Implementing Machine Learning Techniques for Predicting Student Performance in an E-Learning Environment

Adi Suryaputra Paramita [1,*], Laura Mahendratta Tjahjono [2]

[1] Information Systems Department, School of Information Technology Universitas Ciputra Surabaya, Indonesia
[2] Informatics Department, School of Information Technology Universitas Ciputra Surabaya, Indonesia
[1] adi.suryaputra@ciputra.ac.id [*]; [2] laura@ciputra.ac.id;
* Corresponding author

(Received July 3, 2021 Revised August 11, 2021 Accepted August 29, 2021, Available online September 1, 2021)

Abstract

The way people learn has been affected by the COVID-19 pandemic, resulting in a shift from traditional classroom-based learning to online learning. As a result, educational institutions have a new opportunity to use relevant data to predict student performance, which can help improve teaching and learning processes and adjust course curriculum. Universities can leverage machine learning technology to forecast student performance, enabling them to make necessary changes to lecture delivery and curriculum. A study examined open university educational data, using demographic, engagement, and performance metrics to predict student performance with machine learning techniques. The study found that the k-NN strategy outperformed all other algorithms in some cases, while the ANN approach performed better in others.

Keywords: E-Learning, Data Mining, Machine Learning, Student Performance

1. Introduction

The use of Internet technology has revolutionized education and brought about the development of a new online and blended format known as the E-Learning Environment (ELE). This has opened up a new area of study for researchers. Researchers have found that combining onsite learning with the ELE platform can lead to improved comprehension and performance in students [1,2]. The COVID-19 epidemic has further highlighted the importance of ELE, with academic institutions switching to online forms of learning. Despite the benefits of ELE, one of the biggest challenges for educational institutions is accurately assessing student performance. It can be difficult to ensure that students are not cheating and using external sources such as the Internet or printed notes. Therefore, predicting real student performance early on in the course can be beneficial to teachers and course organizers who need to provide support and attention to struggling students [2-4].

Student performance evaluation is a crucial task for educators as it helps them identify areas of improvement and tailor teaching methods to meet the unique needs of each student [5]. Data mining can provide valuable insights into student performance by analyzing large sets of data to identify patterns and trends [6]. By using data mining techniques, educators can not only identify academic strengths and weaknesses but also identify factors that may contribute to student success, such as attendance, engagement, and learning styles [7]. This information can be used to develop targeted interventions and personalize learning experiences, leading to better academic outcomes and higher levels of student engagement.

Over the last decade, Educational Data Mining has emerged as a powerful tool for identifying critical knowledge and patterns in massive educational datasets [3,7]. It involves applying data mining methods to educational datasets. Today, regression, classification, and machine learning algorithms are more accurate and effective at predicting student performance [6]. The accuracy of these predictions is influenced by various factors, including the type, dimension, and variety of the dataset. Machine learning algorithms such as Decision Tree, SVC, k-NN, Random Forest, ANN, and AdaBoost are commonly used for forecasting student performance based on the regression and classification analysis of ELE datasets.

2. Literature Review

In this section, we present the results of evaluating student performance using AI and machine learning. These outcomes are based on various teaching methods, course materials, and accessibility data. We have organized this review in chronological order to demonstrate the progress of this field over time. In a previous study Saheed et al [5] developed a set of linear multi-regression models to predict student performance using educational data. The model included various factors such as past performance, participation in the Learning Management System (LMS), and course-related tasks. This model was tested using a unique approach, which collected 11,556 student entries and 832 courses. The root mean square error of the multi-regression model was 0.147, which was slightly higher than the RMSE of the single regression model [3,6-9]. Overall, these findings suggest that AI and machine learning have the potential to enhance the evaluation of student performance. By considering various factors, such as past performance and course-related tasks, these models can provide more accurate predictions of student outcomes. As this field continues to evolve, we anticipate further advancements in the development of AI-based assessment tools.

Sorour et al. [10] developed a model in 2014 that uses k-NN to predict the success of students in collaborative social learning. They utilized a multivariate classification approach to deal with the challenge of poor categorization. In order to validate their method, they gathered a custom dataset from an online course on Coursera in 2014. The dataset consisted of the total number of user interface clicks and mouse-overs generated by students during the course. The accuracy of categorization for the 2, 3, and 10 score bands were 88 percent, 77 percent, and 31 percent, respectively. In 2012, Bhardwaj and Pal [11] predicted the final exam performance of students using machine learning techniques such as k-NN and support vector machine. They used a dataset comprising 395 data samples from Portugal's University of Minho, which included individual data variables and information about students' families. SVM outperformed k-NN in terms of accuracy.

Karthikeyan et al [12] compared three machine learning strategies, namely collaborative filtering, matrix factorization, and restricted boltzmann machines, in order to validate their performance in predicting student scores. They used a proprietary dataset from pakistan's international technical university consisting of 225 student records. They collected performance-based variables such as prior academic performance and interview score. RBM had the best performance with a root mean square error of 0.3. These studies demonstrate the effectiveness of machine learning techniques in predicting the success of students in collaborative social learning. The use of custom datasets and performance-based variables may improve the accuracy of these predictions. SVM and RBM have shown promising results and could be useful in developing predictive models for other educational contexts.

In 2012, a study conducted by Tair and El-Halees [14] aimed to predict student engagement and its influence on academic performance using various learning-based algorithms. To assess student involvement, the researchers utilized decision tree, classification, regression tree, JRIP decision rules, gradient boosting trees, and naive Bayes classifier techniques on the open university dataset. The analysis focused on demographic, performance, and learning behavior variables from the July 2013 session, which comprised 384 records. Among the methods employed, the J48 decision tree approach demonstrated the highest accuracy of 88.52% and recall of 93.4%, outperforming other techniques. Meanwhile, in a separate study, Burgos et al [15] utilized machine learning algorithms to identify students who were at risk of underperformance. The researchers used the Open University Learning Analytics Dataset (OULAD), which included 32,593 student entries, and employed a combination of activity-based and performance topographies to predict academic success. The machine learning approaches employed by included support vector machine, naive Bayes, random forest, XGBoost, and logistic regression. Of all the algorithms, support vector machine demonstrated the best performance, with an accuracy of 87.98%. Together, these studies demonstrate the potential of machine learning methods to forecast student engagement and performance, and the importance of considering multiple algorithms to identify the most effective approach for a given dataset.

The researchers looked at the OULAD dataset, which contains 32,593 student records and is open-source. The dataset comprised demographic information, clickstream behavior, and assessment results. The proposed deep learning-based technique beat classical regression and SVM algorithms with an accuracy of up to 93 percent[20].

3. Methodology

3.1. Random Forest

The random forest is a powerful technique in machine learning that allows for accurate classification and prediction. It uses an approach known as ensemble learning, which combines multiple classifiers to solve complex problems. A random forest algorithm is composed of many decision trees, and it trains a 'forest' by using bagging or bootstrap aggregation [16,17]. These meta-algorithms combine multiple machine learning techniques to improve accuracy and avoid the limitations of a single decision tree. The random forest algorithm determines the outcome by using the predictions of each decision tree in the forest, which are then averaged or aggregated to provide a more precise forecast [18]. Unlike traditional decision trees, the random forest approach does not require the user to calculate information gain or select root nodes. Instead, it generates multiple decision trees, each of which has a different subset of the dataset and features. The algorithm then calculates the output and vote for each anticipated objective for each of these trees, and the final forecast is based on the one with the most votes. This technique is particularly useful in supervised learning, where a set of rules is used to frame each decision tree, making it ideal for categorization and regression analysis [19].

The benefits of using a random forest approach include improved precision, decreased overfitting, and the ability to generate forecasts without requiring a large number of package shapes. As the number of decision trees in the forest increases, the accuracy of the forecast improves, making it an ideal method for complex problems that require a high level of accuracy. Overall, the random forest approach is a valuable tool in the field of machine learning and can be used in a wide range of applications.

3.2. Naive-Bayes

The Naive Bayes method is a popular and powerful predictive modeling approach, commonly used in data science and machine learning. This method employs a probabilistic model that contains two essential types of probabilities, namely the likelihood of each class and the conditional probability for each class given each x value. These probabilities are calculated from the training data and then used to predict new data using the Bayes theorem. For numerical data, the Naive Bayes method assumes that each attribute follows a Gaussian distribution, making the estimation of probabilities simpler [21-24].

The name "naive Bayes" is derived from the assumption that each input variable is independent of one another, although this may not be entirely true in real-world scenarios. Nonetheless, this technique works remarkably well in a broad range of complex situations [25]. Naive Bayes is a Bayesian graphical model that contains nodes for each column or feature, also known as NB. This model is referred to as "naive" because it disregards prior parameter distributions and assumes that all features and rows are independent. By ignoring the past, this technique provides significant advantages, such as applying any distribution to individual attributes and inferring the most probable values from the data [26].

However, it also has some drawbacks, such as being a maximum likelihood model, meaning that the posterior does not improve iteratively. Despite these limitations, the NB method remains a probabilistic generative model that can generate data given parameters. The nodes generate values that correspond to observable feature values, where numeric attribute values can be a discrete collection of symbols, and categorical attribute values can be a discrete set of symbols. The label column in the Naive Bayes method is used to indicate the location of a node, which can be categorical or real-valued in a classification or regression problem. This method is based on the Bayes theorem, a classification technique that determines the probability of an object possessing certain properties associated with a given class, also known as a probabilistic classifier. In this method, the occurrence of one attribute is unrelated to the occurrence of other characteristics.

3.3. k-Nearest Neighbour

The technique of k-nearest neighbors is a popular approach for categorizing data, which estimates the probability that a data point belongs to one of two categories based on the proximity to the other data points. KNN is a supervised machine learning algorithm, utilized to address classification and regression problems, although it is primarily used to resolve categorization issues. KNN is a non-parametric, slow learning algorithm that does not perform any training during the training phase [26-28].

Due to this, it is referred to as a lazy learning algorithm, which means that it only records the data and does not perform any computations during the training phase. Instead, KNN starts modeling once the dataset is queried. Hence, KNN is a great tool for data mining. K-NN was introduced by Fix et al. in 1951 as a non-parametric machine learning technique that categorizes input based on its immediate neighborhood. Whether this is the best method for data classification is still an open question. KNN generates predictions directly from the training set, making it the simplest and most suitable approach for making predictions.

3.4. Support Vector Machine

A support vector machine is a powerful machine learning model that falls under the category of supervised learning algorithms. It is designed to solve two-group classification problems by using various classification techniques. With the help of labeled training data for each category, SVM models can accurately categorize new text data. One of the significant advantages of SVM over other recent algorithms like neural networks is its speed and ability to work with less data [29]. This makes it a popular method for text classification tasks that require only a few thousand labeled examples. The original concept of SVM was introduced by Vapnik in 1963, which was later extended by Boser et al. in 1992. The basic principle of this algorithm is to generate a large number of hyperplanes in a high-dimensional space with the aim of achieving good separation between them. The high margin between hyperplanes ensures that there is little generalization loss during the classification process [30].

To explain how SVM works, the test sample is first divided into two groups, and each sample is represented as an m-dimensional vector divided into two parts by a (m-1)-dimensional hyperplane. In a linear classification situation, multiple hyperplanes can be used to separate the test samples, but the one with the greatest separation is chosen. In summary, SVM is an effective machine learning model that can be used for text classification tasks with relatively small amounts of labeled data. The algorithm generates hyperplanes in a high-dimensional space to ensure good separation between different categories. Its speed and efficiency make it a popular choice for many real-world applications.

3.5. Artificial Neural Network

The Artificial Neural Network is a revolutionary data processing technique that draws inspiration from the biological nervous system's intricate operations, particularly how the human brain processes information. This innovative strategy is heavily reliant on the structure of the information processing system, which can differ depending on the application [18]. Essentially, a neural network comprises interconnected information processing elements (neurons) that work in harmony to solve complex problems, typically classification or prediction problems. The functioning of a Neural Network closely mimics supervised learning, which is how humans learn by observing examples. First, Neural Networks are programmed to perform specific tasks such as pattern recognition or data classification. They are then fine-tuned over time, much like how humans learn and adapt to new experiences. In biological systems, learning involves modifying existing synaptic connections between neurons. In the case of a Neural Network, weight values are updated in every input, neuron, and output link to facilitate learning [20].

An artificial neural network is a machine learning system that emulates human brain behavior. It is made up of nodes, layers, and connections. Nodes represent artificial neurons that have the ability to understand and transmit input impulses to other neurons. Most ANNs consist of input, output, and hidden layers, each containing artificial neurons linked to one another. These layers of nodes, connections, and weights allow ANNs to perform sophisticated tasks, such as speech recognition, natural language processing, image analysis, and much more.

4. Result and Discussion

In this section, we provide an overview of the data collection and processing methods used in the experimental study. Additionally, this section presents the results of several data mining and machine learning algorithms, including Naive Bayes, Random Forest, k-NN, SVM, and ANN, that were applied to analyze student performance across multiple input feature combinations.

4.1. Dataset

The dataset used in both tests was sourced from Kaggle and pertains to The Open University. The dataset comprises 32,593 student records that represent 15 different countries. It contains a wealth of information, including data on courses that were selected by students, demographics, and details on student interactions with the e-learning platform. Before the data was utilized in the analysis, it was necessary to clean it and extract the desired features. Specifically, data cleaning for classification analysis requires dealing with any missing values and converting textual phrases to numerical values. The dataset is composed of three main input features: demographic (DM), engagement (EG), and performance (PP). The target variable is a binary measure of students' performance, which indicates whether they passed or failed.

4.2. Data Pre-Processing

The final dataset is stored in a file that is separated by commas between each data element. The file contains information related to three different categories: Demographic, Engagement, and Past Performance. Within each category, there are various features that are included in the dataset. These features provide more specific information and help to categorize the data into meaningful groups. In the Demographic category, the dataset includes information about the characteristics of the individuals being studied, such as age, gender, and location. The Engagement category includes information related to how individuals interact with a particular product or service, including things like click-through rates, page views, and time spent on a website. Finally, the Past Performance category includes data related to an individual's historical behavior, such as purchase history or previous engagement with a product or service.

Overall, the comma-separated dataset provides a comprehensive and organized view of the collected data, which can be used for further analysis and to draw insights about the individuals or groups being studied.

Table. 1. Details of Data set features

Demographic features	Values	Description
Gender	1-2	1: Female 2: Male
Highest Education	0 1 2 3 4	0: Below high school 1: High school 2: Diploma 3: Bachelor 4: Post graduate
Age	0 1 2	0: <35 1: 35-55 2: >55
	Engagement	
Total clicks	0-1	1: 0-100 0: N (Null)
	Performance	
Score per assessment	0-1	1: 1 - 100 0: 0
No. of Attempts	0-1	1: 0
Final exam score	0-1	0: >0

The open university education dataset was utilized in a study where algorithms were employed to categorize academic performance based on several factors including demographics, engagement, and past academic performance. The results of the study are presented in Table-2, which displays the performance of various algorithms in different situations. Notably, k-NN algorithm was observed to perform exceptionally well in situations involving engagement (EG), past performance (PP), and a combination of both (DM+EG), with accuracy rates of 0.9979, 0.9907, and 0.9918 respectively. However, in situations that also involved demographics (D), demographics and past performance (DM+PP), engagement and past performance (EG+PP), and all three factors (DM+EG+PP), the ANN algorithm outperformed k-NN, with accuracy rates of 0.7902, 0.9952, 0.9834, and 0.9956 respectively. These results suggest that a combination of algorithms may be necessary to accurately classify academic performance across different scenarios.

Table. 2. Performance comparison of Classification Algorithms to predict final exam result

		DM	EG	PP	DM+EG	DM+PP	EG+PP	DM+EG+ PP
Accuracy	CNN	0.7404	0.9705	0.9366	0.9249	0.9276	0.9774	0.9012
	KNN	0.7021	**0.9979**	**0.9907**	**0.9918**	0.9802	0.9123	0.9689
	ANN	**0.7902**	0.9319	0.9582	0.9904	**0.9952**	**0.9834**	**0.9956**
	SVM	0.7243	0.9238	0.9035	0.9802	0.9165	0.9127	0.9812

	Naive Bayes	0.7865	0.9557	0.9841	0.9744	0.9691	0.9276	0.9234
	Random Forest	0.7189	0.9665	0.9612	0.9153	0.9567	0.9523	0.9155
F1 Score	CNN	0.7065	0.9712	0.9198	0.9978	0.9277	0.9911	0.9322
	KNN	0.7179	0.9067	0.9234	0.9051	0.9743	0.9543	0.9789
	ANN	0.7051	0.9501	0.9543	0.9232	0.9814	0.9648	0.9421
	SVM	0.7512	0.9608	0.9500	0.9361	0.9856	0.9987	0.9155
	Naive Bayes	0.7898	0.9676	0.9056	0.9345	0.9109	0.9678	0.9901
	Random Forest	0.7542	0.9853	0.9965	0.9632	0.9449	0.9234	0.9545
J-Index	CNN	0.7946	0.9705	0.9366	0.9249	0.9153	0.9567	0.9523
	KNN	0.7723	0.9402	0.9099	0.9035	0.9978	0.9277	0.9911
	ANN	0.7321	0.9979	0.9582	0.9904	0.9051	0.9743	0.9543
	SVM	0.6799	0.9238	0.9907	0.9918	0.9232	0.9814	0.9648
	Naive Bayes	0.7821	0.9557	0.9841	0.9744	0.9979	0.9582	0.9056
	Random Forest	0.7698	0.9665	0.9612	0.9153	0.9238	0.9907	0.9965

5. Conclusion

This article utilized a combination of multiple data mining and machine learning techniques to forecast the performance of a particular dataset. To begin with, the data was meticulously cleaned and prepared in a CSV file format, and then different data mining and machine learning algorithms were employed to predict the final exam results of students [18]. The findings of the research were presented in the form of a table, highlighting the most effective algorithms used. The experimental results demonstrated that, when considering various feature variations, the K-NN algorithm was more effective than ANN and SVM, Naive Bayes, and Random Forests. However, there were certain situations where the ANN algorithm proved to be superior to other algorithms [21]. Looking forward, further work will be done to broaden the scope of this research. This will involve measuring additional parameters using a genuine dataset, as well as calculating precision and recall. There will also be a concerted effort to address missing values in the data.

References

[1] C.-C. Kiu, “Data mining analysis on student’s academic performance through exploration of student’s background and social activities,” in 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018, pp. 1–5.

[2] B. K. Francis and S. S. Babu, “Predicting academic performance of students using a hybrid data mining approach,” J. Med. Syst., vol. 43, no. 6, pp. 1–15, 2019.

[3] M. Pandey and V. K. Sharma, "A decision tree algorithm pertaining to the student performance analysis and prediction," Int. J. Comput. Appl., vol. 61, no. 13, pp. 1–5, 2013.

[4] A. D. Kumar and V. Radhika, "A survey on predicting student performance," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 5, pp. 6147–6149, 2014.

[5] Y. K. Saheed, T. O. Oladele, A. O. Akanni, and W. M. Ibrahim, "Student performance prediction based on data mining classification techniques," Niger. J. Technol., vol. 37, no. 4, pp. 1087–1091, 2018.

[6] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006, vol. 1.

[7] S. K. Yadav and S. Pal, "Data mining: A prediction for performance improvement of engineering students using classification," arXiv Prepr. arXiv1203.3832, 2012.

[8] C. Romero and S. Ventura, "Data mining in education," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 1, pp. 12–27, 2013.

[9] K. Kaur and K. Kaur, "Analyzing the effect of difficulty level of a course on students performance prediction using data mining," in 2015 1st International Conference on Next Generation Computing Technologies (NGCT), 2015, pp. 756–761.

[10] S. E. Sorour, T. Mine, K. Godaz, and S. Hirokawax, "Comments data mining for evaluating student's performance," in 2014 IIAI 3rd International Conference on Advanced Applied Informatics, 2014, pp. 25–30.

[11] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," arXiv Prepr. arXiv1201.3418, 2012.

[12] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, "Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation," Soft Comput., vol. 24, no. 24, pp. 18477–18487, 2020.

[13] E. N. Ogor, "Student academic performance monitoring and evaluation using data mining techniques," in Electronics, robotics and automotive mechanics conference (CERMA 2007), 2007, pp. 354–359.

[14] M. M. A. Tair and A. M. El-Halees, "Mining educational data to improve students' performance: a case study," Int. J. Inf., vol. 2, no. 2, 2012.

[15] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," Comput. Electr. Eng., vol. 66, pp. 541–556, 2018.

[16] T. Devasia, T. P. Vinushree, and V. Hegde, "Prediction of students performance using Educational Data Mining," in 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016, pp. 91–95.

[17] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students performance in educational data mining," in 2015 international symposium on educational technology (ISET), 2015, pp. 125–128.

[18] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," Heliyon, vol. 5, no. 2, p. e01250, 2019, doi: https://doi.org/10.1016/j.heliyon.2019.e01250.

[19] A. A. Saa, "Educational data mining & students' performance prediction," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, 2016.

[20] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, 2017.

[21] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," Int. J. Comput. Sci. Manag. Res., vol. 1, no. 4, pp. 686–690, 2012.

[22] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," IEEE Access, vol. 8, pp. 55462–55470, 2020.

[23] V. Ramesh, P. Parkavi, and K. Ramar, “Predicting student performance: a statistical and data mining approach,” Int. J. Comput. Appl., vol. 63, no. 8, 2013.

[24] S. K. Yadav, B. Bharadwaj, and S. Pal, “Data mining applications: A comparative study for predicting student’s performance,” arXiv Prepr. arXiv1202.4815, 2012.

[25] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.

[26] A. Abu Saa, M. Al-Emran, and K. Shaalan, “Factors affecting students’ performance in higher education: a systematic review of predictive data mining techniques,” Technol. Knowl. Learn., vol. 24, pp. 567–598, 2019.

[27] E. Osmanbegovic and M. Suljic, “Data mining approach for predicting student performance,” Econ. Rev. J. Econ. Bus., vol. 10, no. 1, pp. 3–12, 2012.

[28] A. Namoun and A. Alshanqiti, “Predicting student performance using data mining and learning analytics techniques: A systematic literature review,” Appl. Sci., vol. 11, no. 1, p. 237, 2020.

[29] F. Yang and F. W. B. Li, “Study on student performance estimation, student progress analysis, and student potential prediction based on data mining,” Comput. Educ., vol. 123, pp. 97–108, 2018.

[30] A. M. Shahiri and W. Husain, “A review on predicting student’s performance using data mining techniques,” Procedia Comput. Sci., vol. 72, pp. 414–422, 2015.