Black Sea Journal of Statistics

Doi: 10.5281/zenodo.18234488

Open Access Journal
e-ISSN: 3108-4265

Research Article
Volume 1 - Issue 1: 13-17 / June 2025

CLASSIFICATION OF STUDENT ACHIEVEMENT USING DATA MINING TECHNIQUES: A COMPARATIVE STUDY

Hatice DİLAVER[1]*, Kâmil Fatih DİLAVER[2]

[1]*Niğde Ömer Halisdemir University, Department of Eurasia Studies, 51200, Niğde, Türkiye*

[2]*Niğde Ömer Halisdemir University, Faculty of Engineering, Department of Electric and Electronics, 51200 Niğde, Türkiye*

Abstract: This study investigates the application of data mining techniques to classify secondary school students' academic performance. The Student Performance Dataset, obtained from the UCI Machine Learning Repository, was used for analysis. After excluding two of the exam results, the dataset comprised 31 attributes for 395 students. The classification was based on final exam grades: scores between 0–10 were labeled as "unsuccessful" (0) and scores between 11 and 20 as "successful" (1). The dataset was preprocessed to correct CSV format errors, making it suitable for analysis in the WEKA software. Four classification algorithms—Iterative Classifier Optimizer, OneR, LogitBoost, and Artificial Neural Networks—were evaluated using 5, 7, and 10-fold cross-validation. Results showed that OneR achieved the highest average accuracy (92.15%) and sensitivity (96%), while LogitBoost yielded the best specificity (88%). The findings suggest that OneR is the most effective method for classifying student success using this dataset.

Keywords: Data mining, Student performance, Classification, Machine learning, Neural networks

***Corresponding author:** Niğde Ömer Halisdemir University, Department of Eurasia Studies, 51200, Niğde, Türkiye
E mail: haticedilaver509@gmail.com (H. DİLAVER)
Hatice DİLAVER https://orcid.org/0000-0002-4484-5297
Kâmil Fatih DİLAVER https://orcid.org/0000-0001-7557-9238

Received: April 23, 2025
Accepted: May 27, 2025
Published: June 15, 2025

Cite as: Dilaver, H., & Dilaver, K. F. (2025). Classification of student achievement using data mining techniques: a comparative study. *Black Sea Journal of Statistics, 1*(1), 13–17.

1. Introduction

The concept of success, in its most basic sense, can be defined as the attainment of desired outcomes as a result of dedicated efforts directed toward specific goals. There are several factors that influence student achievement and performance. The presence of individuals with diverse life backgrounds in a common classroom environment has often led to a neglect of these individual differences. However, students who are treated equally within the same classroom setting may exhibit distinct pathways to acquiring knowledge and learning. Evidence of this lies in the variability of academic success among students receiving the same instruction. Various factors within the classroom environment affect both the academic performance and learning processes of students (Arslan and Babadoğan, 2005).

An increase in research and methodological developments regarding individual differences and student performance offers the potential for a future upward trend in educational attainment levels. A review of the literature highlights key criteria that influence student performance, including the quality of prior education, parental education levels, average family income, the academic program in which the student is enrolled, satisfaction with the school environment, and the student's current psychological state.

The first study attempting to predict student performance was conducted by Gorr et al. (1994). This study compared Linear and Multiple Regression Analysis with Artificial Neural Networks (ANN) for estimating students' grade point averages. The findings indicated that the ANN method produced more accurate results. In another study, SubbaNarasimha et al. (2000) compared Regression methods and ANN by using two separate datasets to predict academic performance of a selected student group. The results suggested that ANN prediction techniques yielded more accurate estimates in that specific context.

Tosun (2007) examined Decision Trees and ANN methods in his study on student performance. While Decision Trees achieved an accuracy rate of 86%, the same dataset analyzed with ANN resulted in 92% accuracy. More recent literature on academic performance prediction includes Aydemir (2019), who developed prediction models using ANN and other classification methods to estimate passing grades in foreign language courses among university students in Türkiye. The data were divided into training and testing sets, and among the tested models, the Bagging method yielded the most accurate predictions, with a mean absolute error of 1.22 and a correlation coefficient of 0.80.

Another contemporary study by Güre et al. (2020)

compared the prediction capabilities of Random Forest and Multilayer Perceptron methods to identify factors affecting mathematical literacy. The analysis of student scores showed that the Random Forest method predicted outcomes with less error, and the variables identified by high-performing models were considered significant factors influencing mathematical literacy.

Altun e al. (2019) conducted a study aimed at predicting final exam scores based on midterm results among elementary education students. The study compared Multiple Linear Regression and ANN methods. The evaluation showed that regression analysis achieved 94.30% accuracy, while ANN achieved 94.43%, indicating comparable success in performance prediction.

In this study, various data mining techniques were applied to classify student achievement in secondary education based on individual and demographic factors. Among the methods tested, the most successful were Iterative Classifier Optimizer, OneR, LogitBoost, and ANN. The analysis used the Student Performance Dataset obtained from the UCI repository. Early studies on this dataset involved clustering algorithms, while later works used different classification algorithms. In this study, the dataset was refined for classification-based data mining analyses, resulting in high accuracy rates. Among all the methods tested, the OneR algorithm delivered the highest prediction performance, outperforming other data mining algorithms.

2. Materials and Methods

The following hyperparameters were used in the artificial neural network model: learning rate = 0.3, batch size = 100, momentum = 0.2. Z-score normalization was applied to the data before processing.

To classify student achievement levels using data mining techniques, the Student Performance Dataset obtained from the UC Irvine Machine Learning Repository (UCI) was utilized. The dataset comprises records of 395 students (187 male and 208 female) and contains 33 attributes (Cortez and Silva, 2008). Among these attributes, there are three exam results pertaining to the mathematics course. According to the information provided on the website from which the dataset was retrieved, the third exam score is considered the most significant. Therefore, only the third exam score was used for classification purposes.

The exam scores, which range between 0 and 20, were categorized into two groups: students scoring between 0 and 10 were classified as “0” (unsuccessful), while those scoring between 11 and 20 were classified as “1” (successful). As a result, two of the exam results were removed from the dataset, reducing the number of attributes from 33 to 31.

The dataset in CSV format initially contained quotation mark (“) errors, which were corrected to make it compatible for analysis using the WEKA software platform. All analyses were performed through WEKA.

Two of the exam scores in the dataset were excluded, reducing the number of attributes from 33 to 31. Based on the final exam scores, the data were binary classified as “0” for unsuccessful (scores between 0–10) and “1” for successful (scores between 11 and 20). The original .csv file contained quotation mark (“) errors, which were corrected to make it suitable for analysis using the WEKA software. All subsequent analyses were conducted via the WEKA application. The data mining methods applied in the study are detailed in the following sections.

The Iterative Classifier Optimizer is commonly used in optimization and classification problems, where gradual model refinement is necessary. It mimics neural network learning and is well-suited for data requiring iterative feedback for improvement.

2.1. Iterative Classifier Optimizer

The Iterative Classifier Optimizer algorithm updates the model through feedback from the errors obtained during the classification of the first record, allowing for iterative refinement. This algorithm functions similarly to a neural network and can be compared to the structure of the human brain. The records are distributed across the network, and once all input samples are presented, the process is repeated—thus demonstrating a core feature of artificial neural networks (ANNs). The network can be configured and trained for a specific application and begins the learning process by randomly selecting initial weights (Manikandan et al., 2018).

The OneR (One Rule) algorithm selects the single best attribute for classification and builds a one-level decision tree. It is particularly effective in cases with one dominant feature.

2.2. OneR

Proposed by Holte (1993), the OneR (One Rule) algorithm is a rule-based classifier that essentially learns a decision tree. It follows a simplistic error-based rule induction logic, aiming to select the best classification criterion by minimizing error rates. Since it focuses on only a single attribute, OneR is often perceived as a shallow approach (Uzun, 2005). The OneR algorithm operates on the following logic:

1. For each attribute, create rules for each of its values.
2. Count the occurrences of each class.
3. Identify the most frequent class for each value.
4. Define classification rules based on the most frequent class.
5. Calculate the error rate for each rule set.
6. Choose the rule set with the lowest error rate.

LogitBoost employs logistic regression in its boosting framework, aiming to enhance accuracy and prevent overfitting. It is especially useful for dense datasets with overlapping classes.

Using the above steps, OneR is applied to the dataset for classification.

2.3. LogitBoost

LogitBoost is one of the boosting algorithms developed to address the overfitting issues often encountered in AdaBoost when applied to dense datasets. LogitBoost reduces the classification errors during training in a

linear manner, thereby improving generalization performance. It employs a logistic loss function and aims to resolve the overfitting problem by increasing the weight of data instances that contribute to classification errors (Aydın and Arslan, 2017).

2.4. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. They consist of a series of interconnected processing units, or neurons, organized into input, hidden, and output layers, as illustrated in Figure 1. Each neuron receives input data from preceding neurons or external sources, applies an activation function to transform the data, and passes the output to the next neuron (Osborn et al., 2011).

ANNs are adaptive, capable of learning from examples, and can function even with incomplete data. They are among the most effective techniques for classification, pattern recognition, signal filtering, data compression, and optimization. ANNs have demonstrated high success rates in various real-world applications including data mining, navigation, fingerprint recognition, material analysis, quality control, and medical diagnostics (Öztemel, 2003).

3. Results and Discussion

In this study, various data mining methods were tested on the selected dataset using the Weka software (URL-1), executed on a computer equipped with an Intel Core i7-4720HQ processor and 12 GB RAM. To ensure objective evaluation, each of the three different data mining methods was applied using three different fold values (5, 7, and 10). The models developed through these techniques were assessed using accuracy, specificity, and sensitivity as evaluation metrics. Detailed results are presented in Table 1.

The ROC (Receiver Operating Characteristic) curves of the applied methods are illustrated in Figure 2.

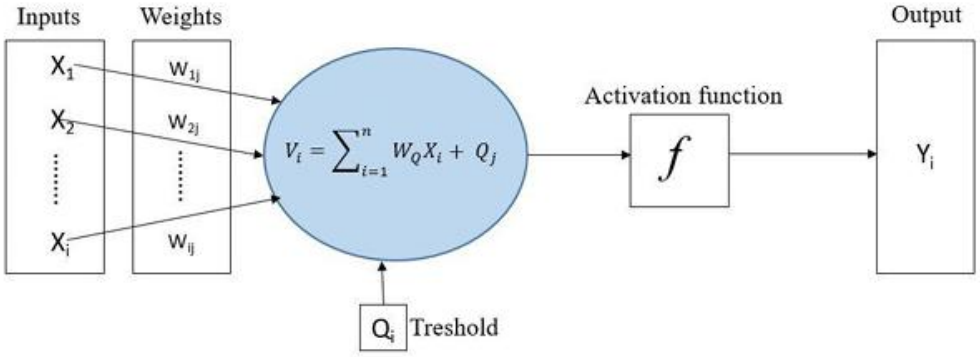

Figure 1. Structure of artificial neural network (Elmas, 2003; Kurnaz and Murat, 2023).

Table 1. Analysis results

Classification Algorithms	Fold	Sensitivity	Specificity	Accuracy
Iterative Classifier Optimizer	5	96%	86%	92.41%
	7	94%	87%	91.90%
	10	94%	86%	91.39%
Average		94.6%	86.3%	91.9%
OneR	5	96%	84%	92.15%
	7	96%	86%	92.15%
	10	96%	84%	92.15%
Average		96%	84.6%	92.15%
LogitBoost	5	94%	88%	92.15%
	7	93%	88%	91.90%
	10	94%	88%	91.89%
Average		93.6%	88%	91.98%
ANN	5	96.4%	76.6%	89.11%
	7	95.6%	69.6%	86.07%
	10	96%	71%	86.8%
Average		96%	72.4%	87.32%

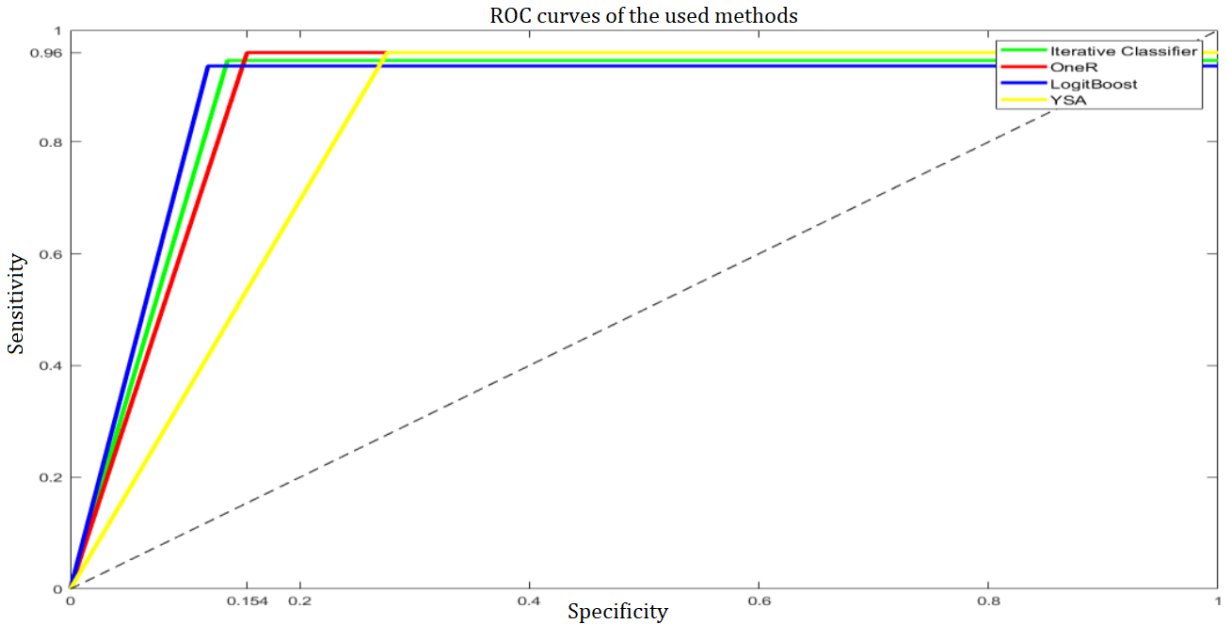

Figure 2. ROC curves of the applied methods.

In the evaluations, classification models were assessed using sensitivity, specificity, and accuracy as performance metrics across 5, 7, and 10-fold cross-validation. Among all the methods tested, the Iterative Classifier Optimizer with 5-fold validation delivered the highest individual accuracy result. However, when considering average accuracy across all folds, the OneR method demonstrated the best overall accuracy performance. Regarding sensitivity, OneR again yielded the highest average. In terms of specificity, LogitBoost outperformed the other methods.

Although the specificity metric for OneR was slightly lower (by approximately 2%) compared to LogitBoost, this difference is considered negligible. Thus, OneR is recommended as the most successful method for this dataset.

4. Conclusion

Students spend a significant portion of their lives within school environments, where they face numerous challenges that can impact their academic performance. These include adaptation difficulties in new environments, academic failures, issues with teachers, peer-related problems, and family-related stressors. Such factors are increasingly acknowledged as key determinants of student success.

In literature, various studies have employed Artificial Neural Networks (ANN) and data mining techniques to analyze and address these challenges. The aim of this study was to identify the main factors negatively influencing the academic performance of secondary school students. The significance of this research lies in its potential to reveal the key elements affecting student achievement in secondary education.

Unlike previous literature, this study incorporated diverse classification algorithms—Iterative Classifier Optimizer, OneR, LogitBoost, and ANN—to classify student performance. Among these, OneR demonstrated the most promising results in predicting academic success.

To improve student achievement, it is essential to first identify and diagnose the key performance-influencing factors, followed by targeted improvements. This approach enables focused educational strategies based on performance-specific indicators. Future research can explore other data mining techniques or focus on identifying the most influential individual factors affecting student success.

Author Contributions

The percentage of the author(s) contributions is presented below. All authors reviewed and approved the final version of the manuscript.

	H.D.	K.F.D.
C	100	
D	100	
S		100
DCP	50	50
DAI	50	50
L	50	50
W	50	50
CR	50	50
SR	50	50
PM	50	50
FA	50	50

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

Altun, M., Kayıkçı, K., & Irmak, S. (2019). Estimation of graduation grades of primary education students by using regression analysis and artificial neural networks. *International Journal of Educational Research, 10*(3), 29–43.

Arslan, B., & Babadoğan, C. (2005). İlköğretim 7. ve 8. sınıf öğrencilerinin öğrenme stillerinin akademik başarı düzeyi, cinsiyet ve yaş ile ilişkisi. *Eurasian Journal of Educational Research, 31*, 35–48.

Aydemir, E. (2019). Ders geçme notlarının veri madenciliği yöntemleriyle tahmin edilmesi. *Avrupa Bilim ve Teknoloji Dergisi, 15*, 70–76.

Aydın, F., & Arslan, Z. (2017). Yapay öğrenme yöntemleri ve dalgacık dönüşümü kullanılarak nörodejeneratif hastalıkların teşhisi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 32*(3), 745–754.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. İçinde *Proceedings of the 5th Annual Future Business Technology Conference* (ss. 5–12).

Elmas, Ç. (2003). *Yapay sinir ağları teori, mimari, eğitim, uygulama* (1. baskı). Seçkin Yayıncılık.

Gorr, W. L., Nagin, D., & Szczypula, J. (1994). Comparative study of artificial neural network and statistical models for predicting student grade point averages. *International Journal of Forecasting, 10*(1), 17–34.

Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). PISA 2015 matematik okuryazarlığını etkileyen faktörlerin eğitsel veri madenciliği ile çözümlenmesi. *Eğitim Bilim, 45*, 251–270.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63–90.

Kurnaz, G., & Murat, N. (2023). Determination of harness production time and defective product formation risk factors with artificial neural network. *Black Sea Journal of Engineering and Science, 6*(4), 325–329. https://doi.org/10.34248/bsengineering.1296187

Manikandan, G., Aravind, V., & Anitha, B. (2018). A survey to identify an efficient classification algorithm for heart disease prediction. *International Journal of Pure and Applied Mathematics, 119*(2), 13337–13345.

Osborn, J., Francisco Javier De C. J., Guzman, D., Butterley, T., Myers, R., Guesalaga, A., & Laine, J. (2011). Using artificial neural networks for open-loop tomography. *Optics Express, 20*(3), 2420–2432. https://doi.org/10.1364/OE.20.002420

Öztemel, E. (2003). *Yapay sinir ağları*. Papatya Yayıncılık.

SubbaNarasimha, P. N., Arinze, B., & Anandarajan, M. (2000). The predictive accuracy of artificial neural network and multiple regression in the case of skewed data: Exploration of some issues. *Expert Systems with Applications, 19*(2), 117–123.

Tosun, S. (2007). *Artificial neural networks, decision tree comparison in classification analysis: An application on students' success* [Yayımlanmamış yüksek lisans tezi]. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Uzun, Y. (2005). *Machine learning algorithms and learning rules with fuzzy logic on medical data* [Yayımlanmamış yüksek lisans tezi]. Selçuk Üniversitesi Fen Bilimleri Enstitüsü.