Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Intelligent System for Student Performance Prediction Using Machine Learning

***Mustafa S. Ibrahim Alsumaidaie*[1], *Ahmed Adil Nafea**[2], *Abdulrahman Abbas Mukhlif*[3], *Ruqaiya D. Jalal*[1], *Mohammed M AL-Ani*[4]**

[1]Department of Computer Science, College of Computer Science and IT, University of Anbar Ramadi, Iraq.
[2]Department of Artificial Intelligence, College of Computer Science and IT, University of Anbar, Iraq.
[3]Registration and Students Affairs, University Headquarter, University of Anbar, Anbar, Iraq.
[4]Center for Artificial Intelligence Technology (CAIT)، Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia.

Received 24/09/2023, Revised 26/01/2024, Accepted 28/01/2024, Published Online First 20/05/2024

Abstract

Accurately predicting student performance remains a significant challenge in the educational sector. Identifying students who need additional support early can significantly impact their academic outcomes. This study aims to develop an intelligent solution for predicting student performance using supervised machine learning algorithms. This proposed focus on addressing the limitations of existing prediction models and enhancing prediction accuracy. In this work employed three supervised machine learning algorithms: Random Forest, Extra Trees, and K-Nearest Neighbors. The steps of research methodology contained (data collection, preprocessing, feature identification, model construction, and evaluation). This paper utilized a dataset comprising 24,000 training instances and 6,000 testing instances, applying various preprocessing techniques for data optimization. The Extra Trees algorithm achieved the highest accuracy (98.15%), followed by Random Forest (94.03%) and K-Nearest Neighbors (91.65%). All algorithms demonstrated high precision and recall. Notably, K-Nearest Neighbors exhibited exceptional computational efficiency with a training time of 0.00 seconds. This study proposed an efficient model for prediction student performance. The high accuracy and efficiency of the proposed system highlight its potential for application in educational data mining. The findings of this proposed to improving student success rates in educational institutions by enabling timely and appropriate interventions.

Keywords: Artificial Intelligence, Educational Data Mining, Extra Trees Algorithm, Student Performance Prediction, Supervised Machine Learning.

Introduction

Student performance prediction is an important role in the field of education and the management of educational institutions. Understanding the importance of anticipating student performance helps educational institutions make strategic decisions and improve overall educational outcomes. Providing an accurate prediction of student performance that educational institutions can benefit from in several aspects. It can help to identify students' needs and intervene early to meet those needs. When there have been expectations about student performance, could be effectively direct efforts and resources to provide the support needed to the students who need it most [1]. This can help

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

reduce the failure rate and improve overall academic success. Student performance expectations can be used to improve the strategic planning process in educational institutions. Based on performance expectations, institutions can develop customized educational programs that target potential weaknesses and enhance students' academic capabilities. Tests and assessments may also be structured based on these expectations to ensure that educational goals are met, and students are motivated to succeed [2].

Student performance detection can contribute to improving the overall level of education by providing more efficient and effective teaching. When teachers and educators have knowledge of students' expected performance, they can plan appropriate lessons and instructional methodologies to suit students' abilities[3]. They can modify teaching methods and assessment strategies to enhance students' understanding and academic achievement. Student performance prediction serves as a support tool that helps teachers identify the weaknesses and strengths of each individual student, allowing them to provide individualized and targeted support for improving performance. Student performance expectations can be used to make comprehensive assessments of the educational institution [4]. Before monitoring the achievement of expectations and comparing them to actual results, educational institutions can determine the effectiveness of their educational strategies and evaluate the quality of teaching and academic programs. This information can help you make strategic decisions to improve the educational process and promote academic excellence. Definitely, predicting student performance is of great importance to educational institutions. It improves educational outcomes and raises the level of education in general. By utilizing these advanced predictive tools and technologies, educational institutions can achieve sustainable improvements in the quality of academic and education success of students[5].

However, traditional methods often challenge in accuracy, scalability, and adaptability to different educational contexts, particularly in quickly evolving learning environments[6]. In the last years, several studies have tried to predict student performance using various ML algorithms. Although advancements, many existing models still try with issues such as overfitting, deficient flexibility to different types of educational data, and the inability to handle large-scale datasets effectively[7,8].

This study proposed an advanced approach that controls ML algorithms - specifically Random Forest, Extra Trees, and K-Nearest Neighbors - to address these issues. This proposed aims to improve prediction accuracy and efficiency of student performance models, thus giving more reliable understandings for teachers and administrators.

The study addresses the challenge of accurately predicting student performance in educational settings. There are a lot of available datasets but still have limitations in prediction accuracy and there are many presented predictive models that fail to achieve high accuracy due to limitations in their algorithms design and the failure to manage the complication and variability of educational data.

The aim of this paper to proposed a combination of three ML algorithms (RF, ET, and KNN). These algorithms are selected depending on their ability to get higher accuracy with better flexibility to different types of educational dataset, and ability to scale efficiently with large datasets.

This research proposed a modern method to predicting student performance utilizing ML algorithms for address limitations of current studies. The main contribution in this study shows in following steps:

- Implementation of a Hybrid Model: This proposed have developed combined three ML algorithms (RF, Extra Trees, and KNN). This hybrid approach proposed for improves prediction accuracy and efficiency.
- High Accuracy in Different Settings: This study supports accuracy across different educational settings. This flexibility makes it a valuable tool for educational institutions with different data characteristics.
- Effective Handling of Large Datasets: The proposed solution is designed to efficiently process and analyze large volumes of educational data, overcoming the scalability issues faced by many existing models.

The remainder of this paper is organized as follows:

- Related work: Section 2 provides a comprehensive review of the literature.
- Proposed Model: Section 3 details the methodology adopted in this study.

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

- Results and Discussion: Section 4 presents the results obtained from the implementation of the algorithms.
- Conclusion: Section 5 concludes the paper. It summarizes the key findings, reiterates the significance of the study.

Related work

Several studies have been conducted in the field of data mining with the aim of enhancing the educational system and securing a prosperous future for students. Studies conducted between 2017 and 2023 are mentioned here, which focus on predicting student performance through machine learning techniques.

Al-Marabah suggested using different classification and data mining(DM) methods to examine and predict the performance of university students. Using multiple performance indicators, the article compared five classifiers, including Naïve Bayes(NB), Neural Network, J48, ID3, and Bayesian Network, using the WEKA tool. The research discovered that the Bayesian network classifier outperformed the other classifiers in terms of accuracy. This study demonstrates how DM can be used to improve educational outcomes by evaluating student performance and predicting future events[9].

Al-Shehri H. et al. [10] suggested, by investigating improved models, an enhance student performance prediction models. The study made use of a landmark data set on student performance in mathematics from Minho University. I checked the performance of two algorithms, SVM and KNN, using the Weka tool. The SVM performed marginally better, with a correlation coefficient of 0.96, than the K-NNN approach, which had a correlation coefficient of 0.95. This study stresses the need to experiment with different algorithms in order to increase the accuracy of prediction models. Furthermore, the results of the study may be useful for adopting early precautions, taking action in a timely manner, or selecting the best student for a particular task.

Tanwar E. et al. [11] this investigation proposed utilizing historical data of Bina Nusantara University graduates. The methods used in this trial encompass generalized linear models, deep learning (DL), and DT. The objective is to identify the key elements that impact the final deduction, aiding students in proactive readiness. The resultant DT emerges as the most straightforward model for effectively communicating to users.

Hussein M et al. [12] introduced a model aimed at predicting forthcoming challenges students might encounter in the subsequent session of a digital design course. They achieved this by employing ML algorithms on data sourced from an advanced learning system named DEEDS. Within individual sessions of the Digital Design course, input data encompassed various metrics such as average time, total activity count, average idle time, mean keystroke count, and total relevant activity per exercise. The study used various ML techniques including artificial neural networks (ANN), SVM, logistic regression (LR), NB classifiers, and DT. The outcomes indicated that ANN and SVM surpassed other algorithms in terms of predictive accuracy. These superior algorithms can be straightforwardly integrated into the TEL system to enhance students' performance in subsequent sessions.

Hammoud A.K. et al. [13] proposed the creation of a model based on decision tree algorithms to recommend characteristics that improve student performance. The study obtained data from 161 students through a questionnaire that included 60 questions about their health, social activities, relationships, and academic performance. For data analysis, three classifiers were used: J48, Random Tree, and REPTree. This form was created using the Weka 3.8 tool. Based on its performance, the study determined that J48 was a better algorithm than the other two. The study shows how data mining algorithms and DT algorithms can be used to detect hidden trends and make recommendations to improve student performance in educational institutions.

Burgos C, et al. [14] presented a model to predict student dropout from e-learning courses using LR modelling and classification. In this paper, the data collected from the Moodle database of the Open University of Madrid, and the study suggests using the LOGIT Act knowledge discovery system model for this purpose. The model achieves an accuracy of 97.13%.

Nagy and R. Molontai [15] suggested using ML algorithms to predict student performance and avoid failure by detecting at-risk students. The research is

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

established on data collected from more than 15,000 undergraduate students at a single university in Budapest. The authors create prediction models based on available data at the time of enrollment, such as high school achievement and personal information, using various ML methods such as decision tree-based, NB, K-NN, linear models, and DL algorithms. The models were tested using 10-fold cross-validation; The best models, gradient-augmented trees, and DL have AUCs of 0.808 and 0.811, respectively.

Vijayalakshmi V, and Venkatachalapathy K. [16] Dada introduced a deep neural network-based student performing forecast model. Educational data mining is a field that uses data mining principles and algorithms to examine and enhance students' academic performance. ML algorithms have become indispensable in almost every industry, including educational data mining. They trained and evaluated the model using the Kaggle dataset and compared the precision of several methods in R programming, such as DT (C5.0), SVM, RF, NB , KNN, and Deep Neural Network(DNN). The DNN algorithm outperformed all other algorithms with an accuracy rate of 84%.

Wahid H, et al. [17] discussed the utilization of learning analytics in virtual learning environments to proactively identify students who are at risk and implement early intervention measures. The study employed a deep artificial neural network that utilized manually gathered click data to reach a classification accuracy ranging from 84% to 93%. According to the research, the neural network achieves higher accuracy than the logistic and base regression models, with accuracy ranging from 79.82% to 85.60% and 79.95% to 89.14%, respectively. The objective of the article was to assist institutions in establishing the fundamental structure for educational support and streamlining decision-making processes in higher education with the aim of promoting sustainable education.

Hassan R, et al. [18] proposed a data mining (DM) and video learning analytics to predict performance of 772 students registered in e-Commerce technology modules and e-commerce at a higher education institution (HEI). This study applied eight classification algorithms to analyze data from a student knowledge system, a mobile application, and a learning management system. In order to further minimize the features, the data experienced transformation and preprocessing, followed by the application of genetic search and primary component analysis. The RF algorithm accuracy predicted the success of students at the conclusion of the semester with a precision of 88.3% when utilising an equal combination of presentation and knowledge gain. This study shown utilizations of video learning analytics and data mining techniques to forecast student achievement in higher education institutions.

Kemper L, et al. [19] proposed the use of two ML methodologies, namely, LR and DT, to predict student dropouts at the Karlsruhe Institute of Technology (KIT). The proposed are built using the examination data, which is freely available in all institutions and does not require any special collection. As a result, this study offers a systematic approach that can be easily implemented in other organisations. This work find that DT only marginally outperforms LR. However, after three semesters, both methods produce an excellent prediction accuracy of up to 95%.

Mubarak A, et al. [20] suggested using video click data to examine the learning behavior of MOOC enthusiasts and predict their success in MOOC courses. Based on a set of implicit properties derived from video streaming data, they propose a DNN model (LSTM) to predict the weekly performance of learners. The goal is to enable trainers to implement measures that will allow them to work in a timely manner and improve the educational process. By evaluating streaming video data and predicting student performance, the work solves the challenge of classifying time series. In the real-cycle data sets used, the proposed LSTM model outperformed simple ANN, SVM, and LR, obtaining an accuracy of 93%.

Sukry S, and Saleh A. [21] proposed predicting student model success in a cloud computing course and identifying at-risk individuals early on. The proposed model is RHEM (Strong Hybrid Array Model), which combines four individual algorithms (NB, Multilayer Perceptron(MLP), KNN, and DT) with four aggregate algorithms (Packaging, Multiclass Classifier Random Subspace, and Rotational Forest). 24 models were developed, trained, and tested to accurately evaluate the performance of the RHEM model. Based on evaluation metrics such as accuracy (91.70%), accuracy (86.1%), F-score rate (87.3%), and receiver

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

operating characteristic area determination (98.6%), the RF + MLP model was the best.

Al-Hassan et al. [22] this study proposed correlation relating online LMS activity data and evaluation scores on students' performance. This study utilized different ML algorithms like sequential minimum optimisation (SMO), LR, MLP, DT (J48), and RF. The data collected utilized from the Distance Education at King Abdulaziz University and Deanship of E-Learning. The results shows the RF algorithm achieved the highest accuracy of 99.17%.

Adnan et al. [23] proposed a predictive model for identifying at risk students on online platforms learning, with virtual learning environments (VLEs), learning management systems (LMS), and open online courses (MOOCs). Various ML and DL techniques are used to train and test the model, and the best-performing method, Random Forest (RF), is used to generate the prediction model. The study is based on examining study characteristics such as rating scores, engagement intensity (click stream data), and time-dependent variables among students. The results reveal that the radio-frequency-based prediction model has accuracy, recall, and F-score at different percentages of the course length, which indicates that it may identify at-risk students initial in the courses for timely involvement to prevent students from dropping out.

Rodriguez Hernandez S. et al. [24] performed an Artificial Neural Network (ANN) test to forecast academic performance in higher education. Furthermore, they examined the significance of many factors that influence academic success in tertiary education. The study sample included 162,030 students of diverse ethnic backgrounds enrolled in both private and public universities affiliated with Columbia University. Researchers have found that employing Artificial Neural Networks (ANN) can effectively classify students' academic performance as either high or low, with an accuracy rate of 82% and 71% respectively. The finding of this study was found ANN outperformed ML techniques in evaluation measures such as the F1 score and recall.

Kumar M, et al. [25] proposed a prediction model based on a historical dataset of student academic achievement. This study presents two important contributions the first step is to create the prediction model by applying various ML algorithms to the dataset and the second contribution to proposed combining different ML methods like Bagging, AdaBoostM1, and Random Subspace are between the cluster meta-models evaluated. The results of this proposed shows the combined with a layered ML approach, the AdaBoostM1 meta-based technology performed best, achieving the highest accuracy of 80.33%. This results shows the power of applying ML and data mining approaches to predict academic performance is important implications for enhancing education and student achievement.

Yac et al. [26] proposed a novel ML model to forecast the final examination results for undergraduate students. The primary source of data for this model was the midterm examination scores. This study compares the performance of different ML techniques, such as RF, nearest neighbors, SVM, LR, NB, and KNN. In this study utilized dataset that contained of academic performance records of 1854 students who were registered in Turkish Language Course 1 at a Turkish state university. This study utilized only three types of input parameters midterm of examination scores, department-specific information, and faculty-related data. This proposed model achieved an accuracy between 70% to 75%. The findings of this study to efficiency of ML methods in forecasting the academic achievements of undergraduate students.

Al-Bouaneen et al. [27] this study proposed ML algorithms for forecast academic performance and identify students who may be at risk of failing. This study proposed focuses on female students registered in the Department of Computer Science at Imam Abdulrahman Bin Faisal University. This study proposed various ML algorithms (SVM, RF, KNN, ANN, and LR) to predict the overall course score at an early stage. This study used a dataset that contained 842 cases with included 168 individuals. The web-based prediction system developed as part of this research has an average absolute error rate of 6.34% and is readily available to educators via an Internet server.

Gaftandzhieva et al. [28] proposed a Moodle Learning Management System (LMS) and Zoom data to predict the final grades of students who are taking the object-oriented programming course at the University of Plovdiv. This study utilized a dataset have contained the final grades, online course activities, and lecture attendance records of 105 students. This study proposed chi-square tests and

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

logistic regression tests to study the correlation between scores and the online activity environment. Machine learning algorithms such as RF, XGBoost, KNN, and SVM were utilized. The RF shows the highest level of prediction accuracy, achieving 78%. The research highlights importance of utilizing data-driven forecasts to help educators and decision-makers identify kids who are at risk and improve their general academic performance.

Abdullah et al. [29] proposed a novel approach to prediction academic achievements of students registered in online training courses, utilizing electronic learning recorded as the primary data source. This study collected a range of variables, encompassing e-learning history, demographic details, and previous academic records, for a sample of undergraduate students from a university in Jordan. This proposed applied various ML techniques (RF, Bayesian Ridge (BR), AdaBoost, and XGBoost). The results of this study showed highest performance was achieved by the combined model of RF and XGBoost. The finding of this reserch that e-learning history data, which include metrics such as login frequency, course participation, and forum participation, shows as a large predictor of the academic performance. Table 1 shows a summary of related work:

Table 1. Summary of Related Work.

Ref.	Data Used	Algorithms/Methods / Accuracy
9	Multiple performance indicators of university students	NB: 91.11%, BN: 92.0%, ID3: 88.0%, J48: 91.11%, Neural Network: 90.2%
10	Student performance in mathematics from the University of Minho	SVM: 0.95%, KNN: 0.96%
11	Alumni data from Bina Nusantara University	Generalised linear model: 66.6%, DL: 67.6%, DT: 60.6%
12	Data from a technologically improved learning system (DEEDS)	ANN: 75%, SVM: 75%, LR: 73%, NB: 75%, DT: 69%
13	Survey data from 161 students	J48: N/A, RT: N/A, REPTree: N/A
14	Data from the Moodle database of the Open University of Madrid	LR modelling: 97.13%, SEDM: 94.23%, FFNN: 85.58%, PESFAM: 70.19%, SVM: 62.50%
15	Data from over 15,000 undergraduate students	DT: 63%, RF: 65.5%, LR: 70.3%, NB: 68.3%, k-NN: 69%, linear models: 67%, DL:73.5%, Gradient Boosted Trees: 70.6%, Adaptive Boost: 68.8%
16	Kaggle dataset	DT (C5.0): 69%, NB: 73%, RF: 79%, SVM: 75%, KNN: 69%, DNN: 84%
17	virtual learning environments	DNN: 84%, LogLR: 93%
18	Student information system, learning management system, and mobile applications data	Classification Tree, RF: 88.3% , KNN, SVM, LR, NB, Neural Network, CN2 Rule: 87.4%.
19	Examination data	LR: A/N, DT: 95%
20	Video click data	LSTM: 93%
21	Cloud computing course data	RHEM (Robust Hybrid Ensemble Mode: 91.70%
22	LMS online activity data	Sequential Minimum Optimization (SMO):99.17% , LR97.04%, MLP: 98.08%, DT (J48): 98.75 , RF: 99.17
23	Various online learning platforms' data	RF: N/A
24	Data from private and public colleges in Colombia	ANN: 82%
25	Historical dataset of student academic achievement	MLP: 78.33%, NB: 67.70%, Locally Weighted Learning: 66.04%, DT: 72.70%, J48: 75.83%
26	Midterm exam results of undergraduate students	RF: 0.746 , KNN: 0.699 , SVM: 0.735, LR: 0.717, NB: 0.713
27	Academic and demographic characteristics of female students	SVM: N/A, Random Forest: A/N, KNN: N/A, ANN: N/A, LR: N/A
28	Moodle LMS and Zoom data	RF: 78%, Extreme Gradient Boosting: 76%, KNN: 70% , and SVM: 73%
29	E-learning records of undergraduate students	RF: N/A, Bayesian hills: N/A, Adaptive Boosting: N/A, Extreme Gradient Boosting:N/A

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Proposed Model

This study utilization of ML algorithms to forecast the probability of a student's success, possible dropout, or addition in their academic activities. The particular emphasis of this paper lies in contrasting various ML techniques and addressing the challenge of class imbalance, evaluating their respective contributions to enhancing predictive accuracy. This paper uses three ML algorithms to classify student outcomes. This section provides a comprehensive elucidation of the system architecture that has been put forward. This architecture encompasses various elements, namely data collection, pre-processing, classification algorithms, and performance metrics. The configuration of this model is depicted, and the procedures for its execution are elucidated in Fig. 1.

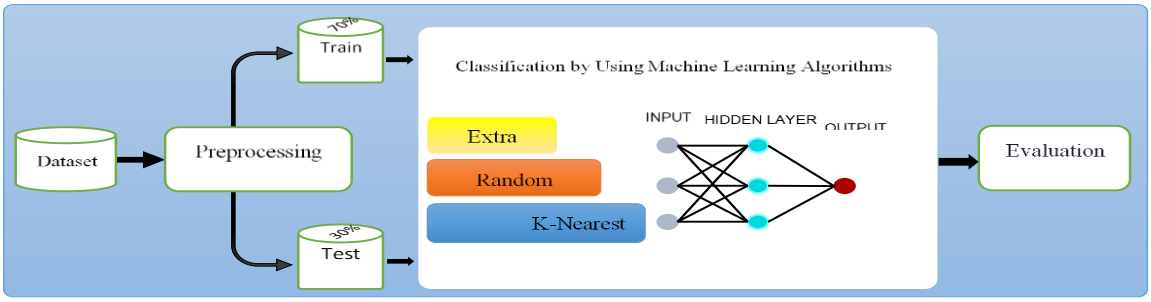

Figure 1. The Flow Chart of the Proposed.

The proposed phases of the flowchart are explained in detail in the following sections:

Supervised Machine Learning Algorithms in Student Performance Prediction

Supervised ML is a subset of machine learning where algorithms are trained on labeled datasets[30,31]. These datasets consist of input-output pairs, where the model learns to map inputs to the desired output. In the context of predicting student performance, the input could be various student-related factors (like attendance, grades in previous exams, participation in class, socio-economic background, etc.), and the output would be the predicted performance (such as grades in upcoming exams or overall academic success). This study used supervised ML for predicting student performance involves several key steps:

A. Data Collection.
B. Feature Selection.
C. Model Training.
D. Model Evaluation.
E. Prediction.

This study employed three supervised ML algorithms: Random Forest, Extra Trees, and K-Nearest Neighbors. In our study, the model construction process was meticulously undertaken with the aim of developing a robust and accurate model for predicting student performance. This involved careful selection and preparation of data, strategic choice and tuning of algorithms, and thorough evaluation and refinement to ensure the model's effectiveness and reliability.

Dataset Description and Data Collection

In this proposed used dataset from numerous global universities spanning the academic years 2021 and 2022. The data originated from a survey conducted among remote students. Encompassing a total of 30,000 student records, the dataset comprises seventeen distinct attributes. These attributes can be categorized into five groups: personal and lifestyle, studying approach, familial connections, satisfaction with the educational environment, and student academic performance. Table 2 presents a breakdown of the attributes employed in formulating the dataset. The student has been categorized Individually as "pass," "fail," or "drop out" based on their final academic outcome so the passing student is the student who has successfully completed the course or program and achieved the minimum required grade or percentage to be considered passing while the failing student has not achieved the minimum required grade or percentage to pass the course or program. In this case the student may need to retake the course or improve performance to meet the passing criteria. The drop-out student is a student who has voluntarily or involuntarily withdrawn from the course or program before its completion. This might be due to various reasons, such as personal issues, financial constraints, or academic difficulties.

Dataset Pre-processing

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

This step is important as it involves modifying and preparing the data. The goal of data preparation is to reduce the quantity of information. Preprocessing is a method used to prepare data for analysis, as raw data is of limited value for analysis. To prevent overfitting, this study performed preprocessing on the data before feeding it into a classification system. Applying preprocessing is a fundamental requirement in constructing a prediction model that yields precise outcomes, constituting a vital phase in the process. The preprocessing techniques applied in this paper are as follows:

Conversion to Strings

The dataset contains a range of different data types. For consistency and to ensure that the Label Encoder can effectively encode all features, this study first converts all data columns to strings using the 'astype' function in Pandas.

Label Encoding

ML algorithms typically require numerical input. Nonetheless, the dataset this research possess comprises categorical attributes denoted as strings. To handle this situation, in this proposed employ the Label Encoder provided by the 'sklearn.preprocessing' module. The Label Encoder serves as a convenient tool to transform labels into a normalized format, ensuring they exclusively encompass values ranging from 0 to n_classes-1. This is done for both the feature columns and the target column.

Feature Scaling

ML algorithms exhibit suboptimal performance when the numerical attributes within the input display significant dissimilarity in scales. To mitigate this issue, this study employ the "StandardScaler," a preprocessing tool available in the "sklearn.preprocessing" module. This utility standardizes the features of the dataset to adhere to a uniform scale (with a mean of 0 and a variance of 1), a prerequisite for achieving optimal effectiveness across numerous ML algorithms.

Handling Imbalanced Classes with SMOTEENN

Class imbalance is a prevalent issue encountered in ML classification tasks, characterized by an uneven distribution of observations among different classes. This study used the SMOTEENN technique to combat this problem. The "SMOTEENN" is a combination of over-sampling the smaller class utilizing Synthetic Minority Over-sampling Technique (SMOTE) and cleaning the results over-sampled dataset with Edited Nearest Neighbours (ENN) a type of under-sampling method.

Splitting the Data

The dataset was split into two subsets, a training and a testing, utilizing the 'train_test_split' function sourced from the 'model_selection' module within 'sklearn.'. The training set works as the data utilized to train the ML models, while the test set is applied to evaluate the performance of these models. The data was split to 80% of training set, while 20% for testing. Additionally, the 'random_state' parameter was configured to a fixed value, guaranteeing the reproducibility of the generated data splits.

Applying Machine Learning Algorithms

The fourth phase in constructing a problem-solving model involves the utilization of ML methods. This study introduced three distinct ML algorithms (Extra Trees, RF, and KNN) along with ensemble techniques to forecast student performance via ML. These algorithmic selections were made for the purpose of contrasting them with prior research endeavors focused on identical research question. Subsequently, each of these algorithms was implemented within the system.

The paper employs three supervised algorithms, including Extra trees, RF, and KNN. The algorithms in question are elucidated in the subsequent subsection:

Extra trees

The Extra Trees Classifier is a ML algorithm that falls within the category of ensemble methods. The method in question is a derivative of the Random Forest algorithm, exhibiting notable similarities with the latter. The term "Extra Trees" is an abbreviation for "Extremely Randomized Trees," which denotes the utilization of randomization techniques in the process of constructing individual decision trees [32],as shown in Fig. 2.

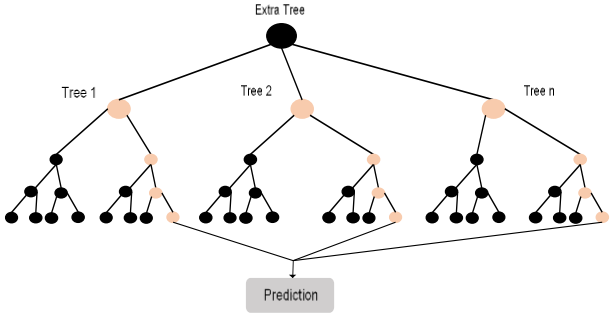

Figure 2. The structure of Extra Tree[33].

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Like Random Forest, the Extra Trees Classifier also operates by creating a collection of DT and combining their predictions to make a final classification decision. However, there are a few key differences between the two algorithms [32].

The main distinction lies in the construction of the individual decision trees. In Random Forest, the decision trees are built using the concept of random feature subsets, where a random subset of features is considered at each node to determine the best split [33]. In contrast, Extra Trees takes the randomization a step further by considering random splits for each feature at each node rather than finding the optimal split based on a chosen criterion such as Gini impurity or information gain, Extra Trees randomly selects splits and chooses the best among them this increased level of random helps to reduce the variance of the model but can potentially increase the bias.

The random presented in Extra Trees has some advantages the first, it speeds up the training process since no time is spent searching for the optimal split at each node while the second, it makes the model less prone to overfitting, as the random adds variety to the decision trees. This random may result in some higher bias compared to Random Forest, as it increases the chances of selecting suboptimal splits.

The Extra Trees Classifier is a flexible algorithm efficient of performing both classification and regression tasks. It has achieved significant popularity across varied fields such as finance, healthcare, and natural language processing. This approach shows to be particularly useful when working with datasets that have a high number of dimensions or contain features that are prone to noise. The method derives advantages from the aggregated decisions of several decision trees, thus enhancing its flexibility to outliers and noisy data [34]. When training an Extra Trees Classifier, it is important to believe the number of trees in the ensemble and the maximum depth of each tree, which can affect the balance between bias and variance. The hyperparameter tuning and cross-validation techniques can be applied to optimize the performance of the model. The Extra Trees Classifier is an ensemble learning algorithm that creates decision trees with increased random in the splitting process. It offers a trade-off between model complication, bias, and variance, making it a valuable tool for various ML tasks[35]. Utilizing Class Imbalance help us to achieve the highest accuracy of 98.15% in predicting student performance.

Random Forest

This algorithm is a supervised ML algorithm utilized for regression and, classification which is fast to work and easy to implement. The source of the work of the RF is a decision tree that is generated by selecting the root from a set of features, and then the root node is divided into various sections depending on the entropy calculation for each feature[36].
The major idea of this approach is to generate a set of features using small decision trees, which are computationally efficient procedures so if it were possible to construct multiple small decision trees parallel, the resulting trees could be combined into a single robust learner by a combination or majority voting approach[37]. Fig. 3 shows the main idea of RF algorithm.

The Random Forest technique offers several advantages. Firstly, it is applicable to both regression and classification tasks. Secondly, it possesses the capability to effectively manage missing values, ensuring accuracy even in the presence of incomplete data. Lastly, it is capable of effectively managing huge datasets with high dimensionality[38].

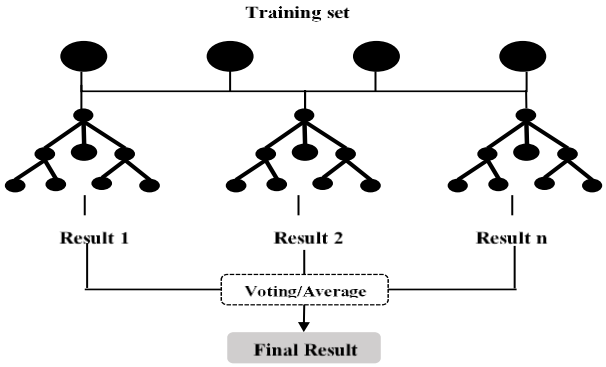

Figure 3. The structure of RF

In our study, the Random Forest classifier was meticulously tuned and applied to predict student performance due to Balanced Dataset and Quality of Data.

K-Nearest Neighbor

This algorithm is one of the supervised ML algorithms and is considered more straightforward than the others. Since learning in this algorithm is supervised, the samples in each vector must be named for training samples. Then, in the classification phase determining the test sample for

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

the classes it belongs to, the distance between the training sample points and the new sample point must be calculated, and the k value that represents the depth of the algorithm must be determined, that is, the number of points in the classification process used in the classification process. Finally, the appropriate classes are selected for each new sample point, depending on most of the neighbours' votes [39].
The value $\boldsymbol{k}$ must be determined before starting the classification because it determines the scope of training (number of neighbours) according to the number of samples present. The distance between the new data point and the training data points for each neighbour is measured using different measurements, including the manhattan and euclidean distance measurements[40].
Where $\boldsymbol{d}$ is the distance between the new data point and the training data points, the $\boldsymbol{t}$ is the number of the training sample, the $\boldsymbol{x}$ new data points, and the $\boldsymbol{y}$ is the training data points; the new data point is assigned to the class with the most significant number of nearest neighbors[41].
In our study, the Random Forest classifier was meticulously tuned and applied to predict student performance due to Balanced Dataset and Quality of Data.

Performance Metrics

This section encompasses the metrics employed for assessing the performance of ML algorithms. In the case of classification algorithms, the evaluation relies on the following metrics: accuracy, recall, precision, f1_score, execution time, and error rate. The computation of these metrics is contingent upon the utilization of the confusion matrix, which encompasses three distinct predictions [42] , as shown in Fig. 4.

Multiclass Confusion Matrix		Predicted		
		S_1	S_2	S_3
Actual	S_1	TP	FN	FN
	S_2	FP	TN	TN
	S_3	FP	TN	TN

Figure 4. The Confusion Matrix.

The descriptions of the confusion matrix predictions are as follows [43] :

TP: The positive samples that have been correctly classified [44]. TP is formally defined in Eq. 1.

$$TP\ Rate\ = \frac{TP}{(TP + FN)} * 100 \quad 1$$

TN: The negative samples that have been correctly classified. TN is formally defined in Eq. 2.

$$TN\ Rate\ = \frac{TN}{(TN + FP)} * 100 \quad 2$$

FN: The positive samples that have been incorrectly classified as negative samples. FN is formally defined in Eq. 3.

$$FN\ Rate\ = \frac{FN}{(FN + TP)} * 100\ \% \quad 3$$

FP: The negative samples have been incorrectly classified as positive samples. FP is formally defined in Eq. 4.

$$FP\ Rate\ = \frac{FP}{(FP + TN)} * 100\ \% \quad 4$$

Additional evaluation metrics are derived from the predictions within the confusion matrix. Here's an explanation of these metrics:
Accuracy Rate: The accuracy rate is the ratio of samples correctly classified from the tested samples. That is, correct classifications of negative and positive samples are detected [45]. It is formally defined in Eq. 5.

$$Accuracy\ Rate = \frac{TN + TP}{TN + TP + FN + FP} * 100\ \% \quad 5$$

Recall Rate: Measuring the number of correctly classified positive samples, also called sensitivity; when the recall rate is high, the number of positive samples classified as negative is few. It is formally defined in Eq. 6.

$$Recall\ Rate\ = \frac{TP}{(FN + TP)} * 100\ \% \quad 6$$

Precision Rate: The ratio of true positives to the sum of true positives and false positives to determine the number of unwanted positives. If the false positives are FP=0, then the algorithm's accuracy is 100%, and the ratio decreases as the value of FP increases[46]. It is formally defined in Eq. 7.

$$Precision\ Rate\ = \frac{TP}{(FP + TP)} * 100\ \% \quad 7$$

f1_score Rate: It Combines both accuracy and recall and scales them to obtain precise predictive accuracy and is formally defined in Eq. 8.

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

$$f1\_score = \frac{2 * (Precision) * (Recall)}{(Recall + Precision)} * 100\ \% \quad 8$$

Results and Discussion

Within this section, this proposed unveils the outcomes stemming from the utilization of three distinct ML algorithms: Extra Trees, RF, and KNN. The primary objective was to forecast student performance, leveraging a dataset comprising seventeen attributes originating from personal and lifestyle factors, study habits, familial considerations, satisfaction with the educational environment, and the student's grades.

Table 2 succinctly encapsulates the pivotal metrics that were assessed for each algorithm. These metrics encompass accuracy, precision, recall, F1 score, and the duration required for training.

Table 2. Performance of Machine Learning Algorithms

Model	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)	Training time (sec)
Extra Trees	98.15%	98.14%	98.15%	98.15%	6.07 s
Random Forest	94.03%	94.09%	94.42%	94.03%	6.32 s
K-Nearest Neighbors	91.65%	91.21%	91.91%	91.65%	0.00 s

The Extra Trees algorithm achieved the highest accuracy of 98.15%, closely followed by the RF algorithm at 94.03%, and the KNN algorithm at 91.65%. Precision and recall metrics, which are crucial for evaluating the true predictive power of the models, also showed high values across all three algorithms, confirming their robustness.

In terms of computational efficiency, the KNN algorithm stands out with a significantly lower training time of 0.00 seconds, despite having a slightly lower performance in terms of accuracy, precision, recall, and F1-score, as shown in Fig. 5.

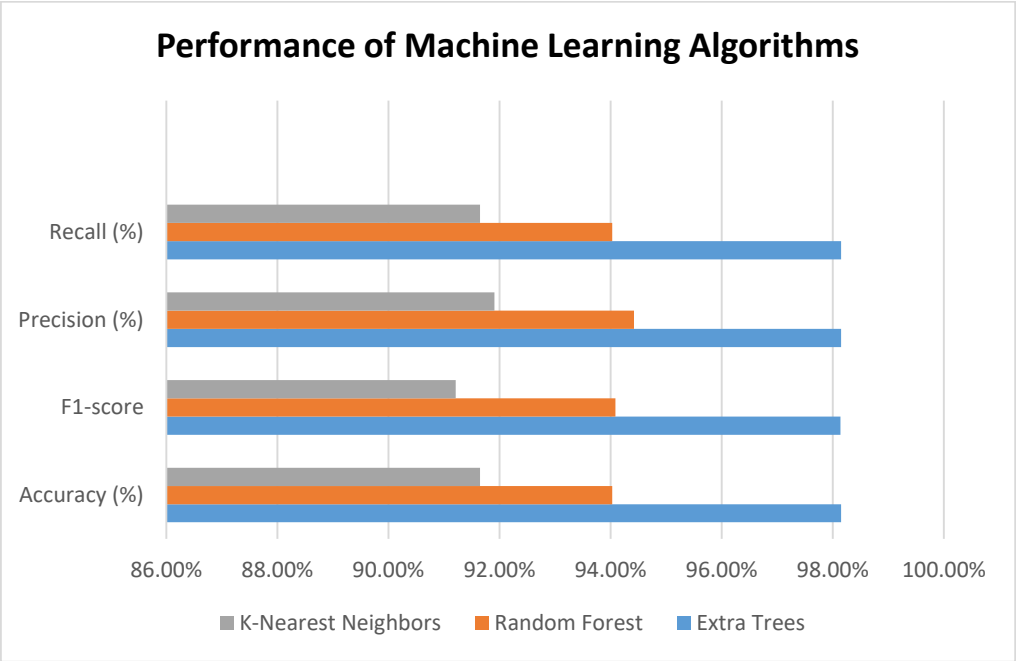

Figure 5. Comparative Performance Analysis of Machine Learning Algorithms.

The confusion matrices for each algorithm were also computed to provide a more detailed view of the performance of each model, as shown in the following Figs. 6, 7, and 8:

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

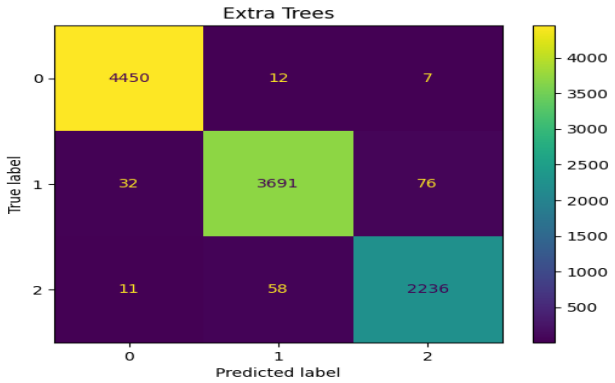

Figure 6. Confusion Matrix of ET.

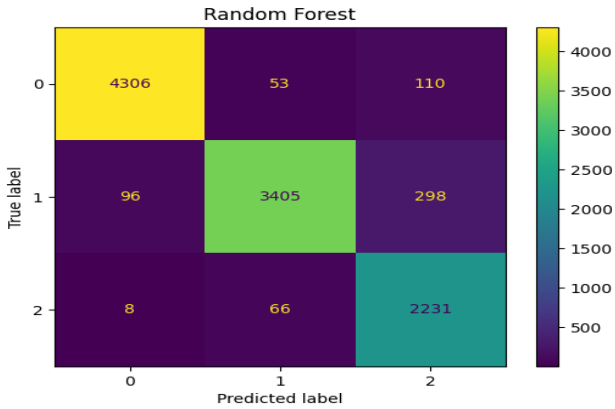

Figure 7. Confusion Matrix of RF.

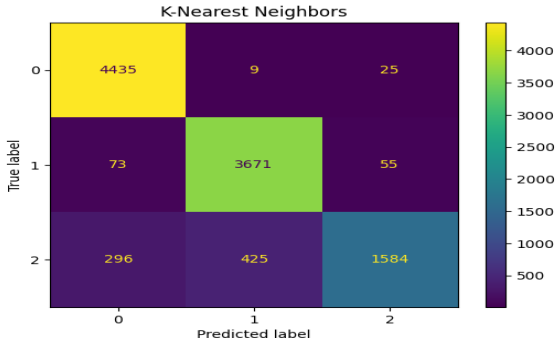

Figure 8. Confusion Matrix of KNN.

The central research problem of this paper was to predict student performance using ML. The questions that guided this research: Can ML models accurately predict student performance, and which attributes are most influential in these predictions.

The results shown the Extra Trees, RF, and KNN have good performed in predicting student performance, with achieved an accuracy of 98.15%, 94.03%, and 91.65%, respectively. The precision, recall, and F1 scores of these algorithms also show a good predicting student performance. This supports strong evidence to application of ML in making data-driven decisions regarding student performance.

Conclusion

The finding of this study to develop an application efficient of predicting student performance. This study built a model to predict abilities of ML algorithms to provide teachers with a powerful tool that can advance in making data-driven decisions. This study utilized combined model, including Extra Trees, RF, and KNN have shown their efficacy in predicting student performance, supporting the capability of ML in this area. This can be an effect of the critical factors academic success can guide teachers and institutions in designing strategies and interventions that cater to students' individual needs. This paper showed the ML models can be predicted on student performance accurately. The results achieved from the Extra Trees, RF, and KNN algorithms ensure that ML can be an effective tool in helping teachers in making data-driven decisions. In future research needs to focus on applying and testing the model across different educational settings and cultures to establish its universality and adaptability.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Anbar.
- Ethics statement:
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- No potentially identified images or data are present in the manuscript.

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Authors' Contribution Statement

M.S.I.A, A.A.N, A.A.M, R.D.J, M.M.A contributed to the working and proposed of the research, to the examination results, and to the writing of the manuscript.

References

1. Farrington CA, Roderick M, Allensworth E, Nagaoka J, Keyes TS, Johnson DW, et al. Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance--A Critical Literature Review. ERIC. 2012; 38 p.
2. Hale EL, Moorman HN. Preparing school principals: A national perspective on policy and program innovations. ERIC. 2003; p 11.
3. Del Río S, López V, Benítez JM, Herrera F. On the use of mapreduce for imbalanced big data using random forest. Inf Sci (Ny). 2014; 285: 112–37. https://doi.org/10.1016/j.ins.2014.03.043
4. Thai-Nghe N, Drumond L, Horváth T, Krohn-Grimberghe A, Nanopoulos A, Schmidt-Thieme L. Factorization techniques for predicting student performance. In: Educational recommender systems and technologies: Practices and challenges. IGI Global. 2012; p. 129–53. https://doi.org/10.4018/978-1-61350-489-5.ch006
5. Daud A, Aljohani NR, Abbasi RA, Lytras MD, Abbas F, Alowibdi JS. Predicting student performance using advanced learning analytics. In: Proc 26th int conf world wide web comp. 2017; p. 415–21. https://doi.org/10.1145/3041021.3054164
6. Nafea AA, Mishlish M, Muwafaq A, Shaban S, Al-ani MM, Alheeti KMA, et al. Enhancing Student ' s Performance Classification Using Ensemble Modeling. Iraqi J Comput Sci Math. 2023; 4(4): 204–14. https://doi.org/10.52866/ijcsm.2023.04.04.016
7. Xu Z, Yuan H, Liu Q. Student performance prediction based on blended learning. IEEE Trans Educ. 2020; 64(1): 66–73. https://doi.org/10.1109/TE.2020.3008751
8. Pallathadka H, Wenda A, Ramirez-Asís E, Asís-López M, Flores-Albornoz J, Phasinam K. Classification and prediction of student performance data using various machine learning algorithms. Mater today Proc. 2023; 80: 3782–5. https://doi.org/10.1016/j.matpr.2021.07.382
9. Almarabeh H. Analysis of students' performance by using different data mining classifiers. Int J Mod Educ Comput Sci. 2017; 9(8): 9. https://doi.org/10.5815/ijmecs.2017.08.02
10. Al-Shehri H, Al-Qarni A, Al-Saati L, Batoaq A, Badukhen H, Alrashed S, et al. Student performance prediction using support vector machine and k-nearest neighbor. IEEE. Canadian Conference on Electrical and Computer Engineering (CCECE). 2017; p 1–4. https://doi.org/10.1109/CCECE.2017.7946847
11. Tanuar E, Heryadi Y, Abbas BS, Gaol FL. Using machine learning techniques to earlier predict student's performance. IEEE. Indonesian Association for Pattern Recognition International Conference (INAPR). 2018;p. 85–9. https://doi.org/10.1109/INAPR.2018.8626856
12. Hussain M, Zhu W, Zhang W, Abidi SMR, Ali S. Using machine learning to predict student difficulties from learning session data. Artif Intell Rev. 2019; 52: 381–407. https://doi.org/10.14569/IJACSA.2016.070531
13. Hamoud A, Hashim AS, Awadh WA. Predicting student performance in higher education institutions using decision tree analysis. Int J Interact Multimed Artif Intell. 2018; 5: 26–31. https://ssrn.com/abstract=3243704
14. Burgos C, Campanario ML, de la Peña D, Lara JA, Lizcano D, Martínez MA. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Comput Electr Eng. 2018; 66: 541–56. https://doi.org/10.1016/j.compeleceng.2017.03.005
15. Nagy M, Molontay R. Predicting dropout in higher education based on secondary school performance. IEEE International Conference on Intelligent Engineering Systems (INES).. 2018;p. 389–94. https://doi.org/10.1109/INES.2018.8523888
16. Vijayalakshmi V, Venkatachalapathy K. Comparison of predicting student's performance using machine learning algorithms. Int J Intell Syst Appl. 2019; 11(12): 34. https://doi.org/10.5815/ijisa.2019.12.04
17. Waheed H, Hassan SU, Aljohani NR, Hardman J, Nawaz R. Predicting Academic Performance of Students from VLE Big Data using Deep Learning Models. Computers in Human behavior, 2020, 104: 106189. http://dx.doi.org/10.1016/j.chb.2019.106189
18. Hasan R, Palaniappan S, Mahmood S, Abbas A, Sarker KU, Sattar MU. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Appl Sci. 2020; 10(11): 3894. https://doi.org/10.3390/app10113894
19. Kemper L, Vorhoff G, Wigger BU. Predicting student dropout: A machine learning approach. Eur J High Educ. 2020; 10(1): 28–47. https://doi.org/10.1080/21568235.2020.1718520
20. Mubarak AA, Cao H, Ahmed SAM. Predictive learning analytics using deep learning model in MOOCs' courses videos. Educ Inf Technol. 2021;

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

26(1): 371–92. https://doi.org/10.1007/s10639-020-10273-6
21. Sakri S, Alluhaidan AS. RHEM: A robust hybrid ensemble model for students' performance assessment on cloud computing course. Int J Adv Comput Sci Appl. 2020; 11: 388–96. https://doi.org/10.14569/IJACSA.2020.0111150
22. Alhassan A, Zafar B, Mueen A. Predict students' academic performance based on their assessment grades and online activity data. Int J Adv Comput Sci Appl. 2020; 11(4). https://doi.org/10.14569/IJACSA.2020.0110425
23. Adnan M, Habib A, Ashraf J, Mussadiq S, Raza AA, Abid M, et al. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. Ieee Access. 2021; 9: 7519–39. https://doi.org/10.1109/ACCESS.2021.3049446
24. Rodríguez-Hernández CF, Musso M, Kyndt E, Cascallar E. Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. Comput Educ Artif Intell. 2021; 2: 100018. https://doi.org/10.1016/j.caeai.2021.100018
25. Kumar M, Mehta G, Nayar N, Sharma A. EMT: Ensemble meta-based tree model for predicting student performance in academics. IOP Conf Ser.: Mater Sci Eng. IOP Publishing; 2021. p. 12062. https://doi.org/10.1088/1757-899X/1022/1/012062
26. Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn Environ. 2022; 9(1): 11. https://doi.org/10.1186/s40561-022-00192-z
27. Alboaneen D, Almelihi M, Alsubaie R, Alghamdi R, Alshehri L, Alharthi R. Development of a web-based prediction system for students' academic performance. Data. 2022; 7(2): 21. https://doi.org/10.3390/data7020021
28. Gaftandzhieva S, Talukder A, Gohain N, Hussain S, Theodorou P, Salal YK, et al. Exploring online activities to predict the final grade of student. Mathematics. 2022; 10(20): 3758. https://doi.org/10.3390/math10203758
29. Abdullah M, Al-Ayyoub M, Shatnawi F, Rawashdeh S, Abbott R. Predicting students' academic performance using e-learning logs. IAES. Int J Artif Intell. 2023; 12(2): 831. https://doi.org/10.11591/ijai.v12.i2.pp831-839
30. Kareem AK, Al-ani MM, Nafea AA. Detection of Autism Spectrum Disorder Using A 1-Dimensional Convolutional Neural Network. Baghdad Sci J. 2023; 20(3): 1182–93. https://doi.org/10.21123/bsj.2023.8564
31. Nafea AA, Omar N, Al-qfail ZM. Artificial Neural Network and Latent Semantic Analysis for Adverse Drug Reaction Detection. Baghdad Sci J. 2024; 21(1): 226-33 . https://doi.org/10.21123/bsj.2023.7988
32. Sharaff A, Gupta H. Extra-tree classifier with metaheuristics approach for email classification. In: Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018. Springer. 2019;p. 189–97. https://doi.org/10.1007/978-981-13-6861-5_17
33. Chu Z, Yu J, Hamdulla A. Throughput prediction based on extratree for stream processing tasks. Comput Sci Inf Syst. 2021; 18(1): 1–22. https://doi.org/10.2298/CSIS200131031C
34. Bhati BS, Rai CS. Ensemble based approach for intrusion detection using extra tree classifier. In: Intelligent Computing in Engineering: Select Proceedings of RICE 2019. Springer. 2020; p. 213–20. https://doi.org/10.1007/978-981-15-2780-7_25
35. Pinto A, Pereira S, Correia H, Oliveira J, Rasteiro DMLD, Silva CA. Brain tumour segmentation based on extremely randomized forest with high-level features. Annu Int Conf IEEE Eng Med Biol Soc . IEEE. 2015; p. 3037–40. https://doi.org/10.1109/EMBC.2015.7319032
36. Alsumaidaie MSI, Alheeti KMA, Alaloosy AK. An Assessment of Ensemble Voting Approaches, Random Forest, and Decision Tree Techniques in Detecting Distributed Denial of Service (DDoS) Attacks. Iraqi J Electr Electron Eng. 2023; 20(1) : p16-24. https://doi.org/10.37917/ijeee.20.1.2
37. Devetyarov D, Nouretdinov I. Prediction with confidence based on a random forest classifier. Conference Artificial Intelligence Applications and Innovations: 6th IFIP WG 125 Int Conf, AIAI 2010, Larnaca, Cyprus, October 6-7, 2010 Proc 6. Springer; 2010. p. 37–44. https://doi.org/10.1007/978-3-642-16239-8_8
38. Boulesteix A, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2012; 2(6) :493–507. https://doi.org/10.1002/widm.1072
39. Lubis AR, Lubis M. Optimization of distance formula in K-Nearest Neighbor method. Bull Electr Eng Inform. 2020; 9(1): 326–38. https://doi.org/10.11591/eei.v9i1.1464
40. Al-Khowarizmi RS, Nasution MKM, Elveny M. Sensitivity of MAPE using detection rate for big data forecasting crude palm oil on k-nearest neighbor. Int J Electr Comput Eng. 2021; 11(3): 2696–703. https://doi.org/10.11591/ijece.v11i3.pp2696-2703
41. Rafiee M. Self-organization map (SOM) algorithm for DDoS attack detection in distributed software defined network (D-SDN). J Inf Syst Telecommun. 2022; 2(38): 120. https://doi.org/10.52547/jist.15644.10.38.120
42. Alsumaidaie MSI, Alheeti KMA, Al-Aloosy AK. Intelligent Detection System for a Distributed Denial-of-Service (DDoS) Attack Based on Time Series. DeSE. IEEE; 2023. p. 445–50.

Published Online First: May, 2024
https://doi.org/10.21123/bsj.2024.9643
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

https://doi.org/10.52866/ijcsm.2023.02.03.002
43. Giuffrida D, Benetti G, De Martini D, Facchinetti T. Fall detection with supervised machine learning using wearable sensors. IEEE Int Conf Ind Inform. ; 2019. p. 253–9.https://doi.org/10.1109/INDIN41052.2019.8972246
44. AL-Ani MM, Omar N, Nafea AA. A Hybrid Method of Long Short-Term Memory and Auto-Encoder Architectures for Sarcasm Detection. J Comput Sci. 2021; 17(11): 1093–8. https://doi.org/10.3844/jcssp.2021.1093.1098
45. Alsumaidaie MSI, Alheeti KMA, Alaloosy AK. Intelligent Detection of Distributed Denial of Service Attacks: A Supervised Machine Learning and Ensemble Approach. Iraqi J Comput Sci Math. 2023; 4(3): 12–24. DOI: https://doi.org/10.52866/ijcsm.2023.02.03.002
46. Nafea AA, Omar N, AL-Ani MM. Adverse Drug Reaction Detection Using Latent Semantic Analysis. J Comput Sci. 2021; 17(10): 960–70. https://doi.org/10.3844/jcssp.2021.960.970

نظام ذكي للتنبؤ بأداء الطلاب باستخدام التعلم الآلي

مصطفى صفوك إبراهيم الصميدعي[1]، احمد عادل نافع[2]، عبد الرحمن عباس مخلف[3]، رقية ضاري جلال[1]، محمد ماهر العاني[4]

[1]قسم علوم الحاسوب، كلية علوم الحاسوب وتكنولوجيا المعلومات، جامعة الأنبار الرمادي، العراق.
[2]قسم الذكاء الاصطناعي، كلية علوم الحاسوب وتكنولوجيا المعلومات، جامعة الانبار، الرمادي، العراق.
[3]التسجيل وشؤون الطلاب، مقر الجامعة، جامعة الانبار، الرمادي، الانبار، العراق.
[4]مركز تكنولوجيا الذكاء الاصطناعي، كلية علوم وتكنولوجيا المعلومات، جامعة كيبانجسان الماليزية، بانجي، سيلانجور، ماليزيا.

الخلاصة

يعتمد الذكاء الاصطناعي (AI) على الخوارزميات التي تمكن الآلات من اتخاذ قرارات بدلاً من البشر، مما يؤدي إلى تحسين تجارب المستخدم عبر مجالات متنوعة. تناقش هذه الدراسة حلاً ذكيًا للتنبؤ بأداء الطلاب وتحديد الطلاب الذين قد يحتاجون إلى دعم إضافي. يستخدم النظام المقترح خوارزميات التعلم الآلي الخاضع للإشراف: مصنف الغابة العشوائية، ومصنف الأشجار الإضافية، ومصنف. تتضمن منهجية البحث جمع البيانات والمعالجة المسبقة وتحديد الميزات وبناء النموذج والتقييم. يتم استخدام مجموعة بيانات مكونة من 24000 مثيل للتدريب و6000 مثيل للاختبار. يتم تطبيق تقنيات المعالجة المسبقة على مجموعة البيانات، ويتم استخدام خوارزميات تعلم الآلة للكشف عن أداء الطلاب. تقوم النماذج المدربة بتقييم نتائج الطلاب بناءً على استفسارات المستخدم. ويتم تقييم دقة وكفاءة النظام المقترح باستخدام المقاييس المناسبة. تحقق خوارزمية ET أعلى دقة تبلغ 98.15%، تليها خوارزمية RF بنسبة 94.03% وKNN بنسبة 91.65%. تُظهر مقاييس الدقة والاستدعاء قيمًا عالية عبر الخوارزميات الثلاثة. تعرض KNN وقت تدريب أقل بكثير يبلغ 0.00 ثانية، مما يوضح كفاءتها الحسابية. بشكل عام، توفر هذه الورقة رؤى فعالة حول تطبيق تعلم الآلة في التنبؤ بأداء الطلاب. يُظهر النموذج المقترح نتائج واعدة في تحديد الطلاب الذين يحتاجون إلى دعم إضافي، مما يتيح التدخلات المناسبة لتعزيز نتائجهم الأكاديمية. تساهم النتائج في التنقيب عن البيانات التعليمية في العراق ولها آثار على تحسين معدلات نجاح الطلاب في المؤسسات التعليمية.

الكلمات المفتاحية: الذكاء الاصطناعي، التنبؤ بأداء الطالب، استخراج البيانات التعليمية، خوارزمية الأشجار الإضافية، التعلم الآلي الخاضع للإشراف.