Computers and Education: Artificial Intelligence 3 (2022) 100066

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization

Jovial Niyogisubizo [a,*], Lyuchao Liao [a], Eric Nziyumva [a], Evariste Murwanashyaka [b], Pierre Claver Nshimyumukiza [c]

[a] Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, 350118, China
[b] Institute of Rock and Soil Mechanics, University of Chinese Academy of Sciences, Wuhan, China
[c] Integrated Polytechnic Regional College (IPRC), Karongi, Rwanda

ARTICLE INFO

Keywords:
Student dropout prediction
Agents of education
Learning analytics
Artificial intelligence
Educational data mining

ABSTRACT

Student dropout is a serious problem globally. It affects not only the individual who drops out but also the former school, family, and society in general. With the current development of science and technology, big data is emphasized as the most significant technology in data analysis. From the recorded educational data, efficient prediction of students' dropout is currently a hot topic of research. Previous studies have focused only on the students' dropout based on specific levels such as individual, middle school, and university level. However, ensemble learning approaches have not received much research attention so far to predict students' dropout in university classes based on rare datasets. In this paper, we propose a novel stacking ensemble based on a hybrid of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict student's dropout in university classes. On the dataset collected from 2016 to 2020 at Constantine the Philosopher University in Nitra, the proposed method has demonstrated greater performance when compared with the base models using testing accuracy and the area under the curve (AUC) evaluation metrics under the same conditions. Based on the findings of this study, students at the risk of dropping out the school can be identified based on influential factors and different agents of education can refer to this information for early intervention in the uncontrolled behavior that can lead to the risk of dropping out and take proactive precautionary measures before the issue arise.

1. Introduction

Student dropout is considered the most complex and significant issue in the education system (Kim & Kim, 2018). This problem causes economic, social, academic, political, and financial damage to the main agents of education, i.e., from the students to the governmental and promotional agencies for effective and efficient strategies to minimize the indexes of school dropout, since the measures adopted up to now did not have the positive effects on this problem (Martinho et al., 2013).

Previous research has established various definitions of student dropout. The most common definition focuses on whether students will continue to be active until the end of the week or if the current week is the last week in which students are active. Early identification of students at risk of dropping out is critical to reducing the problem and allowing for the necessary conditions to be targeted. As a result, timing considerations are important for the dropout problem. Some studies have found out that 75% of dropouts occur in the first few weeks (Moreno-Marcos et al., 2018). Dropout prediction is frequently regarded as a time series prediction problem or a sequence labeling problem (Mubarak et al., 2020). These can correspond to students' final status (Jin, 2020). On the other hand, the time dimension can be indirectly included in the prediction of dropout by using the input features available in a specific time window, which allows for the selection of a suitable form of intervention (Drlik et al., 2021).

Specifically, students' dropout causes educational deficiencies which can severely affect the social and economic well-being of current and future generations (Rumberger, 1987). Besides, several losses to the society may be encountered because the productive capacity of a nation

* Corresponding author.
E-mail address: jovialniyo93@gmail.com (J. Niyogisubizo).

https://doi.org/10.1016/j.caeai.2022.100066
Received 25 November 2021; Received in revised form 26 March 2022; Accepted 26 March 2022
Available online 28 March 2022

can be challenged by lack of skilled labor force, and the issues of dropout can lead to poor standards of living, unemployment issues, and disruptive behaviors in the society (Catterall, 1987). Based on different demerits of students' dropouts, this issue is considered as a huge hindrance to educational development by researchers, policymakers, and educators (Kim & Kim, 2018). To address this issue, a warning to dropout can enable schools to identify the behaviors that can accelerate the risks of dropping out and take precautionary proactive measures to solve the issue before occurrence (Balfanz et al., 2007). When students intend to drop out, they apply their decision without carefully taking into consideration the dangers that can arise, either asking the guidance from their parents, relatives, experts, or successful colleagues (Sara et al., 2015).

The kind intervention followed by the dropout early warning system can help those who are at the risk of dropping out to remain focused on their studies until they graduate and prepare for a better future (Dynarski & Gleason, 2002). Several governments have developed and implemented dropout early warning systems to deal with this issue. To list a few measures, the department of education and early childhood development in the state of Victoria in Australia developed the student mapping tool to help private and public schools to analyze students at the risk of disengagement and dropout (Lamb & Rice, 2008). Similarly, the state of Wisconsin in the United States developed the same system to predict students' dropouts (Knowles, 2015). Furthermore, dropout early warning systems were implemented in the United States during 2014–2015 for more than half of public high schools (Sullivan, 2017).

Machine learning is a promising tool for building a predictive model for student dropout and offers early warning to responsible authorities to take alternative measures to students at the risk of dropping out the school (Del Bonifro et al., 2020). Recently, several studies emphasized the prediction of students' performances using data mining (DM)/machine learning (ML) models. For example, student dropout in higher education was predicted using the largest known dataset on higher education attrition, which tracks over 32,500 students' demographics and transcript records at one of the nation's largest public universities and the overall results demonstrated that ML models have a significant impact on student retention and success while pointing to several promising directions for future work (Aulck et al., 2016). Similarly, DM methods were employed to analyze student dropouts in the junior years. Among the DM methods used such as logistic regression, decision trees, and neural networks, the decision trees demonstrated the greater ability to predict accurately student dropout (Jadrić et al., 2010).

Although different DM/ML algorithms have been used to predict school dropout, they suffer from some intrinsic limitations. In some cases, the class imbalance could be one of the potential difficulties in implementing school dropout prediction using ML (Orooji & Chen, 2019). In the binary outcome representing students' dropouts, the proportions of the two classes (i.e., dropouts and non-dropouts) tend to be imbalanced (e.g., 1.4 percent of dropouts vs. 98.6 percent of non-dropouts in Korean high school in 2016) (Lee & Chung, 2019). Besides, ML approaches are criticized because they employ a "black box" tactic to predict school dropout and lack proper interpretation of the model for humans (Orooji & Chen, 2019).

To address the issues presented in the literature, this study proposes a novel stacking ensemble based on a hybrid of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict student's dropout in university classes. The main contributions of this paper are summarized as follows:

- To the best of the author's knowledge, this is one of the pioneer studies to explore the stacking ensemble ML between RF, XGBoost, GB, and FNN for students' dropout prediction in university classes.
- A novel two-layer stacking ensemble based on a hybrid of RF, XGBoost, GB, and FNN is proposed to improve the overall classification performance. This helps in estimating the students at risk of quitting an academic course.
- A comparative study is conducted to evaluate the performance of the proposed approach with four base models using performance indicators such as Accuracy, Precision, Recall, F1-Score, and Area under the Curve.

The remainder of this paper is structured as follows. First, a summary of related works focusing on accurate predictions of student dropout using ML techniques is presented in Section 2. Section 3 gradually introduces dataset description and preprocessing. This section describes the main characteristics of the investigated dataset. Section 4 discusses the proposed models for predicting student dropout, which used various ML classifiers. Section 5 discusses the obtained results and the evaluation of the models using various performance measures. Finally, Section 6 concludes the study, discusses its implications and suggests possible future directions.

2. Related works

In the last decades, artificial intelligence (AI) has shown the ability to change many aspects of our society and our lives because it offers the technological basis for new services and tools that help decision-making in daily life (Nagy & Molontay, 2018). Education is not immune to this revolution, indeed, AI and ML algorithms can play a significant role in improving several aspects of the learning process (Lykourentzou et al., 2009). The main reason for emphasizing these technologies is the possibility of building new predictive systems which can be utilized to help students to plan for their future and improve their academic careers (Del Bonifro et al., 2020).

One of the most frequently researched topics in educational data mining (EDM) and learning analytics (LA) disciplines is predicting a student's learning success or failure (Prenkaj et al., 2020). EDM is concerned with the analysis of study-related data to comprehend student behavior. These techniques are typically used to provide more effective learning environments by revealing useful information for modifying course structure or to aid in the prediction of student performance and behavior (Baker & Inventado, 2014). LA, on the other hand, is concerned with the measurement, collection, analysis, and reporting of student data and backgrounds to understand and improve learning and the environments in which it occurs (Siemens & Baker, 2012). EDM and LA methods are typically at the heart of current prediction approaches. Predicting the likelihood of students completing or failing a course, particularly in the early weeks (Alamri et al., 2019), has been one of the hottest research topics in learning analytics, as has EDM (Romero et al., 2010). Once a reliable performance prediction is available, it can be used to identify weak students and provide feedback to students, as well as predict students' failure (Skalka & Drlik, 2020).

Moreover, ML techniques were emphasized to prevent student dropout in distance learning (Kotsiantis et al., 2003). On data provided by the informatics course of the Hellenic Open University, the Naïve Bayes algorithm was found as the best model to predict student dropout (Kotsiantis et al., 2003). Furthermore, DM techniques were developed to predict school failure and dropout. Using the real data gathered on 670 middle-school students from Zacatecas, México, white-box classification methods such as induction rules and decision trees were employed to perform the tasks proposed and the classification accuracies were compared to find the best performing model. The overall results demonstrated that decision trees outperform the benchmark models built with induction rules (Márquez-Vera et al., 2013).

Decision trees and rule-based classifiers, on the other hand, are white-box models that are more understandable and easily interpretable because they expose the reasoning process underlying the predictions. Clustering algorithms or association rule mining are other options. Correlation analysis of course grades and attributes defining credits obtained by students and their average grades can also be useful (Lang et al., 2017).

According to (Lang et al., 2017), classification and regression methods, neural networks, Bayesian networks, support vector machines, logistic regression, and linear regression could be used to solve the student performance prediction problem. These models are frequently referred to as black-box models because they are difficult to understand and interpret. They are all heavily reliant on feature extraction. Feature extraction, also known as attribute selection, is the compilation of a subset of unique predictive features for the predictive problem in modeling. The process aids in identifying relevant attributes in the dataset that contribute to the accuracy of the prediction model, such as the student's most recent activity on the corresponding course to predict a dropout (Queiroga et al., 2020). Strategies that work well for one type of dataset may not work well for another. In this case, it is frequently necessary to manually develop new feature extraction strategies (Li, 2018).

Additionally, supervised classification algorithms were developed to predict student dropouts in higher education (Serra et al., 2018). On the dataset provided by the University of Bari Aldo Moro, during 2013–16, and were provided by the Osservatorio Studenti-Didattica of Miur-Cineca, students at high risk of dropping out of university were identified (Serra et al., 2018). Supervised ML algorithms such as support vector machines, logistic regression, and Gaussian Naïve Bayes classifiers were compared and the support vector machine outperformed conventional ML models (Serra et al., 2018). Besides, artificial neural networks were used as an intelligent system to predict school dropout risk groups in higher education classrooms (Martinho et al., 2013). Specifically, Fuzzy-ARTMAP Neural Network which is one of the AI techniques considered a richer set of features involving family status and living conditions for each student and he conclusions of the study revealed a success rate of the dropout group around 92% and overall accuracy of over 85% which highlights the reliability and accuracy of the system (Martinho et al., 2013).

The study by (Chen et al., 2020) presents a structural topic modeling analysis of 3963 articles published in Computers & Education over the last four decades. Throughout the paper, important questions related to the latent topics and trends in educational technologies have been analyzed. In conclusion, useful insights and implications that can act as powerful guides to contributors in computer and education have been provided. Additionally, a comprehensive study presenting the impact of AI in education, contributors, collaborations, research topics, challenges, and future directions over the last two decades using topic-based bibliometrics has been conducted to fill the gap of various aspects not yet presented by other researchers on the impact of AI in Education (Chen et al., 2022). The results of this study reveal an increasing interest in using AI for educational purposes from the academic community based on different topics such as intelligent tutoring systems for special education, natural language processing for language education, educational robots for AI education, EDM for performance prediction, discourse analysis in computer-supported collaborative learning, neural networks for teaching evaluation, affective computing for learner emotion detection, and recommender systems for personalized learning (Chen et al., 2022).

3. Dataset description and preprocessing

The raw dataset used to conduct experiments in this study was collected directly from Constantine the Philosopher University in Nitra records from 2016 to 2020 (Kabathova & Drlik, 2021). The initial data contains 261 samples and 12 features of students registered in the preparatory lesson on database systems. The features in the raw datasets include information about [access], [tests], [tests_grade], [exam], [project], [project_grade], [assignments], [result_points], [result_grade], [graduate], [year] and [acad_year]. The output variable has two values either 1 for non-dropout or 0 for students who dropped out of the university course. Among all students, 210 (80.46%) passed the course successfully and have not dropped out the school and 51 (19.54) failed to pass. In order to find the screened features for model building, the dataset cleaning process was performed to deal with, irrelevant, noisy, and inconsistent data (Márquez-Vera et al., 2016). The null and unrealistic values were dropped. Additionally, for categorical features, one-hot encoding was performed to transform integers to binary vectors. Besides, the feature selection process was conducted to decide the input variables for model building. As shown in Fig. 1, a correlation heat map was designed to analyze the correlations between input features and output variables. There exist a medium correlation between important features and grade points. Features such as tests, access, and project have a strong correlation with grade points. As a result, strongly correlated features are considered in the model building due to their highest impact on student outcomes. In the last stage of data preprocessing, the dataset was normalized using a standard scaler to eliminate the mean and scale it to unit variance (Obonya & Kapusta, 2018).

4. Methodology

In this study, a novel stacking ensemble made up of a hybrid of RF, XGBoost, GB, and FNN is proposed to predict student's dropout in university classes. A hybrid of four models is proposed to make a powerful meta-learner (Xing et al., 2016). Stacking generalization is defined as an ensemble modeling technique to merge several classification models via a meta-classifier (Wolpert, 1992). The process of combining multiple classification models applies non-linear weightings for low-level predictors to minimize the generalization error rate and enhance the prediction accuracy (Wolpert, 1992). The proposed approach consists of two layers. In the first layer, temporal predictions of the RF, XGBoost, and GB are generated using a complete training dataset to extract the merits of each base classifier. In the second layer, the predictions generated in the first layer are fed to the FNN model to compute the final prediction of student dropout using cross-validation (Jiang et al., 2020). As shown in Fig. 2, the proposed approach has four important stages namely, feature engineering and selection with rationale, dataset splitting, final prediction, and evaluation.

In the proposed stacking ensemble, the raw data with messy and irregular features will be processed through multiple classification models, and valid features will be extracted. Stacking's learning ability stems primarily from the representation of features, which is consistent with the structure of neural networks (NN). The first layer in Stacking is analogous to the first N-1 layer in a NN, while the second layer in stacking is analogous to the last output layer in a NN.

Stacking's first layer can be thought of as a highly complex nonlinear feature converter. In stacking, different classifiers represent heterogeneity for different features. To effectively extract features from raw data, the first layer's base classifiers must meet two requirements such as high accuracy and high diversity. RF, XGBoost, and GB are chosen as the first layer's base models in this study. All base classifiers accomplish learning tasks by combining multiple learners, but their modeling concepts are completely different. These three base models were chosen and combined in the first layer of the proposed model because of their similarities and differences, and they all performed well in cross-validation as well as returning the best accuracy.

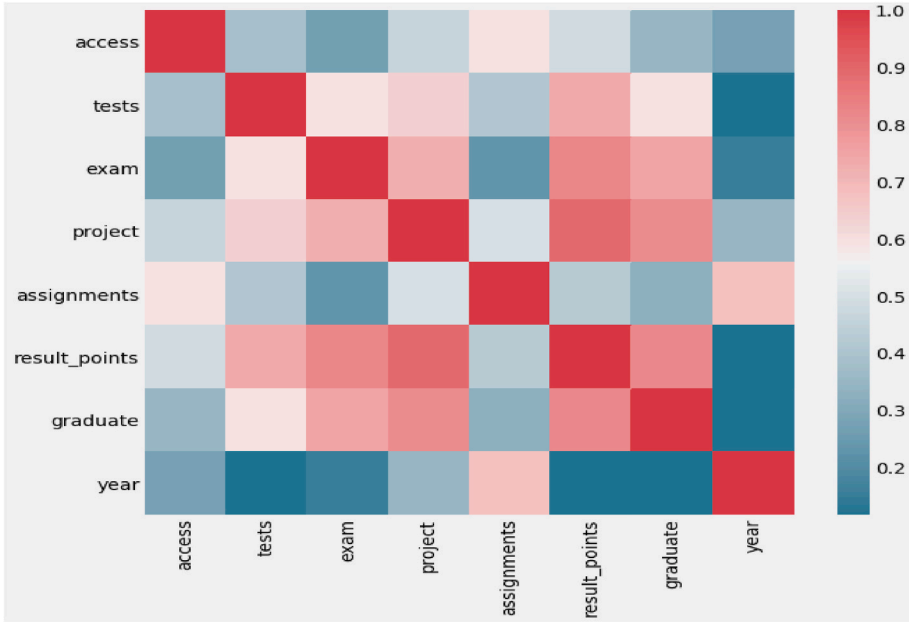

Fig. 1. Correlation heat map between input and output variables.

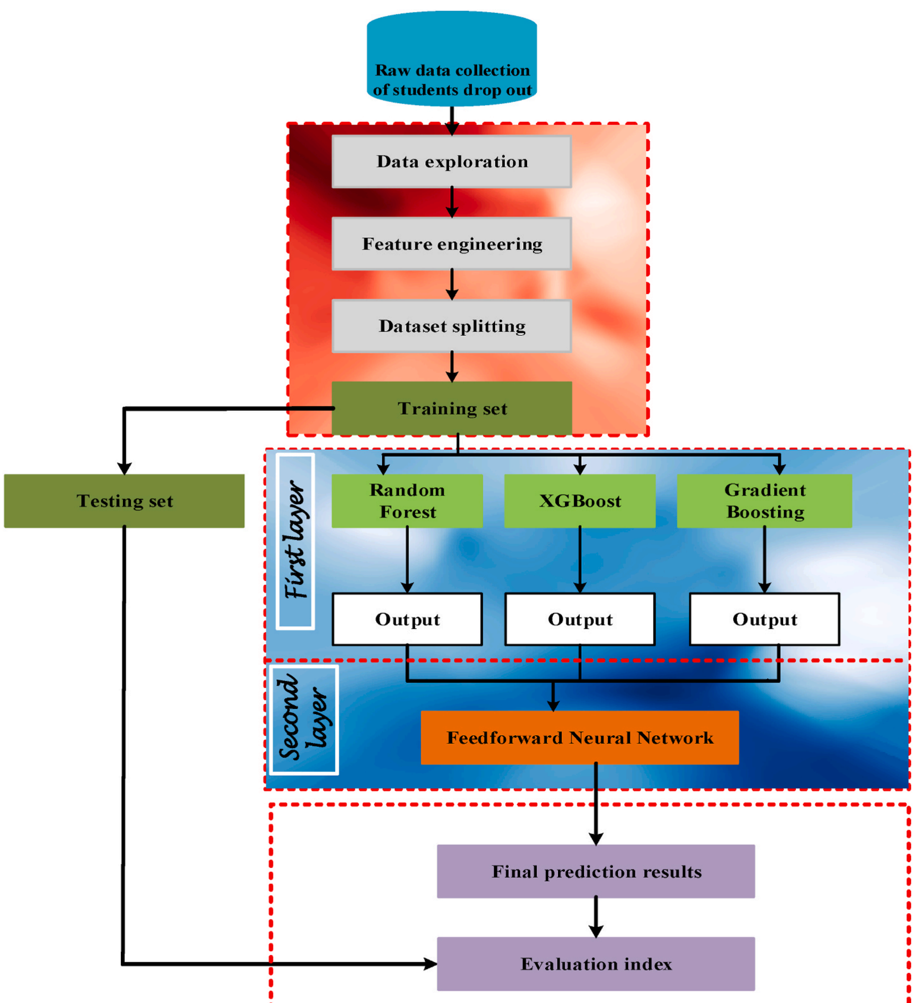

Fig. 2. General representation of proposed stacking framework model.

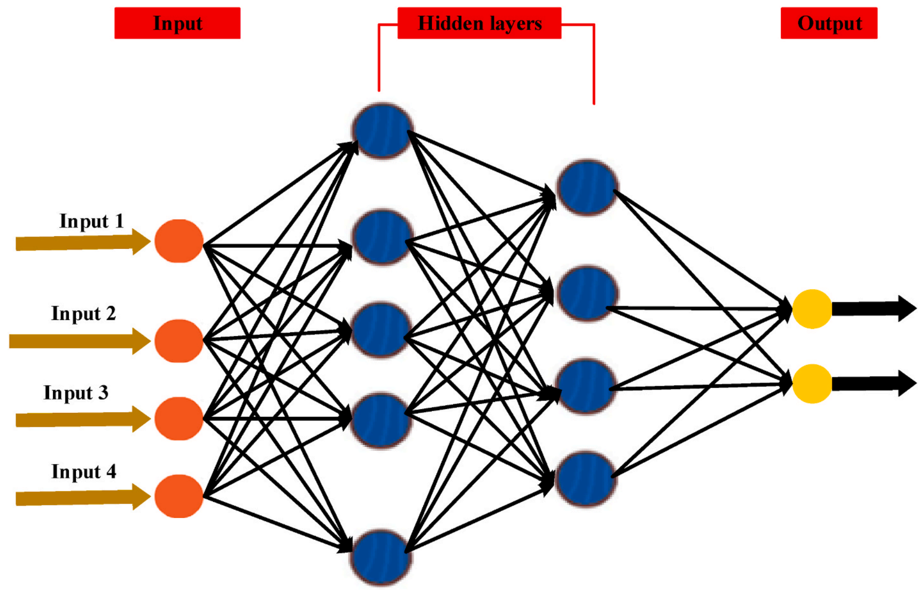

Fig. 3. General representation of FNN.

Because features are extracted using complex non-linear transformations in the second layer of our model, complex classifiers in the output layer are unnecessary. FNN is a good candidate because of its simple structure and additional benefits. Furthermore, the FNN integrated into the second layer can prevent over-fitting even further.

4.1. Preliminaries on classification base models in proposed stacking generalization

In this section, four classification ML models are presented based on their distinguished structure.

4.1.1. Artificial neural network (ANN)

ANN is a commonly used model for prediction systems and it maps several inputs data into a set of suitable outputs (Thanh et al., 2019). The ANN is widely used by many researchers due to its ability of parallel computation, ease of implementation, and swift operation. This model contains a set of many algorithms in which the intelligence of human beings is integrated into the computing machines that can solve extremely complex problems with excellent accuracy performance (Sun et al., 2016). ANN structure consists of three layers, namely, the input layer which functions as raw data receiver, the hidden layer for which the nodes of one layer and next layer are connected entirely, and the output layer which functions as a result displayer (Shahin et al., 2008). The process of learning originates from the error backpropagation utilizing the gradient descent research approach while the predicted output is represented by the symbolic function $\widehat{y}$ denoted by Eq. (1) (He et al., 2009)

$$\widehat{y} = \varphi_0 \left\{ \sum_{i=1}^{F} w_{jp}^{o} \left[\varphi_H \left(\sum_{i=0}^{n} w_{ij}^{H} x_i \right) \right] \right\} \tag{1}$$

For which w_{ij}^{H} represents the hidden while w_{jp}^{o} represents the output layer weights. φ_0 denotes output layer while φ_H denotes the activation function for the hidden layers. x_i represents the input of F features at a given sample time t. To reduce the error function, the gradient-based function optimization algorithm is employed (Hunter et al., 2012). Though the ANN model is popular in producing better accuracy results with a greater number of hidden layers and neurons, it can lead to model overfitting due to inefficiency in the unidirectional learning mechanism with high dimensional space (Hecht-Nielsen, 1992). Besides, the optimization of several hyper-parameters namely, the activation function, learning rate, batch size, etc., requires high levels of expertise and computing performance ability (Hecht-Nielsen, 1992). It is worth mentioning that "In the FNN architecture of ANN, the preceding layer's output is taken as an input for the following layer" (Ivakhnenko, 1971). Fig. 3 represents the simplified structure of FNN.

4.1.2. Gradient boosting (GB) model

GB or gradient boosting tree is an ensemble ML method utilized to solve both classification and regression problems (Ikeagwuani, 2021). The GB was first introduced by (Friedman, 2001) as multiple additive trees that can be used to improve the decision tree approach by utilizing stochastic gradient boosting. The main objective of GB is to decrease a loss function by stirring on the opposite side of the gradient (Friedman, 2001). This means that the GB builds the new base learners to principally correlate with the negative gradient of the loss function in the prediction process. A loss function is an indicator of how good a model is in the prediction process given a number of different parameters. Generally, when the loss function is small, the performance of the model becomes accurate and efficient (Friedman, 2001). The input training samples are represented as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ while the output is the probability of predicting one among the target classes (crash injury severity levels). The detailed description of this algorithm is summarized in Algorithm 1. In the GB model, the gradient descent diminishes complex loss functions that cannot be reduced directly. The loss function to decrease is denoted as $\boldsymbol{L}$. Initialize the model with a unique forecast value $\boldsymbol{F}_0(x)$ with the average of training target values. For initial iteration $m = 1$, calculate the gradient of $\boldsymbol{L}$ in relationship to prediction value $\boldsymbol{F}_1(x)$ and then fit a base learner to the gradient constituents. Compute the magnitude multipliers and update the function to get the prediction value. The process continues recursively until the final result $\boldsymbol{sign}(F_m(x))$ from the collection of all regression trees produced throughout the iteration is computed. The GB employs cross-validation and Out-of-bag (OOB) approaches to find out the optimal number of boosting iterations. OOB permits on-the-fly calculation with no need for persistent model fitting (Friedman, 2001).

Algorithm 1. GB algorithm.

Inputs:

N training samples $\mathbf{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Loss function to minimize $\boldsymbol{L}(y, \boldsymbol{F}(x))$;
Weak learners $\boldsymbol{h}_{\boldsymbol{m}}(x)$;
Step size $\boldsymbol{\gamma}$;
Number of learning iteration M

Process:

1. $\boldsymbol{F}_0(x) = \arg\min_\gamma \sum_{i=1}^{n} \boldsymbol{L}(\boldsymbol{y}_i, \boldsymbol{\gamma})$ %Initialize the model with a constant value
2. for m = 1, … , M:
3. $\tilde{y}_i = -\left[\frac{\partial \boldsymbol{L}(y_i\, \boldsymbol{F}(\boldsymbol{x}_i))}{\partial \boldsymbol{F}(\boldsymbol{x}_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \ldots \boldsymbol{n}$ %Using the negative gradient of the loss function to compute the pseudo residuals.
4. Set $(x_i, \tilde{y}_i), i = 1, \ldots \boldsymbol{n}$ as training data of next tree and construct a new tree mode $\boldsymbol{h}_m(x)$
5. $\boldsymbol{\gamma}_m = \arg\min_\gamma \sum_{i=1}^{n} \boldsymbol{L}(\boldsymbol{y}_i, F_{m-1}(x_i) + \boldsymbol{\gamma} \boldsymbol{h}_m(x_i))$ %Find the best $\boldsymbol{\gamma}_m$
6. $F_m(x) = F_{m-1}(x) + \boldsymbol{\gamma}_m \boldsymbol{h}_m(x)$ %Update the function

end

Output: $\boldsymbol{Y} = \boldsymbol{sign}(F_m(x))$ $\%sign(x) = \begin{cases} 1, x > 0 \\ 0, x = 0 \\ -1, x < 0 \end{cases}$

4.1.3. *Random forest*

Random forest (RF) is a ML algorithm that was first invented by Tin Kam Ho (Yuan & Hu, 2016). It is used to train weak learners so that identical problems get solved and combine them to obtain accurate results (Liaw & Wiener, 2002). RF was extended by (Breiman, 2001) as an accurate classifier made by a group of tree predictors in such a way that there exists the dependence of each tree on independent values of a sampled random vector. Fig. 4 explains the general idea of the RF procedure. It is made up of four important parts: Firstly the training data sample is selected from a given dataset with the help of bootstrap Z times (Breiman, 2001). From this process, each sample set has an equal chance of being selected to represent the training set in every round the sample

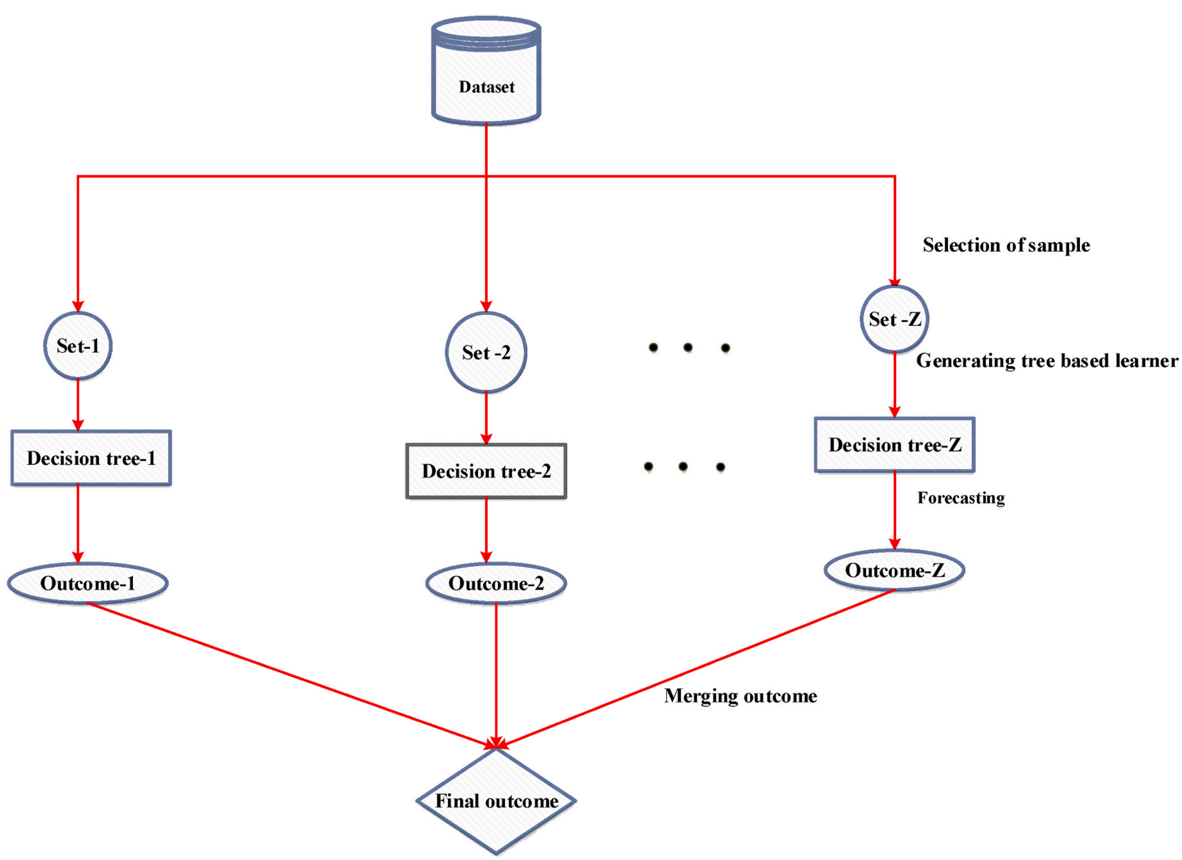

Fig. 4. The structure of Random Forest.

is selected. In the second phase, the decision trees are formed through the process represented as tree-based learner generation. In this phase, the node splitting process is responsible to select randomly the features to represent the set of tree-based predictors. The third phase consists of a forecasting process where the set of selected features is used to present the outcome of each tree predictor. Finally, the outcomes from each tree-based learner are merged where any predictor has equal proportion to the final result.

4.1.4. Extreme Gradient Boosting (XGBoost)

The XGBoost model is one of the most commonly used algorithms introduced by (Chen & Guestrin, 2016) to solve prediction problems. The objective function of XGBoost relies on regularization to the cost function terms such as tree depth and leaf nodes' weights (Zhu et al., 2021). This means that this model has the capacity of enhancing the performance of building trees when the iteration process reduces. Regularization to the cost function which makes XGBoost a regularized boosting technique is mathematically denoted by (Chen & Guestrin, 2016):

$$L(f) = \sum_{i=1}^{n} L(\hat{y}_i,\ y_i) + \sum_{m=1}^{M} \Omega(\delta_m) \tag{2}$$

With

$$\Omega(\delta) = \alpha|\delta| + \frac{1}{2}\beta\omega^2 \tag{3}$$

For which $|\delta|$ represents the leaves number of a classifier and ω denotes the number of vector values assigned to every leaf. To penalize the complexity of the model, the regularizer Ω is identified. The regularization process is explained by Ridge and Lasso's regularization of coefficients β and α combined. It might otherwise be traditional gradient tree boosting if all regularizer parameters are not considered or set to zero (Wang et al., 2020). Contrary to other Boosting trees, XGBoost uses the second-order Taylor expansion of the loss function as shown in Eq. 4

$$L(f) \approx \sum_{i=1}^{n} \left[(L(\hat{y}_i,\ y_i)) + g_i\delta_t(x_i) + \frac{1}{2}h_i\delta_t^2(x_i) \right] + \Omega(\delta_t) \tag{4}$$

For which $g_i = \partial_{\tilde{y}} L(\hat{y}_i,\ y_i)$ and $h_i = \partial_{\tilde{y}}^{\,2} L(\hat{y}_i,\ y_i)$. At step t, Constant terms are removed to obtain the next approximation.

$$\hat{L}(f) = \sum_{i=1}^{n} \left[g_i\delta_t(x_i) + \frac{1}{2}h_i\delta_t^2(x_i) \right] + \Omega(\delta_t) \tag{5}$$

After defining I_j as the instance set at leaf j and expandingΩ. Eq. (5) can be rewritten as shown by Eq. (6).

$$\hat{L}(f) = \sum_{j=1}^{T} \left[\left(\sum i \in I_j g_i \right) \omega_j + \frac{1}{2} \left(\sum i \in I_j h_i + \beta \right) \omega_j^2 \right] + \alpha|\delta| \tag{6}$$

Conclusively, the gain used by XGBoost instead of entropy or information gain for splits in a decision tree is represented by Eqs. (7) and (8) respectively.

$$\begin{aligned} G_j &= \sum i \in I_j g_i \\ H_j &= \sum i \in I_j h_i \end{aligned} \tag{7}$$

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \beta} + \frac{G_R^2}{H_R + \beta} - \frac{(G_R + G_L)^2}{H_R + H_L + \beta} \right] - \alpha \tag{8}$$

For which the first term represents the score of the left child, the second denotes the score of the right child and the third represents the score if we do not split (Chen & Guestrin, 2016). Whereas α is the complexity cost if we add a new split. Therefore, if missing values are found XGBoost deals with them by suggesting a default direction for each split.

4.2. Performance evaluation

In order to evaluate the results in this study classification metrics the confusion matrix and its related performance metrics such as accuracy (ACC), precision (PR), recall (Rec), the area under receiver operating characteristics curve (AUC), and F1 Score (F1) were employed (Powers, 2020). Below is a brief description of all metrics used in the evaluation stage.

$$ACC = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Negative + False\ Positive} \tag{9}$$

$$PR = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{10}$$

$$Rec = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{11}$$

$$F1\ \ Score = \frac{2{*}Rec{*}PR}{Rec + PR} \tag{12}$$

- **The area under the curve (AUC)**

The AUC is the frequently used performance measure for classification ML models, where the curve represents the Receiver Operating Characteristics (ROC) curve which can include the totality of prediction performance of a classification method for all classification thresholds (Muschelli, 2020). The curve allows visualizing the behavior of classification models and denotes the trade-off of the classifier between the number of correct positive and incorrect positive predictions (Muschelli, 2020). To be more specific, the ROC curve represents a two-dimensional graph in which the True Positive Rate (TPR) denotes the y-axis while False Positive Rate (FPR) represents the x-axis (Koizumi et al., 2019). AUC values range between 0.5 and 1.0. When the value of AUC is near 1, it shows that the model has better performance while a value less than 0.5 indicates poor performance and the steepness of ROC should be high as it represents high TPR while less FPR (Koizumi et al., 2019).

5. Experimental results and discussion

In this study, the implementation and experiments of the proposed approach were mainly implemented in Python using its popular package scikit-learn (0.24.2) developed by Google (Yang et al., 2021). In order to predict student's dropout in university classes accurately, the stacking ensemble based on a hybrid of RF, XGBoost, GB, and FNN to predict student's dropout in university classes is proposed. In addition, a cross-validation approach is adopted while training the model to avoid the issues of over-fitting (Zhang & Yang, 2015). For prediction purposes, the input data were randomly divided into training and testing datasets with percentages of 80% and 20%, respectively (Joseph & Vakayil, 2021, pp. 1–11). Ten-fold cross-validation is applied to address the problem of over-fitting.

Table 1
Accuracy results of proposed methods.

Method	Training Accuracy (%)	Testing Accuracy (%)	Computational Time (second)
Feed Forward Neural Network	96.67	76.67	10.95
Random Forest	91.67	91.66	6.16
Gradient Boosting	86.67	86.66	10.19
Extreme Gradient Boosting	91.67	91.66	10.81
Stacking ensemble	93.59	92.18	8.27

5.1. Model training and hyper-parameters tuning

The proposed method is an integrated model with a number of crucial parameters to tune and systematic grid search is implemented to optimize important parameters and improve accuracy results (Feurer & Hutter, 2019). As mentioned before, ensemble learning classification methods (RF, XGBoost, and GB) are integrated into the first layer of the proposed novel model due to their ability to combine multiple weak learners and produce enhanced accuracy results compared to a single learner (Li et al., 2021). In the first layer, the parameters that must be optimized to train both base models include, the number of weak learners (n_estimators), the learning rate (Learning-rate), and the maximum depth of the tree (max_depth) (Clark & Pregibon, 2017). Generally, there exists a trade-off between model accuracy and computational time during the process of parameter optimization (Aparna & Nair, 2016). In the GB model, after conducting several experimental tests and checking previous literature, the number of weak learners was chosen to be 1000, 0.06 Learning-rate, and 11 maximum depth of the tree (Bentéjac et al., 2021). These parameters were selected because they returned higher testing accuracy results. Besides, In the XGBoost and RF model, parameters that returned higher testing accuracy results include 1000 number of trees, 0.09 learning rate, and 8 maximum depth of the tree (Chen et al., 2015).

In the second layer of our method, the input features were trained using FNN. The input layer of FNN consists of five neurons, each representing one input variable, while the output layer consists of one neuron which represents the output variable (Eldan & Shamir, 2016). To determine the number of hidden layers and the number of neurons in each layer, a replication searching process was followed to find out an optimized method with accurate prediction performance (Pontes et al., 2016). After performing several iterations, the optimal topology which produced efficient prediction results was found to be two hidden layers comprising of five neurons with tangent sigmoid activation function and two neurons with Softmax activation (Pontes et al., 2016). The leading training function which resulted in the best prediction accuracy in the FNN is the Levenberg–Marquardt (Yu & Wilamowski, 2018). This was selected after comparing with other training algorithms such as Bayesian regularization backpropagation, scaled conjugate gradient, variable learning rate backpropagation, resilient backpropagation, and BFGS quasi-Newton (Yu & Wilamowski, 2018). After optimizing the parameters for the models in the first layer and the remaining in the second layer, a novel stacking ensemble model proposed was tested using the same testing set, and finally, the prediction performance was assessed based on the contingency table.

Table 2
Overall prediction results for students' dropout and non-dropout.

Model	Outcome	Precision	Recall	F1-Score
Feed Forward Neural Network	Dropout	0.88	1.00	0.93
	Non Dropout	1.00	0.96	0.98
	Overall	0.83	0.82	0.80
Random Forest	Dropout	1.00	0.64	0.78
	Non Dropout	0.90	1.00	0.95
	Overall	0.92	0.92	0.91
Gradient Boosting	Dropout	0.69	0.79	0.73
	Non Dropout	0.93	0.89	0.91
	Overall	0.87	0.87	0.87
Extreme Gradient Boosting	Dropout	0.85	0.79	0.81
	Non Dropout	0.94	0.96	0.95
	Overall	0.92	0.92	0.92
Stacking ensemble	Dropout	0.89	0.96	0.85
	Non Dropout	0.95	0.94	0.95
	Overall	0.93	0.93	0.92

5.2. Comparison of prediction performance

To evaluate the effectiveness of the proposed model, the experimental results of our method are compared to the base models in the first layer. The results are compared based on training and testing results produced by FNN, RF, GB, XGBoost, and Stacking ensemble. As shown in Table 1, the training accuracy results range from 86.67% to 96.67%. On the other hand, the testing accuracies range from 76.67% to 92.18%. The FNN suffers from overfitting issues (Jabbar & Khan, 2015) while the proposed method shows improved prediction performance with a training accuracy of 93.59% and a testing accuracy of 92.18%. According to Table 1, it can be observed that training the model is computationally demanding for all models with different duration. The proposed stacking ensemble is the fastest which scientifically reduced the computational cost compared to the other techniques with a consuming time of 8.27 s while the FNN computation time is longer than the remaining models. Because of the hidden layers, the FNN naturally recorded a significantly longer time. However, given the proposed problem's time frame, the runtime was not prohibitively long. In the worst-case scenario, it took around 10.95 s to run with all of the attributes enabled.

Fig. 5 demonstrates graphically the above training and testing accuracies for all methods.

Although accuracy is the frequently used performance indicator to evaluate model results, relying only on accuracy could lead to misinterpretation since the model can predict only the dominant class and ignore the minor class (Riquelme et al., 2015). To address this issue, precision, recall, and F1-Score are other evaluation metrics considered in this study. These indicators play a significant role in predicting every

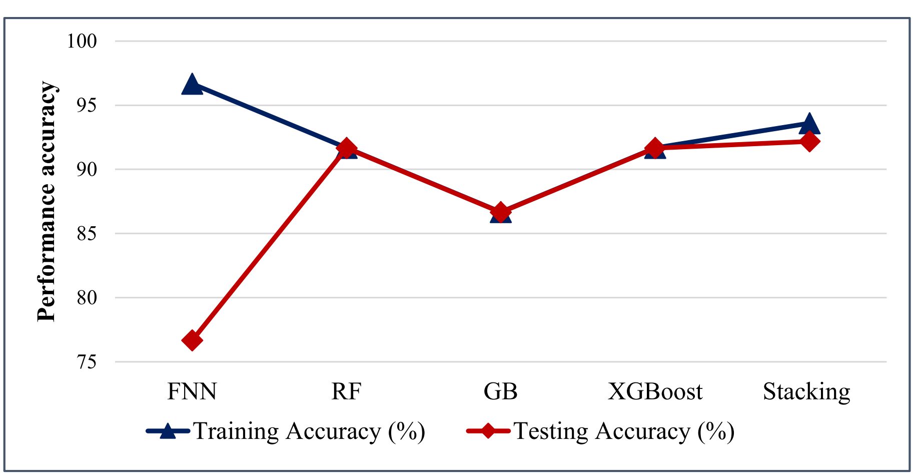

Fig. 5. Training and testing accuracy for all methods.

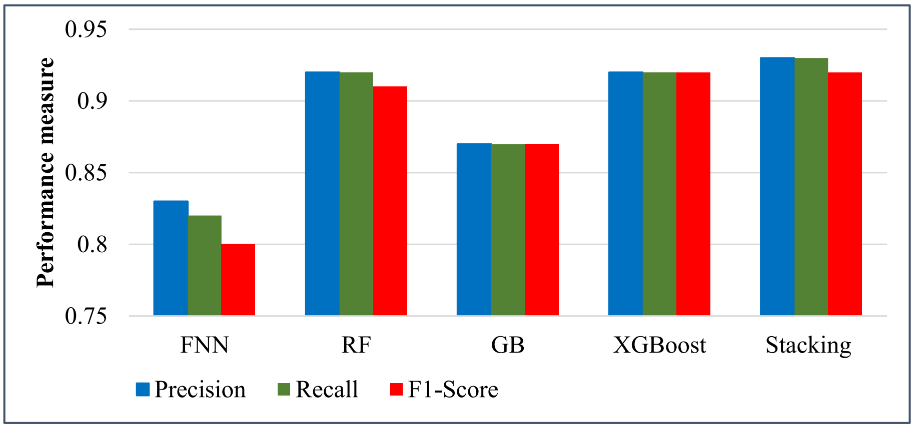

Fig. 6. Overall prediction results for students' dropout and non-dropout.

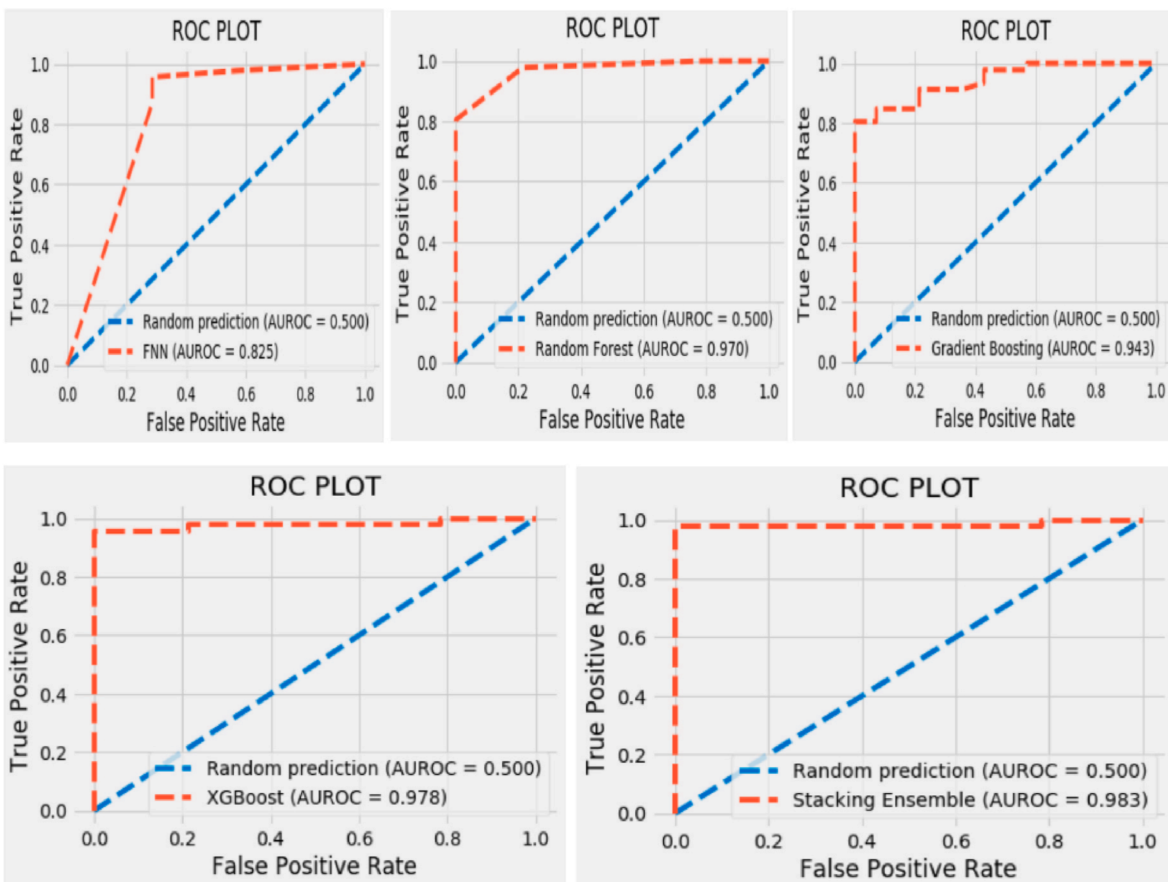

Fig. 7. ROC curves and AUC Results for all methods.

class and provide better insight into prediction results. As shown in Table 2 and Fig. 6 respectively, the overall results for all classes are demonstrated. From the definition of precision and recall, a model that shows improved performance in all classes must be considered. Similarly, the F1-Score which can use both precision and recall is an added advantage while evaluating the model. From Table 2, it is clear that the overall precision, recall, and F1-Score results for our method are 0.93, 0.93, and 0.92 respectively, which outperforms the base models.

Generally, the recall results of 0.93 obtained for stacking ensemble on testing set indicates that the model can predict nearly 93% of dropout. Similarly, the precision value of 0.93 means that we are correct about those predictions about 93% while predicting both university students' dropout and non-dropout after course delivery. These results indicate that our method is capable of handling efficiently every class and eventually, our model is more optimum for students' dropout in university class prediction. Regarding the F1-Score, its values range from 0 to 1, for which 0 represents extremely poor results while the value near 1 shows efficient results. Our method managed to reach the overall F1-Score of 0.92. Better prediction performance of our method indicates that it is a viable option for predicting students' dropout in university classes.

Even if the results shown above for our model mark an acceptable performance, the AUC is the most important evaluation indicator used to assess the overall performance of all models used in this paper (Walter et al., 2015). As shown on Fig. 7, the AUC results and the corresponding ROC curves for FNN, RF, GB, XGBoost, and stacking ensemble are 0.825, 0.97, 0.943, 0.978, and 0.983 respectively. A higher AUC score obtained from our model implies a better and acceptable classification performance because points representing model classification better than random guess are located above the diagonal line.

The discussion of the various performance metrics confirms that the selected binary classification models can be used to predict students' dropout or Non-dropout at the individual course level, even when the dataset is scarce and has a limited number of input features. However, before the classifiers can be used to predict student dropouts, a broader set of performance metrics must be investigated. This statement is consistent with the findings of other research papers published in the domains of AI and ML.

Nonetheless, there are some limitations to the presented case study and its findings. The first is, as previously stated, the size of the dataset is limited. In contrast to many other ML model application domains, the amount of data in the educational domain cannot be easily increased by combining different resources. The reason for this is that individual records should frequently represent individual students' learning outcomes or behavior. As a result, the datasets are frequently smaller than what the ML algorithms require. This limitation is frequently overcome only by employing the same precise research design used in classical educational technology research, in which a sufficient number of records is estimated or guaranteed before the experiment itself.

The same holds for the number and quality of the independent features. The repetition of certain activities is frequently used in the learning process. If these activities are interconnected or conditioned, their inclusion in the ML technique should be thoroughly scoped before the data collection process begins. Weakness can also be identified in the selection of the time threshold for which the prediction is made. Because the time variable was not directly included, it was necessary to identify milestones when the students' performance would be difficult to predict. The end of the second third of the term was used as a compromise in terms of the natural distribution of the individual activity categories in the course sections. However, better results would be expected if the activities in the course sections were designed with the requirements of the ML techniques in mind.

Another limitation of this study is that different runs of the courses provided different data to be analyzed. As a result, it was difficult to determine which attributes are important enough to predict the student's performance in general. The next limitation is the selection of the classifiers used in this case study. Literature reviews of the current state of the research and trends in LA and EDM provide numerous examples of more or less advanced ML techniques that can be used to predict early student dropout at the course level. However, none of them have produced noticeably better results thus far. Because the main goal of this study was not to find the best one, the final choice of the classifier used in this case study allows mentioning that good prediction could also be achieved using stacking ensemble, but the performance metrics must be evaluated.

The case study's final flaw is a type of appropriate intervention that is not discussed in depth. The case study has already demonstrated that the student's dropout at the university level can be predicted based on the activities chosen. The reason for this unflattering state, on the other hand, remains unknown and necessitates further investigation. It would be interesting to investigate whether other types of activities have a similar impact on student engagement and which types of activities can be exchanged during the intervention phase.

6. Conclusions

Students' dropout is considered one of the most complicated and negative issues faced in the learning process (Freeman & Simonsen, 2015). Different agents of education such as parents, stakeholders, government institutions, etc., are concerned with this issue. Although several measures have been taken to tackle this issue, the effects of student dropout are still inevitable. Being able to predict accurately student dropout could help to alleviate its social and economic costs (Robison et al., 2017). The analytical ML solutions can significantly be employed to determine and predict such influential factors among social welfare, learner's outcome, learning conditions, age, gender, family status, student's sponsorship, etc., and thus to explain student dropout.

An accurate prediction can boost learners to focus on their studies, to avoid being at the risk of dropping out the school and become determined on their studies. Besides, the responsible authorities at school can refer to this information when they are deciding about promotion, repeating, or dismissing students who failed and think about other alternatives for those who failed by either repeating the course they failed or getting the second sitting exam. Efficient prediction can also help in giving alerts to teachers for early intervention in the uncontrolled behavior that can lead to the risk of dropping out and take proactive precautionary measures before the issue arise. This study contributes mainly to the performance improvement of university student dropout prediction. This allows estimating the risk of quitting an academic course and the reduction in student outcomes. Additionally, this research also contributes to the development of effective stacking ensemble-based method to minimize the negative impacts of inaccurate prediction of student dropout using a single model.

The study presented here helps to solve the problem of student dropouts at the course level. The results demonstrated that, despite a small dataset, appropriately selected indicators that do not require access to system logs can be beneficial if different performance metrics are evaluated. The predictive models were fed with data gathered about students' online learning environment activities and partial achievements. Simultaneously, a proposed methodology is reliable for predicting course completion when there is enough time for educators to intervene timely.

The results obtained for our method can help in reducing the dropout rate based on identified students that are likely to be affected and the influential factors. Moreover, once the students who are at the risk of being affected are identified, the agents of education can gather their forces to take efficient measures of eradicating the issue. Furthermore, the different learner's motivation strategies can be emphasized to improve the performance and help learners finish their programs successfully. For future research, other computation approaches such as deep learning and other hybrid models can be used to predict student dropout and compare the results with the findings of this study. Other influential factors not presented in this study must also be considered and a further feature analysis study is recommended to help the agents of education in handling successfully the issues of student dropout.

Declaration of competing interest

None declared.

Acknowledgments

This work is supported in part by projects of the National Science Foundation of China (41971340, 41471333, 61304199), and projects of Fujian Provincial Department of Science and Technology (2018Y3001, 2018H0001, 2019I0019).

References

Alamri, A., Alshehri, M., Cristea, A., Pereira, F. D., Oliveira, E., Shi, L., & Stewart, C. (2019). Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. *International Conference on Intelligent Tutoring Systems*.

Aparna, K., & Nair, M. K. (2016). Effect of outlier detection on clustering accuracy and computation time of CHB K-means algorithm. In *, Vol. 2. Computational intelligence in data mining* (pp. 25–35). Springer.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364*.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). Springer.

Balfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist, 42*(4), 223–235.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Catterall, J. S. (1987). On the social costs of dropping out of school. *High School Journal, 71*(1), 19–30.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: Extreme gradient boosting. *R package version 0, 1*(4), 1–4. *4-2*.

Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education. *Computers & Education, 151*, Article 103855.

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of artificial intelligence in education: Contributors, collaborations, research topics, challenges, and future directions. *Journal of Educational Technology & Society, 25*(1).

Clark, L. A., & Pregibon, D. (2017). Tree-based models. In *Statistical models in S* (pp. 377–419). Routledge.

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In *International conference on artificial intelligence in education*.

Drlik, M., Munk, M., & Skalka, J. (2021). Identification of changes in VLE stakeholders' behavior over time using frequent patterns mining. *IEEE Access, 9*, 23795–23813.

Dynarski, M., & Gleason, P. (2002). How can we help? What we have learned from recent federal dropout prevention evaluations. *Journal of Education for Students Students Placed At Risk, 7*(1), 43–69.

Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks. *Conference on Learning Theory*.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning* (pp. 3–33). Cham: Springer.

Freeman, J., & Simonsen, B. (2015). Examining the impact of policy and practice interventions on high school dropout and school completion rates: A systematic review of the literature. *Review of Educational Research, 85*(2), 205–248.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65–93). Elsevier.

He, T., Dong, Z., Meng, K., Wang, H., & Oh, Y. (2009). Accelerating multi-layer perceptron based short term demand forecasting using graphics processing units. In *2009 transmission & distribution conference & exposition: Asia and Pacific*.

Hunter, D., Yu, H., Pukish, M. S., III, Kolbusz, J., & Wilamowski, B. M. (2012). Selection of proper neural network sizes and architectures—a comparative study. *IEEE Transactions on Industrial Informatics, 8*(2), 228–240.

Ikeagwuani, C. C. (2021). Estimation of modified expansive soil CBR with multivariate adaptive regression splines, random forest and gradient boosting machine. *Innovative Infrastructure Solutions, 6*(4), 1–16.

Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. *IEEE transactions on Systems, Man, and Cybernetics*, (4), 364–378.

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 163–172.

Jadrić, M., Garača, Ž., & Čukušić, M. (2010). Student dropout analysis with application of data mining methods. *Management: Journal of Contemporary Management Issues, 15*(1), 31–46.

Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics and Its Applications, 541*, Article 122272.

Jin, C. (2020). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments*, 1–19.

Joseph, V. R., & Vakayil, A. (2021). *SPlit: An optimal method for data splitting.* Technometrics.

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences, 11*(7), 3130.

Kim, D., & Kim, S. (2018). Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability, 10*(4), 954.

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining, 7*(3), 18–67.

Koizumi, Y., Murata, S., Harada, N., Saito, S., & Uematsu, H. (2019). SNIPER: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*.

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *International conference on knowledge-based and intelligent information and engineering systems*.

Lamb, S., & Rice, S. (2008). *Effective strategies to increase school completion report: Report to the Victorian department of education and early childhood development.* Communications Division, Department of Education and Early Childhood Development.

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (2017). *Handbook of learning analytics*. New York: SOLAR, Society for Learning Analytics and Research.

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), 3093.

Li, Y. (2018). Feature extraction and learning effect analysis for MOOCs users based on data mining. *International Journal of Emerging Technologies in Learning (iJET), 13*(10), 108–120.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Li, M., Yan, C., & Liu, W. (2021). The network loan risk prediction model based on Convolutional neural network and Stacking fusion model. *Applied Soft Computing, 113*, Article 107961.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education, 53*(3), 950–965.

Márquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 8*(1), 7–14.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems, 33*(1), 107–124.

Martinho, V. R., Nunes, C., & Minussi, C. R. (2013). Prediction of school dropout risk group using neural network. In *2013 federated conference on computer science and information systems*.

Martinho, V. R. D. C., Nunes, C., & Minussi, C. R. (2013). An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. In *2013 IEEE 25th international conference on tools with artificial intelligence*.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies, 12*(3), 384–401.

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 1–20.

Muschelli, J. (2020). ROC and AUC with a binary predictor: A potentially misleading metric. *Journal of Classification, 37*(3), 696–708.

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*.

Obonya, J., & Kapusta, J. (2018). Identification of important activities for teaching programming languages by decision trees. *DIVAI 2018*.

Orooji, M., & Chen, J. (2019). Predicting Louisiana public high school dropout through imbalanced learning techniques. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*.

Pontes, F. J., Amorim, G., Balestrassi, P. P., Paiva, A., & Ferreira, J. R. (2016). Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing, 186*, 22–34.

Powers, D. M. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061*.

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys, 53*(3), 1–34.

Queiroga, E. M., Lopes, J. L., Kappel, K., Aguiar, M., Araújo, R. M., Munoz, R., Villarroel, R., & Cechinel, C. (2020). A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences, 10*(11), 3998.

Riquelme, N., Von Lücken, C., & Baran, B. (2015). Performance metrics in multi-objective optimization. In *2015 Latin American computing conference (CLEI)*.

Robison, S., Jaggers, J., Rhodes, J., Blackmon, B. J., & Church, W. (2017). Correlates of educational success: Predictors of school dropout and graduation for urban students in the Deep South. *Children and Youth Services Review, 73*, 37–46.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (2010). *Handbook of educational data mining*. CRC press.

Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. *Review of Educational Research, 57*(2), 101–121.

Sara, N.-B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. In *ESANN 2015 proceedings, European symposium on artificial neural networks*. Computational Intelligence.

Serra, A., Perchinunno, P., & Bilancia, M. (2018). Predicting student dropouts in higher education using supervised classification algorithms. In *International conference on computational science and its applications*.

Shahin, M. A., Jaksa, M. B., & Maier, H. R. (2008). State of the art of artificial neural networks in geotechnical engineering. *Electronic Journal of Geotechnical Engineering, 8*(1), 1–26.

Siemens, G., & Baker, R. S.d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*.

Skalka, J., & Drlik, M. (2020). Automated assessment and microlearning units as predictors of at-risk students and students' outcomes in the introductory programming courses. *Applied Sciences, 10*(13), 4566.

Sullivan, R. (2017). *Early warning signs. A solution-finding report*. Center on Innovations in Learning, Temple University.

Sun, K., Huang, S.-H., Wong, D. S.-H., & Jang, S.-S. (2016). Design and application of a variable selection method for multilayer perceptron neural network with LASSO. *IEEE Transactions on Neural Networks and Learning Systems, 28*(6), 1386–1396.

Thanh, H. V., Sugai, Y., Nguele, R., & Sasaki, K. (2019). Integrated workflow in 3D geological model construction for evaluation of CO2 storage capacity of a fractured basement reservoir in Cuu Long Basin, Vietnam. *International Journal of Greenhouse Gas Control, 90*, Article 102826.

Walter, W. D., Onorato, D. P., & Fischer, J. W. (2015). Is there a single best estimator? Selection of home range estimators using area-under-the-curve. *Movement Ecology, 3* (1), 1–11.

Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H., & Gao, L. (2020). Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. *Acta Geotechnica, 15*(11), 3135–3150.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior, 58*, 119–129.

Yang, F., Wang, X., Ma, H., & Li, J. (2021). Transformers-sklearn: A toolkit for medical language understanding with transformer-based models. *BMC Medical Informatics and Decision Making, 21*(2), 1–8.

Yuan, Y., & Hu, X. (2016). Random forest and objected-based classification for forest pest extraction from UAV aerial imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 41*, 1093.

Yu, H., & Wilamowski, B. M. (2018). Levenberg–marquardt training. In *Intelligent systems* (pp. 12-11–12-16). CRC Press.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics, 187*(1), 95–112.

Zhu, X., Chu, J., Wang, K., Wu, S., Yan, W., & Chiam, K. (2021). Prediction of rockhead using a hybrid N-XGBoost machine learning framework. *Journal of Rock Mechanics and Geotechnical Engineering*.