Journal of
Computer & Electrical and Electronics Engineering Sciences
Original Article
DOI: 10.51271/JCEEES-0027

Comparison of machine learning classification models for predicting student academic performance

Mahmut Ünver
Department of Computer Technologies, Kırıkkale Vocational School, Kırıkkale University, Kırıkkale, Turkiye

Cite this article as: Ünver, M. (2025). Comparison of machine learning classification models for predicting student academic performance. *J Comp Electr Electron Eng Sci, 3*(2), 41-46.

Received: 08.07.2025 • **Accepted**: 27.08.2025 • **Published**: 03.10.2025

ABSTRACT

In this study, various models were developed using machine learning algorithms to predict the academic performance of students in the Programming Course at Kırıkkale University. The data, collected through a survey from 170 associate degree students, consists of 9 input attributes, including demographic information, high school education information, academic status, and socio-economic characteristics, and one output attribute. The data were analysed using the WEKA software. According to the results, the J48 algorithm was identified as the most successful model with an accuracy rate of 96.11%. The Random Forest and k-NN algorithms also closely followed J48 with accuracy rates of 95.83%, demonstrating high predictive performance. The Naive Bayes algorithm demonstrated the lowest classification accuracy with an accuracy rate of 78.06%. In terms of error rates, the Random Forest algorithm showed the best performance with the lowest MAE, RMSE, RAE, and RRSE values. This study demonstrates the applicability of machine learning techniques in education and proves that they can be used as a tool for early detection of at-risk students. In future studies, it is planned to improve the performance of the models and test their generalizability by using larger datasets and different machine learning algorithms.

Keywords: Machine learning, educational data mining, academic performance prediction, WEKA

INTRODUCTION

Predicting student success in the education system offers significant advantages to both students and teachers. Measuring and understanding the effects of numerous variables on student performance is important for enhancing academic success. Today, the biggest problem for students graduating from higher education is the problem of not being able to find a job. In addition, the fact that the education and competencies they receive do not match the personnel competencies in the sector is another important problem. The aim of educational institutions is to ensure that students receive appropriate education, competencies and skills in order to eliminate these problems. In this way, the student can be competitive in the sector and the labor market. It is important to determine in advance whether the student will be successful in this regard. In this process, the student's education can be intervened in and his/her success can be ensured. Thanks to the early determination of performance, the performance of a student who may be unsuccessful can be increased, and he/she can be directed to different skills and competencies thanks to educational coaches. Additional training can be provided. Moreover, this success will not only be the success of the student but also of the educational institution. Analyzing students' academic progress throughout their educational journey provides university administrators with valuable insights into each student's probability of success. In conventional practice, instructors evaluate this by monitoring classroom interactions and recognizing students who exhibit a higher risk of dropout, thereby enabling early intervention.

However, in actualy educational settings, the frequency and quality of interactions between instructors and students are gradually decreasing. This trend is largely due to the increasing number of working students and the growing accessibility of online learning materials. As a result, it has become more challenging to detect at-risk students through conventional methods.

To improve the accuracy of student success predictions, data mining and machine learning techniques offer effective solutions. In today's digital age, educational institutions collect large and diverse datasets. However, much of this data often remains underutilized. Leveraging these valuable resources requires advanced analytical tools, among which machine learning algorithms (MLAs) stand out as particularly powerful.

The goal of this study is to identify the specified variables affecting student success and to choose the most effective machine learning algorithm to predict this success. For this purpose, data obtained from a survey conducted on 170 students studying at Kırıkkale University were used. The survey questions constitute the input attributes. The dataset comprises three main feature categories: individual attributes, educational context, and socio-economic variables. Using the

Corresponding Author: Mahmut Ünver, munver@kku.edu.tr

CfsSubsetEval attribute evaluator and the GreedyStepwise search method in the WEKA software, the 9 most important attributes affecting student success were selected (mathematics course grade, graduated high school type, city of origin, parental education status, preparatory course, income status, gender, residence, and working in an additional job status).

The most fundamental course for computer programming students is the programming course. Success in the programming course is an indicator of the student's academic success. Therefore, the programming course was selected for performance estimation.

In the study, models were developed using the J48 decision tree, Naive Bayes, logistic regression, K-nearest neighbors (k-NN), and Random Forest algorithms, which are known to give successful results in the literature, to predict the programming course grade (output attribute). The performances of the models were compared using TP (true positive) rate, precision, F-measure, accuracy, and error criteria [MAE (mean absolute error), RMSE (root mean square error), RAE (relative absolute error), and RRSE (root relative squared error)]. This study is unique in terms of comparing the performance of different machine learning algorithms in predicting the academic success of Kırıkkale University students and determining the most suitable algorithm.

The following sections of the article include a literature review, materials and methods, the performance of the developed models, comparison of the results, and conclusion sections. A flowchart showing the general steps of the manuscript is shown in **Figure 1**.

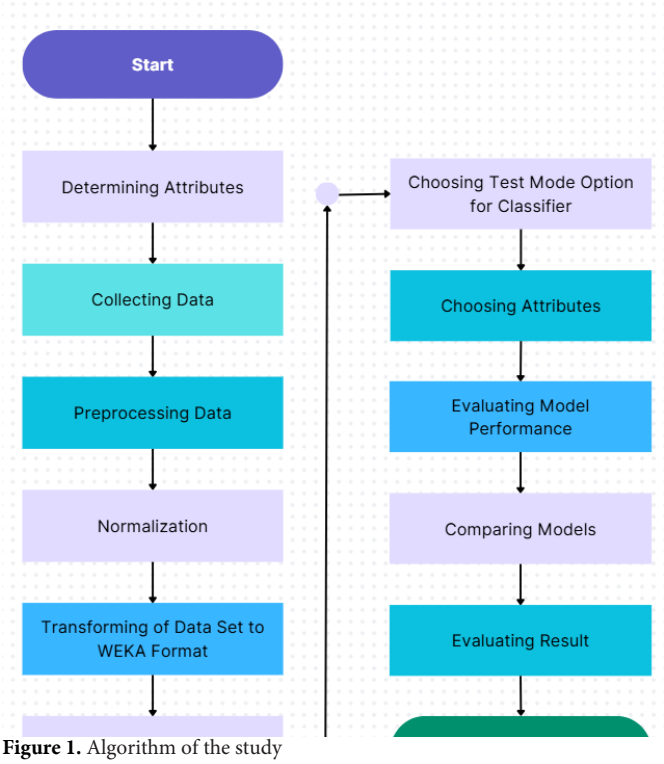

Figure 1. Algorithm of the study

LITERATURE REVIEW

In recent years, machine learning techniques have started to be widely used in the field of education. Algorithms such as support vector machines (SVM), decision trees (DT), k-nearest neighbors (k-NN), and artificial neural networks (ANN) are generally preferred in student performance prediction studies.

Kotsiantis et al. (2004) developed algorithms to predict student success in distance education. In the study, where Naive Bayes, decision trees, and other methods were compared, it was shown that the Naive Bayes method gave better results in many scenarios.

Rastrollo-Guerrero et al. (2020) evaluated machine learning techniques used to analyze and predict student performance. In this study, where supervised learning algorithms were compared, it was stated that SVM in particular gave the best results.

Hasan et al., (2020) stated that the success rate of the models increased when multiple data sources were used.

Aghalarova and Bozkurt Keser (2021) aimed to predict the academic achievements of secondary school students in mathematics and Portuguese courses with the artificial neural network algorithm. In this study, the imbalanced data problem was solved with the SMOTE algorithm, and hyperparameter optimization was performed with the random search method. With the multilayer perceptron (MLP) algorithm, 97.0% accuracy was achieved in mathematics and 92.3% in Portuguese.

In the study conducted by Chen and Zhai (2023), the performance of machine learning methods was investigated. To enable a comparative analysis of various methods, four evaluation metrics along with visual representations are introduced. The experimental findings indicate that the random forest algorithm demonstrates consistently strong generalization across all selected datasets.

Pinto and Paquette (2024) examined the use of deep learning techniques in the field of educational data science. In the literature review on the applications of deep artificial neural networks in education, it was found that deep learning algorithms are effective in areas such as detecting student behaviors and analyzing open-ended student responses.

Xiao et al. (2023) evaluated the applications of large language models in education. The uses of technologies such as deep learning, pre-training, and reinforcement learning in education were examined, and it was stated that large language models have the potential to increase teaching quality and transform educational models.

In the study by Nahar et al. (2021), a dataset containing data from 395 students from a university's undergraduate program was used. Using attributes such as students' demographic information, academic records, and behavioral characteristics, models were developed with classification algorithms such as decision table, Naive Bayes, random forest, OneR, JRip, k-NN, and multilayer perceptron in WEKA software. The performances of the models were compared based on the accuracy rate. According to the results of the study, the random forest algorithm showed the best performance with an accuracy rate of 97.5%.

In the study by Nedeva & Pehlivanova (2021), WEKA software was used. The authors used data they obtained from a survey applied to 115 engineering students. They developed classification models with BayesNet, multilayer perceptron (MLP), sequential minimal optimization (SMO), and J48 algorithms in WEKA software; The models were compared with TP rate, precision, F-measure, accuracy and error criteria (MAE, RMSE, RAE and RRSE). According to the results of

the study, the multilayer perceptron (MLP) algorithm showed the best performance with an accuracy rate of 97.39%.

In exhibited the lowest performance Abu Zohair's study (2019), they used a classifier to predict the performance of postgraduate students at The British University in Dubai in each course. The success of classifier A reached the highest accuracy rate with 79% and 77%.

METHODS

The study was conducted with the approval of the Kırıkkale University Social and Human Sciences Researches Ethics Committee (Date: 17.04.2025, Decision No: 330840). All procedures were carried out in accordance with ethical standards and the principles of the Declaration of Helsinki.

In this study, models that predict academic success were developed using machine learning algorithms in the WEKA environment, using survey data from 170 students studying at Kırıkkale University.

Data Collection and Preparation

In the data collection phase, data obtained from a survey conducted on 170 associate degree students studying at Kırıkkale University were used. The questionnaire consists of questions about students' demographic information, academic backgrounds, socio-economic status, and learning environments. The data were transferred to an Excel file and organized with the attributes in Table 1.

Table 1. Attributes and values for the data set

Type	Attributes	Values
Input	Mathematics_success_grade	AA: 5, BA: 4, BB: 3, CA: 2, CB: 1, Fail: 0
Input	Preparatory_course	Yes: 1, No: 0
Input	Graduated_high_school_type	Other_vocational_program: 1, IT_vocational_program: 2, AL:3, SL:4, FL:5
Input	City_of_origin	outside_Kırıkkale: 0, Kırıkkale: 1
Input	Gender	F: 1, M: 0
Input	Parental_education_status	Graduate: 4, Undergraduate: 3, Associate: 2, High_School:1, Other:0
Input	Income_status	0-20000:1, 20001-30000:2, 30001-50000:3, >50001:4
Input	Residence	Kırıkkale: 1, commuter: 2
Input	Working_additional_job	Yes: 1, No: 0
Output	Programming_algorithm_success_grade	AA: 5, BA: 4, BB: 3, CA: 2, CB: 1, Fail: 0

The values of the attributes were normalized, and the dataset was converted into a WEKA-compatible ARFF file.

Data Preprocessing

At this stage, the collected data were prepared for analysis. Missing data were removed from the dataset. The data preprocessing stage comprises multiple steps, which can be illustrated as follows: data collection, data cleaning, data transformation, and data normalization.

Data collection: Data obtained from the survey results were collected in Microsoft Excel format.

Data cleaning: Surveys with missing data were removed from the dataset.

Data transformation: Data in the Excel file were converted into ARFF (attribute-relation file format) format, which can be used by WEKA, using Python.

```
data = pd.read_excel('anket_verileri.xlsx')

arff_data:
  'description': 'Ogrenci Basari Tahmini',
  'relation': 'ogrenci_basari',
  'attributes':
    ['Mathematics_Success_Grade', ('5', '4', '3', '2', '1', '0')],
    ['Preparatory_Course ', ('1', '0')],
    ['Graduated_High_School_Type ', ('1', '2', '3', '4', '5')],
    ['City_of_Origin', ('0', '1')],
    ['Gender', ('1', '0')],
    ['Parental_Education_Status', ('4', '3', '2', '1', '0')],
    ['Income_Status', ('1', '2', '3', '4')],
    ['Residence', ('1', '2')],
    ['Working_Additional_Job', ('1', '0')],
    ['Programming_Algorithm_Success_Grade', ('5', '4', '3', '2', '1', '0')]
  'data': data.values.tolist()
with open ('ogrenci_basari.arff', 'w') as f:
  arff.dump(arff_data, f)
```

Data normalization: Attributes with numerical values were normalized in the [0,1] range. This process was performed using the "Filters -> Unsupervised -> Attribute -> Normalize" filter via the WEKA interface.

Attribute Selection

To achieve the research objective, attribute selection was performed before classification. The 9 most important attributes were determined using attribute selection algorithms in WEKA. The "CfsSubsetEval" attribute evaluator and the "GreedyStepwise" search method were used for this process. The attribute selection output is shown in Table 2.

Table 2. Attribute selection output

Attributes	Selected folds count	Order of importance
Mathematics_success_grade	10	1
Preparatory_course	0	9
Graduated_high_school_type	2	5
City_of_origin	10	1
Gender	1	7
Parental_education_status	1	7
Income_status	2	5
Residence	10	1
Working_additional_job	4	4

Classification Algorithms

In this study, five different machine learning algorithms, which are known to give good results in student success prediction in the literature (Batool et al., 2023) and are included in the WEKA software, were selected to predict the students' Programming_Algorithm (output attribute). These algorithms are:

J48 (decision trees): A classification algorithm that creates easy-to-interpret and understandable decision trees based on the C4.5 algorithm, which is an improved version of the ID3 algorithm.

Naive Bayes: A probabilistic classification algorithm based on Bayes' Theorem that assumes that the attributes are independent of each other (naive assumption).

Logistic regression: A statistical method that uses the logistic function to model probabilities when the dependent variable is categorical.

K-nearest neighbors (kNN): An instance-based learning algorithm that determines the class of a data point by looking at the class of its nearest "k" neighbors and using majority voting.

Random forest: A powerful and flexible ensemble learning algorithm that works by training multiple decision trees and combining their predictions.

These selected algorithms were applied through the WEKA software using nine input attributes (Mathematics_Success_Grade, Preparatory_Course, Graduated_High_School_Type, City_of_Origin, Gender, Parental_Education_Status, Income_Status, Residence, Working_Additional_Job) and one output attribute (Programming_Algorithm_Success_Grade).

Performance Evaluation Metrics

To evaluate the performance of each classification, the confusion matrix obtained for each classification was used. The classification values for student performance are as follows; AA, BA, BB, CA, CB, and Fail. Therefore, each confusion matrix is created according to 25 classification results. From the confusion matrix, the following parameters can be read for each class; true positive (TP) prediction, true negative (TN) prediction, false positive (FP) prediction, and false negative (FN) prediction.

The parameters used to compare the algorithms are calculated as follows (Sokolova & Lapalme, 2009), (Powers, 2020), (Tharwat, 2021):

True Positive Rate (TP Rate)

It is the ratio of correctly assigned instances to a class.

$$TP\ rate=TP/(TP+FN) \quad (1)$$

Precision (P)

This metric represents the proportion of correctly identified positive cases relative to all instances predicted as positive.

$$P=TP/(TP+FP) \quad (2)$$

F-measure (FM)

It is calculated as the harmonic mean of the Precision (P) and Recall (R) metrics.

$$Fm=2\ PR/(P+R) \quad (3)$$

Here, $R=TP/(TP+FN)$.

Accuracy

This metric indicates the proportion of correctly classified instances out of the total number of cases.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (4)$$

The various error metrics employed in classification methods are presented below.

Mean Absolute Error (MAE)

It is an estimate of how different the predictions are from the values.

$$MAE=(1/n)*\Sigma\ |pi-ai| \quad (5)$$

Here, n is the number of errors, and |pi - ai| are the absolute errors.

Root Mean Square Error (RMSE)

It quantifies the discrepancy between the forecasted values and the empirically observed outcomes.

$$RMSE=\sqrt{[(\Sigma\ (pi-ai)\hat{}2)/n]} \quad (6)$$

pi are the predicted values, ai are values at time/place i (Nedeva & Pehlivanova, 2021).

Relative Absolute Error (RAE)

RAE indicates the ratio of the errors made by the model to the errors that would be made by a simple mean predictor (i.e., a model that predicts the mean value of the output variable for all instances). This metric allows us to understand how much better (or worse) the model's predictions are than a naive mean prediction. A low RAE value means better model performance.

$$RAE=(\Sigma\ |pi-ai|)/(\Sigma\ |ai-\bar{a}|) \quad (7)$$

Where:

- **pi:** The predicted value of the i-th instance by the model
- **ai:** The actual value of the i-th instance
- **ā:** The mean of the actual values of all instances
- **n:** The total number of instances

Root Relative Squared Error (RRSE)

The RRSE is the square root of the ratio of the sum of the squares of the errors made by the model to the sum of the squares of the errors that would be made by a simple mean predictor. Like RAE, RRSE allows us to understand how much better (or worse) the model's predictions are than a naive mean prediction. A low RRSE value means better model performance.

$$RRSE=\sqrt{[(\Sigma\ (pi-ai)^2)/(\Sigma\ (ai-\bar{a})^2)]} \quad (8)$$

Where:

- **pi:** The predicted value of the i-th instance by the model
- **ai:** The actual value of the i-th instance
- **ā:** The mean of the actual values of all instances
- **n:** The total number of instances

For each algorithm, the default parameter settings of WEKA were rearranged.

- For the J48 algorithm, confidenceFactor=0.25 and minNumObj=2 were used.
- For the Naive Bayes algorithm, useKernelEstimator=False and useSupervisedDiscretization=False values were used.
- For the Logistic Regression algorithm, ridge= 1.0E-8 and maxIts=-1 values were used.

- For the k-NN model, K=1, distance weighting=No distance weighting, and nearestNeighbourSearchAlgorithm= euclideanDistance were selected.
- For the random forest algorithm, numIterations=100, maxDepth=0, and numFeatures=0 values were used.

The models used were trained separately with 10-fold Cross-validation and Percentage split (70%) options.

DEVELOPED MODELS AND THEIR PERFORMANCES

While developing the models, after loading the dataset, the classifier was first selected. Then the class attribute was selected, and the parameters of the classifier were adjusted to give the most accurate result. Each of the models was first developed with cross-validation (10 folds) Test options, and then with percentage split (70%). All 5 models gave more accurate results with percentage split (70%). The correctly classified instances values of the models according to the used test option are shown in Table 3.

Table 3. Correctly classified instances of model performances

Algorithm	Cross-validation (10 folds) (%)	Percentage split (70%) (%)
J48	94.8333	96.1111
Naive Bayes	77.3333	78.0556
Logistic regression	80.2500	80.5556
k-NN (k=1)	94.4167	95.8333
Random forest	94.4167	95.8333

The comparison of the weighted average in the detailed accuracy by class and the performance values of the obtained models are shown in Table 4.

According to the results presented in Table 3 and Table 4, the performances of the five different machine learning models developed were compared using TP rate, precision, F-measure, accuracy, MAE, RMSE, RAE, and RRSE metrics. The J48 algorithm showed the highest success with an accuracy rate of 96.1111%, a TP rate of 0.961, a Precision of 0.963, and an F-measure of 0.968. In addition, the random forest algorithm had the lowest error values (MAE: 0.012, RMSE: 0.0786, RAE: 4.3656%, RRSE: 21.1193%).

It turned out that the k-NN and Random Forest algorithms have the lowest TP Rate (0.958). In addition, the Correctly values of these algorithms (95.8333) produced the closest values to the J-48 algorithm.

The Naive Bayes algorithm had the lowest values with a correctly rate of 78.0556, a TP Rate of 0.781, a precision of 0.782, an F-measure of 0.778, a MAE of 0.111, a RMSE of 0.2394, a RAE percentage of 40.3395, and a RRSE percentage of 64.3366.

In summary, according to the results, the J48 (decision trees) algorithm stands out as the most successful model in predicting student success, while the Naive Bayes algorithm gave the lowest performance values.

The comparison of the developed algorithms according to their accuracy rates is shown in Figure 2.

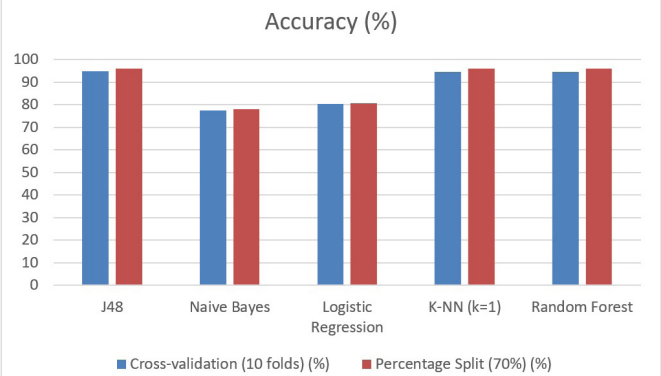

Figure 2. Accuracy rates (%) for algorithms

MAE (mean absolute error) and RMSE (root mean squared error) are error metrics used to evaluate the performance of regression and classification models, and it is desired for these values to be low for an accurate model. The MAE and RMSE values of the developed models are shown in Figure 3. According to the MAE value, the random forest, k-NN, and J48 models produced the lowest values, from smallest to largest. According to the RMSE value, the Random Forest, k-NN, and J48 models produced the lowest values.

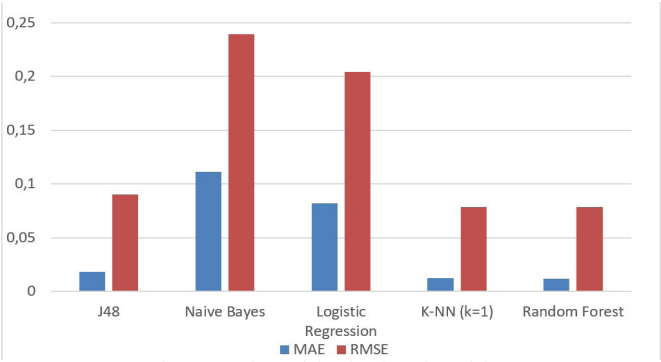

Figure 3. MAE and RMSE values of the developed models

RESULTS

In this study, models were developed using different machine learning algorithms to predict the success of Kırıkkale University students in the programming course. The survey data collected from 170 students consist of 9 input attributes, including demographic information, high school education information, academic status, and socio-economic characteristics, and the output attribute labeled as "programming course grade" (AA, BA, BB, CA, CB, Fail). After the data preprocessing steps, attribute selection was made in the WEKA software, and the 9 attributes with the

Table 4. Error and performance values of models

	TP rate	Precision	F-measure	MAE	RMSE	RAE (%)	RRSE (%)
J48	0.961	0.963	0.968	0.018	0.0901	6.5396	24.2241
Naive Bayes	0.781	0.782	0.778	0.111	0.2394	40.3395	64.3366
Logistic regression	0.806	0.823	0.805	0.082	0.2046	29.7822	54.9894
k-NN (k=1)	0.958	0.964	0.959	0.0121	0.0786	4.3912	21.118
Random forest	0.958	0.964	0.959	0.012	0.0786	4.3656	21.1193

highest impact were determined (mathematics course grade, graduated high school type, city of origin, parental education status, preparatory course, income status, gender, residence, and working in an additional job status).

In the study, five different machine learning algorithms (J48, Naive Bayes, logistic regression, k-NN, and random forest) known to give good results in student success prediction in the literature were used. The models were trained with 10-fold cross-validation and 70% to 30% train-test dataset separation, and their performances were evaluated using accuracy, TP rate, precision, F-measure, MAE, RMSE, RAE, and RRSE metrics.

According to the results obtained, the J48 algorithm was the most successful model with an accuracy rate of 96.11%. This model was followed by the random forest and k-NN algorithms with an accuracy rate of 95.83%. The Naive Bayes algorithm showed the lowest performance with an accuracy rate of 78.06%. When evaluated in terms of error rates, the random forest algorithm produced the lowest MAE (0.012), RMSE (0.0786), RAE (4.3656%), and RRSE (21.1193%) values, making it the model with the lowest error rate. The J48 algorithm also drew attention with its low error rates (MAE: 0.018, RMSE: 0.0901, RAE: 6.5396%, RRSE: 24.2241%).

In this study, it was observed that machine learning algorithms can achieve high accuracy rates in predicting student success. The results of our study show that the J48 and random forest algorithms have high potential in predicting the programming course success of Kırıkkale University students. These models can be used to identify at-risk students, especially those who need to be accurately predicted. In addition, in light of the findings obtained, it can be said that the mathematics course grade, city of origin, and residence attributes have a significant impact on student success.

In future studies, more advanced models can be created by using more comprehensive questionnaires and adding additional attributes such as student absenteeism information and study habits. In addition, the performances of the models can be compared by trying different attribute selection methods and different machine learning algorithms. Finally, the models used in this study were developed to predict the programming course success of Kırıkkale University students. Conducting similar studies in different universities and different courses is important in terms of the generalizability of the results obtained.

Limitations

In this study, it was observed that machine learning algorithms can achieve high accuracy rates in predicting student success. There are also some limitations to this study. First, the dataset is limited to 170 students. With a larger dataset, it is possible to increase the performance of the models and obtain more reliable results. Second, the questionnaire questions used may not cover all the factors that may affect students' academic success.

CONCLUSION

Consequently, this study shows that machine learning algorithms can be used as an effective tool in predicting student success and can help educators in identifying at-risk students and taking necessary measures.

ETHICAL DECLARATIONS

Ethics Committee Approval

The study was carried out with the permission of the Kırıkkale University Social and Human Sciences Researches Ethics Committee (Date: 17.04.2025, Decision No: 330840).

Informed Consent

All students signed and free and informed consent form.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

Abu Zohair, L.M. (2019). Prediction of Student's performance by modelling small dataset size. *Int J Educ Technol High Educ, 16*(1), 1-18. doi:10.1186/s41239-019-0160-3

Aghalarova, S., & Keser, S. B. (2021). Önerilen yapay sinir ağı algoritması ile ortaokul öğrencilerin akademik performansının tahmini. *Veri Bilimi*, *4*(2), 19-32.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: a survey study. *Edu Informat Technol*, *28*(1), 905-971. doi: 10.1007/s10639-022-11152-y

Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. *Edu Informat Technol*, *28*(9), 12039-12057. doi:10.1007/s10639-023-11672-1

Hasan, R., Palaniappan, S., Mahmood, S., Sarker, K. U., & Abbas, A. (2020). Modelling and predicting student's academic performance using classification data mining techniques. *Int J Business Informat Systems*, *34*(3), 403-422. doi:10.1504/IJBIS.2020.108649

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Appl Artific Intell*, *18*(5), 411-426. doi:10.1080/08839510490442058

Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. H. M. S. (2021). Mining educational data to predict students performance. *Edu Informat Technol*, *26*(5), 6051-6067. doi:10.1007/s10639-021-10575-3

Nedeva, V., & Pehlivanova, T. (2021). Students' performance analyses using machine learning algorithms in WEKA. IOP Conference Series: Materials Science and Engineering, *1031*(1), 012061. doi:10.1088/1757-899X/1031/1/012061

Pinto, J. D., & Paquette, L. (2024). Deep Learning for Educational Data Science. In Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines (pp. 111-139). Springer Nature Switzerland. doi:10.1007/978-3-031-64487-0_6

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*. doi:10.48550/arXiv.2010.16061

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: a review. *Appl Sci*, *10*(3), 3. doi:10.3390/app10031042

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Informat Process Manag*, *45*(4), 427-437. doi:10.1016/j.ipm.2009.03.002

Tharwat, A. (2021). Classification assessment methods. *Appl Comput Informat*, *17*(1), 168-192. doi:10.1016/j.aci.2018.08.003

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: a showcase of ChatGPT in education applications. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 610-625). Association for Computational Linguistics. doi:10.18653/v1/2023.bea-1.52