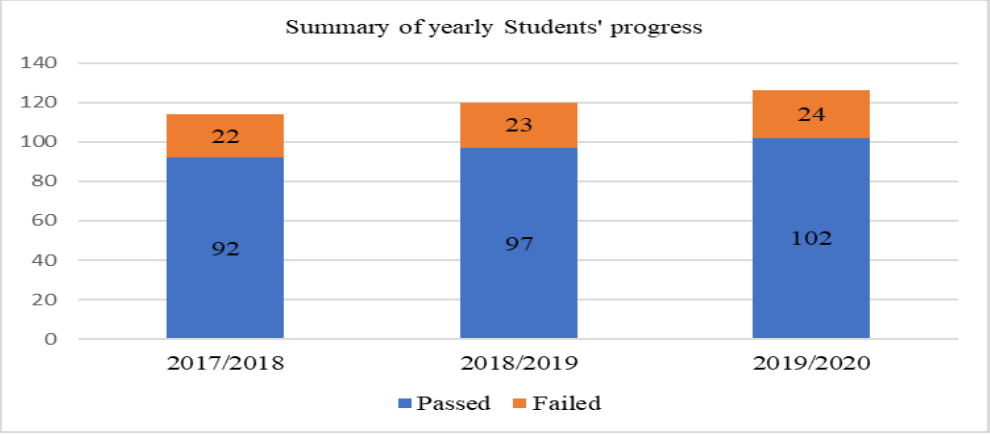

Figure 1. Summary of Students academic progress in years

3.2 Experimental setup for predictive modeling

Tests were conducted using four commonly used Machine learning techniques of Random Forest, Naive Bayesian, Logistic Regression, and Decision Tree. As described in the literature review section of this study, these algorithms are used to classify students' performance into two categories: Fail – those students who cannot achieve passing points; Pass – those students who can earn passing points.

The Python 3.9.5 scripts were used for the construction of predictive models. The Python libraries used are Keras, Sklearn, NumPy, and Seaborn.

4 Results and Discussion

These datasets were used for training and testing the proposed models. Thus, 75% of the datasets were used for model training, and the remining 25% were used to test the model. In addition, the following confusion matrix was used to evaluate the performance of classification algorithms.

Table 3. Confusion matrix

Class	**Predicted Class**	
	P	N
P	True Positive (TP)	False Negative (FN)
N	False Positive (FP)	True Negative (TN)

The information indicated above (Table 3) contains metrics related to the prediction model dropout. The metrics are described as follows:

- True Positive (TP): The number of students correctly classified as "Pass."
- True Negative (TN): The number of dropout students classified correctly as "Fail."
- False Positive (FP): The number of passed students incorrectly classified as "Fail."
- False Negative (FN): The number of failed students classified incorrectly as "Pass."

4.1 Applied evaluation criteria

The metrics selected for measuring the performance of predictive models include the following:

- Accuracy

The accuracy is calculated by dividing the number of correct classes predicted by the total number of types (classes). Thus, it measures the overall rate or classification correctness.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

- Recall

It is the proportion of real positives predicted to be positive. Recall ensures that the predictive model is not overlooking a few students who are Fail or Pass. The Recall is used to evaluate the actual success rate/ students who passed successfully.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (2)$$

- Precision

Precision is the proportion of real negatives expected to be negative. It determines the fraction of true positives among true positives and false positives predicted

The precision is used to assess the dropout of students, and it is determined as follows:

$$\text{Precision} = \frac{\text{TP}}{(\text{TN} + \text{FP})} \tag{3}$$

- F1score

The F1score determines the harmonic mean of recall and precision of a predictive model. Therefore, the F1score is suitable for classification problems where the target labels are imbalanced.

The F1_{score} shows the balance between two measures of classification. It represents a measure widely used trade-off calculation for imbalanced datasets. It is determined as follows:

$$\text{F1score} = \sqrt{\text{Recall} * \text{Precision}} \tag{4}$$

The classification error rate represents a proportion of instances misclassified over the entire set of cases. Its value can be calculated as follows:

$$\text{error} = \frac{(\text{FP} + \text{FN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{5}$$

Table 4. Used prediction models and results of corresponding performance measurement.

	Prediction Model			
Performance Measures	Decision Tree	Naïve Bayes	Random Forest	Logistic Regression
Accuracy	0.93	0.88	0.93	0.94
Precision	0.91	0.89	0.96	0.95
Recall	0.92	0.97	0.96	0.98
F1 Score	0.94	0.96	0.96	0.97

The predictions found through the logistic regression model have an accuracy of 94.44%, 95% precision, 98% recall and 97% F1 score in predicting students' dropout. Out of the 90 records used for testing, the logistic regression model predicted 71 students correctly as successful non dropouts and four students who were actually passed were incorrectly classified as failed.

In addition, the model has correctly predicted 14 students as dropouts. One Student was a dropout, but the Student was detected by the model incorrectly as passed. The proposed predictive model has a high degree of reliability in predicting the data, with an average error rate of just 0.056.

Prediction results by Random Forest were very close but a bit less than the winning model. It was identified as the second-best model in its prediction performance at 93%, 96%, 98%, and 97% accuracy, precision, recall, and F1 score, respectively.

5 Conclusion

Early students' dropout prediction can help academic institutions to provide a timely intervention and apply appropriate planning and training to improve students' success rate.

This study focused on prediction of students' academic dropout using different machine learning techniques. Decision Tree, Random Forest, Logistic Regression and Naïve Bayes have been used for training and testing the model. The proposed prediction method benefits course instructors, institutes, and the University to decide on students' performance and apply appropriate intervention for improving students' academics in advance. This study found out that the Logistic Regression model performed better than the remaining models used in this study in predicting students' early dropouts.

A re-examination of the proposed model using more datasets possibly extracted from academic Big Datasets could be needed to acquire improved accuracy.

Future research needs to include unstructured datasets from students' online activity such as click streams, discussion forums, campus activities, and libraries. In addition, evaluation of other predictions using deep learning methods would be essential to assess the effectiveness of the forecast. The combination of Machine Learning techniques for early dropout prediction needs to be utilized for feature selection, extraction of dropout factors, and calculating percentage of dropout rates. In addition, it would not be enough to evaluate the effectiveness of the machine learning model by observing the accuracy metrics. According to Wu and Flach (2005), the AUC criterion that expresses the area under the ROC-Curve could evaluate the success of Machine learning models in addition to the evaluation metrics used by this study.

Acknowledgments

The work reported in this paper was conducted with the kind support of the University Pardubice grant No SGS_2021_008 of the Student Grant Competition. The outcome of this study is part of the ongoing CRP project under the University of Pardubice.

References

1. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869-8879.
2. Zhang, L., & Rangwala, H. (2018). Early identification of at-risk students using iterative logistic regression. *International Conference on Artificial Intelligence in Education*, (pp. 613-626). Springer, Cham.
3. Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C. J., & Yang, S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, *28*(2), 206–230.
4. Ofori, F., Maina, E., & Gitonga, R. (2020). Using machine learning algorithms to predict students performance and improve learning outcome: A literature based review. *Journal of Information and Technology*, *4*(1), 33–55.
5. Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience*, *2018*.

6. Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
7. Chacha, B. R. C., López, W. L. G., Guerrero, V. X. V., & Villacis, W. G. V. (2019). Student dropout model based on logistic regression. *International Conference on Applied Technologies* (pp. 321–333). Springer, Cham.
8. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences*, *11*(9), 552.
9. Emirtekin, E., Karatay, M., & Kısla, T. (2020). Online course success prediction of students with machine learning methods. *Journal of Modern Technology and Engineering*, *5*(3), 271–282.
10. Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22–32.
11. Imran, A. S., Dalipi, F., & Kastrati, Z. (2019). Predicting student dropout in a MOOC: An evaluation of a deep neural network model. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence* (pp. 190-195).
12. Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., & Khan, S. U. (2021). Predicting at-Risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539.
13. Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, *11*(7), 3130.
14. Alyahyan, E., & Dustegor, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–21.
15. Alipio, M. (2020, April 15). Predicting academic performance of college freshmen in the philippines using psychological variables and expectancy-value beliefs to outcomes-based education: A path analysis. https://doi.org/10.35542/osf.io/pra6z.
16. Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, *104*, 106189.
17. Peach, R. L., Yaliraki, S. N., Lefevre, D., & Barahona, M. (2019). Data-driven unsupervised clustering of online learner behaviour. *NPJ science of learning*, *4*(1), 1–11.
18. Bujang, S. D. A., Selamat, A., & Krejcar, O. (2021). A predictive analytics model for students grade prediction by supervised machine learning. *IOP Conference Series: Materials Science and Engineering*, *1051*(1), 012005.
19. Zohair, L. M. A. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16*(1), 1–18.
20. Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS one*, *12*(2), e0171207.
21. Wu, S., & Flach, P. (2005, August). A scored AUC metric for classifier evaluation and selection. In Second Workshop on ROC Analysis in ML, Bonn, Germany.