Research Article

A Study of Prediction Accuracy of English Test Performance Using Data Mining and Analysis

Yujie Duan

Zhengzhou Tourism College, Zhenghou, Henan 451464, China
un79y9@yeah.net

Received: 26th October 2022; Accepted: 15th March 2023; Published: 1st April 2023

Abstract: This paper focused on the effect of data mining in predicting students' English test scores. With the progress of data mining analysis, there are more applications in teaching, and data mining to achieve the prediction of students' test scores is important to support the educational work. In this paper, the C4.5 decision tree algorithm was improved by combining Taylor's series, and then the data of students' English tests in 2019-2020 were collected for experiments. The results showed that the scores of "Comprehensive English" and "Specialized English" had a great influence on the score of CET-4, and the improved C4.5 algorithm was more efficient than the original one, maintained a fast computation speed even when the data volume was large, and had an accuracy of more than 85%. The results demonstrate the accuracy of the improved C4.5 algorithm for predicting students' English test scores. The improved C4.5 algorithm can be extended and used in reality.

Keywords: *College English Test-4; Data mining; Decision tree; English test; Score prediction*

1. Introduction

Data mining refers to the extraction of valuable and regular information from massive data [1], which involves statistical analysis and visualization [2]. At present, it is an extremely popular item and has extensive applications in many fields such as finance, industry, and health care [3]. With the development of information technology and digitalization in education [4], massive data related to student performance has been accumulated in university databases, in which there is a lot of useful information that can provide some guidance for the development of teaching methods [5], such as understanding the factors that affect student performance, predicting students' future performance, and identifying students who may not graduate on time [6]. Data mining can guide educational work by mining student data [7, 8].

2. Related Works

Mutrofin *et al*. [9] analyzed the determination of students' majors based on freshman enrollment data and found that deep learning had the outstanding performance when using the Tanh function by comparing decision trees and deep learning methods. In 2021, Su *et al*. [10] studied students' viewing behavior and flipped classroom performance. The experimental group used Facebook for flipped learning, while the control group used a learning management system. They found that the experimental group had a higher average score than the control group. In 2021, Bendjebar *et al*. [11] conducted an early prediction of learners. They predicted at-risk learners in distance education by analyzing the dynamic characteristics of learners and proved the validity of the prediction method by evaluating it on a real data set. In 2021, Kartikadarma *et al*. [12] analyzed the ratio of primary school students and study groups using artificial intelligence techniques. They found through experiments that the K-means method could provide a reference for governments to formulate policies. In 2022, Dhanalakshmi *et al*. [13] applied Apriori rule mining to forecast

Yujie Duan, "A Study of Prediction Accuracy of English Test Performance Using Data Mining and Analysis", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 1-8, Vol. 7, No. 1, 1st April 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.02.001, Available: http://aetic.theiaer.org/archive/v7/v7n2/p1.html.

special children with mental retardation, autism, and cerebral palsy to help teachers provide vocational training for their students. In 2022, Abdelkader *et al.* [14] assessed the satisfaction of students in the online learning process and compared the quality of 11 feature selection algorithms using K-nearest neighbors and support vector machines as classifiers to find the optimal number of dimensions of feature subsets and the best method. In 2022, Umer *et al.* [15] analyzed data from a learning management system using machine learning methods to forecast student failure in the course. Their findings confirmed the importance of demographics in academic prediction. In 2022, Rutherford *et al.* [16] analyzed changes in student motivation to learn during COVID-19 and found through data analysis that students lowered their math expectations and the affective cost of math. College English Test-4 (CET-4) is an examination of students' comprehensive English proficiency, and the pass rate of CET-4 can reflect the level of English teaching in schools to a certain extent, which is an examination that universities attach great importance. This paper predicted students' CET-4 scores with the data mining method and verified its effectiveness by analyzing it on the actual data set. This work can guide English teaching in schools and improve the CET-4 passing rate.

3. Methodology

3.1. Data Mining and Decision Trees

As society evolves, people are generating more and more data. To process these data effectively, data mining techniques have been developed. Description refers to understanding the hidden laws of the data through data mining. Prediction refers to inferring the rules from the current data and predicting the unknown new data. According to the functions, data mining can be divided into association analysis, classification, clustering, etc. [17]. Classification and prediction are two common functions. Classification refers to finding models that can discriminate data; then, the models are used to predict unknown data. Classification and prediction have been widely used in industries such as finance and healthcare [18]. Predicting students' English test scores can also be considered as a classification problem. Based on the historical data of students, they can be divided into two categories: those who can pass the CET-4 and those who cannot, and then, future passes can be predicted based on the obtained rules.

Bayesian and decision trees are the commonly used methods in classification problems [19]. Decision trees have been widely used in solving classification problems because of their simple operation and intuitive description of the results [20]. Decision trees generate a tree diagram by analyzing the data to obtain readable rules, and then these rules are used to predict new data. Their basic principle is described below.

Data set D contains k attributes, $\{A_1, A_2, \cdots, A_k\}$. D is used as the root node. If samples in D belong to the same class, D is used as the unique leaf node; otherwise, node splitting is required. If A_j is the splitting attribute, A_j is used as the root node, and D is divided into m subsets, corresponding to m branches. This is how a decision tree is constructed. The final IF-THEN rule is the path from the root node to the leaf nodes, and every leaf node represents a rule conclusion.

ID3 and C4.5 are commonly used decision tree algorithms [21]. ID3 is a classical algorithm, which is based on the principle of calculating the information gain of each attribute and selecting the largest attribute as the branch node to classify the samples. Suppose there are s data in dataset S, there are m classes, $\mathrm{C_i}(\mathrm{i} = 1,2,\cdots \mathrm{m})$, and $\mathrm{S_j}$ samples exist in $\mathrm{C_i}$. The expected information of S is written as:

$$I(S_1, S_2, \cdots, S_m) = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{1}$$

$$p_i = \frac{\mathrm{S_i}}{\mathrm{s}}, \tag{2}$$

where $\mathrm{p_i}$ is the probability that any sample belongs to $\mathrm{C_i}$.

It is assumed that attribute $A = \{a_1, a_2, \cdots, a_v\}$ splits S into $\{S_1, S_2, \cdots, S_v\}$ and the number of S_i belonging to class $\mathrm{C_i}$ is $\mathrm{S_{ij}}$. The entropy of the subset that attribute A splits is:

$$E(A, S) = \sum_{j=1}^{v} \frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{s} I\left(S_{1j} + S_{2j} + \cdots + S_{mj}\right). \tag{3}$$

The information gain is:

$$Gain(A) = I(S_1, S_2, \cdots, S_m) - E(A). \tag{4}$$

The ID3 algorithm classifies based on the information gain, which can lead to overfitting of the results. To improve this drawback, the C4.5 algorithm was developed. The C4.5 algorithm is an optimization and improvement of the ID3 algorithm. It classifies based on the information gain ratio and avoids the imbalance of the tree by different trimming techniques. The information gain ratio is calculated by:

$$SplitInfo_A(S) = -\sum_{j=1}^{m}\frac{|S_j|}{|S|}\log_2\left(\frac{|S_j|}{|S|}\right), \tag{5}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(S)}, \tag{6}$$

where $\mathrm{SplitInfo_A(S)}$ is the split information of attribute A and GainRatio(A) is the gain ratio of attribute A.

3.2. Improved C4.5 Algorithm

There are many logarithmic operations in the C4.5 algorithm. To simplify it, this paper incorporates Taylor's series. The definition of Taylor's series is as follows [23].

$$\mathrm{f(x)} = \sum_{\mathrm{n}=0}^{\infty}\frac{\mathrm{f^{(n)}(a)}}{\mathrm{n!}}(\mathrm{x}-\mathrm{a})^{\mathrm{n}}, \tag{7}$$

where $n!$ is the factorial of n and $f^{(n)}(a)$ is the n-order derivative of f at point a. When $f(x)$ is a natural logarithm, i.e., $f(x) = \ln(1+x)$, its Taylor's series is:

$$\ln(1+x) = \sum_{n=0}^{\infty}\frac{(-1)^{n+1}}{n!}x^n. \tag{8}$$

When x is infinitely small, the above equation can be further simplified as:

$$\ln(1+x) = \sum_{n=0}^{\infty}\frac{(-1)^{n+1}}{n!}x^n \approx x. \tag{9}$$

Therefore, the C4.5 algorithm is simplified according to Taylor's series, the information gain ratio after improvement is calculated by:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(S)} = \frac{\sum_{i=1}^{n}\frac{|sc_i|\times(|S|-|sc_i|)}{|S|}-\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{|sc_{ij}|\times(|S|-|sc_{ij}|)}{|S_j|}}{\sum_{j=1}^{m}\frac{|S_j|\times(|S|-|S_j|)}{|S|}}. \tag{10}$$

Also, the number of attribute values for every attribute in the data set, i.e., M, is added to the formula to compensate for the simplification error. The final calculation formula is:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(S)} \times M = \frac{\sum_{i=1}^{n}\frac{|sc_i|\times(|S|-|sc_i|)}{|S|}-\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{|sc_{ij}|\times(|S|-|sc_{ij}|)}{|S_j|}}{\sum_{j=1}^{m}\frac{|S_j|\times(|S|-|S_j|)}{|S|}} \times M\cdot \tag{11}$$

A large number of logarithmic operations in the original C4.5 algorithm significantly increases the operation time, but after the improvement, only the basic four arithmetic operations are left in the formula, reducing the time of function calls and thus improving efficiency.

4. Results

4.1. System Requirements

The experiment was conducted on the Weka platform with an experimental environment of Windows 10. Eclipse development tools and the Java programming language were used.

4.2. Dataset

Students who were enrolled in Zhengzhou Tourism College in 2019 were analyzed, and the data used were the English course grades in the academic years 2019 and 2020. First, 500 data were extracted for the experiment, and the raw data are presented in Table 1.

The raw data was processed according to Table 1. Taking spoken English as an example, it included spoken English 1 and spoken English 2, so the average value was calculated, and the same was done for comprehensive English and English viewing, listening, and speaking. Then, for the course scores, [85,100] was evaluated as excellent and recorded as 1, [75,85) was evaluated as good and recorded as 2, [66,75) was evaluated as fair and recorded as 3, [60,66) was evaluated as passed and recorded as 4, and below 60 was evaluated as failed and recorded as 5. A CET-4 score above 425 points was considered qualified; therefore, the CET-4 score was expressed as $\begin{cases} score \geq 425, pass \\ score < 425, fail \end{cases}$. In the data pre-processing, "pass" was denoted as 1, and "fail" was denoted as 0. The processed experimental data are shown in Table 2.

Table 1. The scores of students

Sample number	**Spoken English 1**	**Comprehensive English 1**	**English viewing, listening, and speaking 1**	**Study of Britain and America**	**English viewing, listening, and speaking 2**	**Comprehensive English 2**	**Specialized English 1**	**Spoken English 2**	**Score of CET-4**
1	87.0	82.0	88.0	81.0	92.0	85.0	77.0	70.0	448
2	92.0	93.0	99.0	83.0	90.0	92.0	93.0	87.0	441
3	89.0	83.0	94.0	86.0	97.0	84.0	92.0	85.0	429
4	89.0	86.0	94.0	94.0	96.0	88.0	93.0	87.0	452
5	85.0	83.0	90.0	90.0	94.0	84.0	93.0	84.0	426
6	92.0	78.0	87.0	82.0	96.0	89.0	91.0	86.0	425
7	89.0	87.0	95.0	92.0	99.0	91.0	93.0	84.0	459
8	94.0	83.0	91.0	84.0	91.0	90.0	82.0	84.0	452
9	88.0	80.0	91.0	60.0	74.0	76.0	69.0	64.0	471
10	86.0	76.0	92.0	85.0	97.0	91.0	92.0	83.0	439
11	86.0	79.0	84.0	84.0	91.0	83.0	91.0	80.0	473
12	88.0	81.0	92.0	95.0	93.0	82.0	93.0	80.0	460
13	84.0	78.0	84.0	94.0	94.0	89.0	90.0	81.0	446
14	88.0	93.0	93.0	84.0	95.0	92.0	92.0	65.0	435
15	78.0	74.0	75.0	80.0	82.0	73.0	72.0	66.0	416
16	70.0	75.0	72.0	76.0	80.0	71.0	68.0	65.0	423
……	……	……	……	……	……	……	……	……	……
500	86.0	85.0	93.0	92.0	94.0	76.0	90.0	78.0	433

Table 2. Experimental data after pre-processing

Sample number	**Spoken English**	**Comprehensive English**	**English Viewing, Listening, and Speaking**	**Study of Britain and America**	**Specialized English**	**Score of CET-4**
1	2	2	1	2	2	1
2	1	1	1	2	1	1
3	1	2	1	1	1	1
4	1	1	1	1	1	1
5	2	2	1	1	1	1
6	1	2	1	2	1	1
7	1	1	1	1	1	1
8	1	1	1	2	2	1
9	2	2	2	4	3	1
10	2	1	1	1	1	1
11	2	2	1	2	1	1
12	2	2	1	1	1	1
13	2	2	1	1	1	1
14	2	1	1	2	1	1
15	3	3	2	2	3	0
16	3	3	2	2	3	0
……	……	……	……	……	……	……
500	2	2	1	1	1	1

4.3. Results

According to the formula, the information gain ratio of different attributes was calculated, and the results are displayed in Table 3.

Table 3. Information gain ratio of different attributes

Attribute	Information gain ratio
Spoken English	0.147
Comprehensive English	0.316
English Viewing, Listening, and Speaking	0.175
Study of Britain and America	0.121
Specialized English	0.279

Table 3 shows that "Comprehensive English" had the highest information gain ratio, 0.316. Therefore, it was served as the root node to construct a decision tree and then extract the classification rules. The rules with "pass" results are as follows.

(1) IF Comprehensive English = "excellent" AND Specialized English = "excellent" THEN CET − 4 score = "pass"

(2) IF Comprehensive English = "excellent" AND English Viewing, Listening, and Speaking = "excellent" THEN CET − 4 score = "pass"

(3) *IF Comprehensive English = "excellent" AND Specialized English = "good" AND Spoken English = "excellent" THEN CET* − 4 *score = "pass"*

(4) *IF Comprehensive English = "excellent" AND Specialized English = "good" AND English Viewing, Listening, and Speaking = "good" THEN CET* − 4 *score = "pass"*

(5) *IF Comprehensive English = "good" AND Specialized English = "good" AND Spoken English = "good" THEN CET* − 4 *score = "pass"*

(6) *IF Comprehensive English = "good" AND Specialized English = "excellent" AND Study of Britain and America = "excellent" THEN CET* − 4 *score = "pass"*

(7) *IF Comprehensive English = "good" AND specialized English = "excellent" AND English ViewingListening, and Speaking = "excellent" THEN CET* − 4 *score = "pass"*

(8) *IF Comprehensive English = "good" AND Spoken English = "good" AND English Viewing, Listening, and Speaking = "good" THEN CET* − 4 *score = "pass"*

(9) *IF Comprehensive English "good" AND English Viewing, Listening, and Speaking = "excellent" AND Study of Britain and America = "excellent" THEN CET* − 4 *score = "pass"*

The rules with "fail" results are as follows.

(1) *IF comprehensive English = "average" AND Specialized English = "average" AND Study of Britain and America = "good" THEN CET* − 4 *score = "fail"*

(2) *IF comprehensive English = "average" AND Specialized English = "average" AND Spoken English = "average" THEN CET* − 4 *score = "fail"*

(3) *IF Comprehensive English = "average" AND Specialized English = "average" AND Spoken English = "good" THEN CET* − 4 *score = "fail"*

(4) *IF Comprehensive English = "average" AND Specialized English = "qualified" AND Study of Britain and America = "average" THEN CET* − 4 *score = "fail"*

(5) *IF Comprehensive English = "average" AND Specialized English = "average" AND English Viewing, Listening, and Speaking = "qualified" THEN CET* − 4 *score = "fail"*

(6) *IF Comprehensive English = "average" AND spoken English = "average" AND English Viewing, Listening, and Speaking = "average" THEN CET* − 4 *score = "fail"*

(7) *IF Comprehensive English "average" AND English Viewing, Listening, and Speaking = "good" AND Study of Britain and America = "quatified" THEN CET* − 4 *score = "fail"*

Whether students could pass CET-4 was predicted using the extracted rules. Experiments were conducted on 500, 1000, 1500, 2000, and 2500 pieces of data to compare the original C4.5 algorithm with the optimized one. The results are shown in Table 4.

It was seen from Table 4 that the optimized C4.5 algorithm had a clear advantage in operation time. When the data volume for prediction was 500, the operation time of the original algorithm was 0.023 s, and the operation time of the optimized algorithm was 0.019 s, which was 17.39% less than the original one. As the data volume increased, the operation time of both algorithms showed a slow increase. When the data volume was 2500, the operation time of the original C4.5 algorithm was 0.083 s, which was 0.06 s more than that when the data volume was 500, and the operation time of the optimized algorithm was 0.028 s, which was 66.27% less than that of the original C4.5 algorithm and only 0.009 s more than that when the data volume was 500. These results confirmed the reliability of the optimized algorithm in enhancing computational efficiency.

Table 4. Algorithm performance comparison

Data volume/piece	**Operation time/s**		**Prediction accuracy/%**	
	The C4.5 algorithm	**The optimized C4.5 algorithm**	**C4.5**	**The optimized C4.5 algorithm**
500	0.023	0.019	89.54	89.87
1000	0.034	0.021	88.33	88.64
1500	0.045	0.023	87.92	88.03
2000	0.077	0.025	87.54	87.65
2500	0.083	0.028	86.87	87.12

With the increase of the amount of data predicted, the prediction accuracy of both algorithms showed a decrease, but the decrease was not significant. When the data volume was 500, the accuracy of the original algorithm was 89.54%, and the accuracy of the optimized algorithm was 89.87%, which was improved by only 0.33%. These results demonstrated that improving the C4.5 algorithm did not significantly affect the prediction accuracy but improved the computational efficiency.

5. Conclusion

This paper focused on the prediction of English test scores. The score of CET-4 was predicted using data mining methods. An improved C4.5 method was designed. The mining of English course scores in the first academic year found that “Comprehensive English” and “Specialized English” courses had a large influence on whether students could pass the CET-4 or not, while “Spoken English” and “Study of Britain and America” had a small influence. CET-4 does not include the oral test, so the level of spoken English had a small influence on the performance; since the “Study of Britain and America” course involved less professional knowledge, its influence on CET-4 was also small. The results suggested that students should focus on the study of “Comprehensive English” and “Specialized English” courses to improve their English level and pass CET-4. The results of the algorithm performance analysis indicated that the improved C4.5 algorithm accurately predicted students’ CET-4 scores, always achieving an accuracy of more than 85%. Compared with the original algorithm, the optimized C4.5 algorithm significantly reduced the computation time and improved the computational efficiency, which can be applied to actual teaching work.

References

[1] Mustafa Abdalrassual Jassim, “Analysis of the Performance of the Main Algorithms for Educational Data Mining: A Review”, *IOP Conference Series: Materials Science and Engineering*, Print ISSN: 1757-8981, Online ISSN: 1757-899X, pp. 1-10, Vol. 1090, No. 1, March 2021, Published by IOP Publishing, DOI: 10.1088/1757-899X/1090/1/012084, Available: https://iopscience.iop.org/article/10.1088/1757-899X/1090/1/012084.

[2] M. Besher Massri, Joao Pita Costa, Marko Grobelnik, Janez Brank, Luka Stopar *et al.*, “A Global COVID-19 Observatory, Monitoring the Pandemics Through Text Mining and Visualization”, *Informatica: An International Journal of Computing and Informatics*, Print ISSN: 0350-5596, Online ISSN: 1854-3871, pp. 49-55, Vol. 46, No. 1, March 2022, Published by the Slovenian Society Informatika, DOI: 10.31449/inf.v46i1.3375, Available: https://www.informatica.si/index.php/informatica/article/view/3375/1741.

[3] David Perez-Guaita, Guillermo Quintas, Zeineb Farhane, Roma Tauler and Hugh J. Byrne, “Corrigendum to "Data mining Raman microspectroscopic responses of cells to drugs in vitro using multivariate curve resolution-alternating least squares" [Talanta 208 (2020) 120386]“, *Talanta: The International Journal of Pure and Applied Analytical Chemistry*, ISSN: 0039-9140, pp. 1, Vol. 236, September 2022, DOI: 10.1016/j.talanta.2021.122682, Available: https://www.sciencedirect.com/science/article/pii/S0039914021006032?via%3Dihub.

[4] A Andreasyan and A Balyakin, “Transformation of education through Big Data: digital twins case study“, *Journal of Physics: Conference Series*, Print ISSN: 1742-6588, Online ISSN: 1742-6596, pp. 1-6, Vol. 2210, No. 1, March 2022, Published by IOP Publishing Ltd, DOI: 10.1088/1742-6596/2210/1/012003, Available: https://iopscience.iop.org/article/10.1088/1742-6596/2210/1/012003/pdf.

[5] Raya Mohammed Mahmood and Sefer Kurnaz, “Employing Data Mining to Predict Professional Identity”, *Journal of Information Science and Engineering*, ISSN: 1016-2364, pp. 193-203, Vol. 36, No. 2, 2020, Published by Institute of

Information Science, Academia Sinica, Taiwan, DOI: 10.6688/JISE.202003_36(2).0001, Available: https://www.airitilibrary.com/Publication/alDetailedMesh?DocID=10162364-202003-202003050003-202003050003-193-203.

[6] D K Arun, V Namratha, B V Ramyashree, Yashita P Jain and Antara Roy Choudhury, "Student Academic Performance Prediction using Educational Data Mining", In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 27-29 January 2021, Coimbatore, India, Print on Demand(PoD) ISBN:978-1-7281-9299-4, Electronic ISBN:978-1-7281-5875-4, Published by IEEE, DOI: 10.1109/ICCCI50826.2021.9457021, Available: https://ieeexplore.ieee.org/document/9457021.

[7] Saba Batool, Junaid Rashid, Muhammad Wasif Nisar, Jungeun Kim, Hyuk-Yoon Kwon *et al*. "Educational data mining to predict students' academic performance: A survey study", *Education and Information Technologies*, Print ISSN: 1360-2357, Online ISSN: 1573-7608, pp. 1-67, 9th July 2022, Published by Springer Nature, DOI: 10.1007/s10639-022-11152-y, Available: https://link.springer.com/article/10.1007/s10639-022-11152-y.

[8] Ariana Yunita, Harry B. Santoso and Zainal Arifin Hasibuan, "Research Review on Big Data Usage for Learning Analytics and Educational Data Mining: A Way Forward to Develop an Intelligent Automation System", *Journal of Physics: Conference Series*, Print ISSN: 1742-6588, Online ISSN: 1742-6596, pp. 012044, Vol. 1898, No. 1, June 2021, Published by IOP Publishing, DOI: 10.1088/1742-6596/1898/1/012044, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1898/1/012044.

[9] Siti Mutrofin, M. Mughniy Machfud, Diema Hernyka Satyareni, R.V. Hari Ginardi and Chastine Fatichah, "Komparasi Kinerja Algoritma C4.5, Gradient Boosting Trees, Random Forests, dan Deep Learning pada Kasus Educational Data Mining", *Jurnal Teknologi Informasi dan Ilmu Komputer*, Print ISSN: 2355-7699, pp. 807, Vol. 7, No. 4, August 2020, Published by Fakultas Ilmu Komputer Universitas Brawijaya, DOI: 10.25126/jtiik.2020742665, Available: https://jtiik.ub.ac.id/index.php/jtiik/article/view/2665.

[10] Yu-Sheng Su and Chin-Feng Lai, "Applying Educational Data Mining to Explore Viewing Behaviors and Performance With Flipped Classrooms on the Social Media Platform Facebook", *Frontiers in Psychology*, Online ISSN: 1664-1078, pp. 1-8, Vol. 12, 29th April 2021, Published by Frontiers, DOI: 10.3389/fpsyg.2021.653018, Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.653018/full.

[11] Safia Bendjebar, Yacine Lafifi and Hassina Seridi-Bouchelaghem, "An improvement of a data mining technique for early detection of at-risk learners in distance learning environments", *International Journal of Knowledge and Learning*, Print ISSN: 1741-1009, Online ISSN: 1741-1017, pp. 185-202, Vol. 15, No. 2, 3rd February 2022, DOI: 10.1504/IJKL.2021.10042404, Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJKL.2022.121958.

[12] Etika Kartikadarma, Sri Jumini, Nurulisma Ismail, Barany Fachri, Dadang Sudrajat *et al.*, "Educational Data Mining to Improve Decision Support on the Ratio of students and Study Groups in Elementary Schools in Indonesia using K-Means Method", *İlköğretim Online*, Print ISSN: 1305-3515, pp. 691-698, Vol. 20, No. 1, January 2021, Published by Ilkogretim Online, DOI: 10.17051/ilkonline.2021.01.59, Available: http://ilkogretim-online.org/fulltext/218-1609990185.pdf?1613721048.

[13] R Dhanalakshmi, B Muthukumar and RA Canessane, "Analysis of Special Children Education Using Data Mining Approach", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Print ISSN: 0218-4885, Online ISSN: 1793-6411, pp. 125-140, Vol. 30, No. Supp01, April 2022, Published by World Scientific Publishing, DOI: 10.1142/S0218488522400074, Available: https://www.worldscientific.com/doi/10.1142/S0218488522400074.

[14] Hanan Elrefaey, Ahmed G. Gad, Amr A. Abohany and Shaymaa Sorour, "An Efficient Data Mining Technique for Assessing Satisfaction Level With Online Learning for Higher Education Students During the COVID-19", *IEEE Access*, ISSN: 2169-3536, Vol. 10, pp. 6286-6303, January 2022, Published by IEEE, DOI: 10.1109/ACCESS.2022.3143035, Available: https://ieeexplore.ieee.org/document/9681058.

[15] Rahila Umer, Sohrab Khan, Jun Ren, Shumaila Umer and Ayesha Shaukat, "Prediction of students' failure using VLE and demographic data: case study on Open University data", *International Journal of Business Intelligence and Data Mining*, Print ISSN: 1743-8187, Online ISSN: 1743-8195, pp. 235-249, Vol. 20, No. 2, January 2022, DOI: 10.1504/IJBIDM.2022.120829, Available: https://www.inderscience.com/info/inarticle.php?artid=120829.

[16] Teomara Rutherford, Kerry Duck, Joshua M. Rosenberg and Raymond Patt, "Leveraging mathematics software data to understand student learning and motivation during the COVID-19 pandemic", *Journal of Research on Technology in Education*, Print ISSN: 1539-1523, Online ISSN: 1945-0818, pp. S94-S131, Vol. 54, No. S1, June 2021, Published by Taylor & Francis, DOI: 10.1080/15391523.2021.1920520, Available: https://www.tandfonline.com/doi/full/10.1080/15391523.2021.1920520.

[17] Changkun Liu, Xinrong Wu, Changhua Yao, Jibin Guo, Haoren Fan *et al.*, "Research on Discovery of Radio Communication Relationship Based on Correlation Analysis", *IOP Conference Series: Earth and Environmental Science*, Print ISSN: 1755-1307, Online ISSN: 1755-1315, pp. 1-9, Vol. 440, No. 4, March 2020, Published by IOP Publishing, DOI: 10.1088/1755-1315/440/4/042006, Available: https://iopscience.iop.org/article/10.1088/1755-1315/440/4/042006.

[18] Rasool Azeem Musa, Mehdi Ebady Manaa and Ghassan Abdul-Majeed, "Predicting Autism Spectrum Disorder (ASD) for Toddlers and Children Using Data Mining Techniques", *Journal of Physics: Conference Series*, Print ISSN: 1742-6588, Online ISSN: 1742-6596, pp. 1-8, Vol. 1804, No. 1, 2021, Published by IOP Publishing, DOI: 10.1088/1742-6596/1804/1/012089, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1804/1/012089.

[19] Dam Sao Mai, Phan Hong Hai and Bui Khoi, "Optimal model choice using AIC Method and Naive Bayes Classification", *IOP Conference Series Materials Science and Engineering*, Print ISSN: 1757-8981, Online ISSN: 1757-899X, pp. 1-8, Vol. 1088, No. 1, February 2021, Published by IOP Publishing, DOI: 10.1088/1757-899X/1088/1/012001, Available: https://iopscience.iop.org/article/10.1088/1757-899X/1088/1/012001.

[20] Archana R. Panhalkar and Dharmpal D. Doye, "Optimization of decision trees using modified African buffalo algorithm", *Journal of King Saud University - Computer and Information Sciences*, Print ISSN: 1319-1578, pp. 4763-4772, Vol. 34, No. 8, February 2022, Published by Elsevier, DOI: 10.1016/j.jksuci.2021.01.011, Available: https://www.sciencedirect.com/science/article/pii/S1319157821000136.

[21] Yingbo An and Huasen Zhou, "Short term effect evaluation model of rural energy construction revitalization based on ID3 decision tree algorithm", *Energy Reports*, Print ISSN: 2352-4847, pp. 1004-1012, Vol. 8, July 2022, DOI: 10.1016/j.egyr.2022.01.239, Available: https://www.sciencedirect.com/science/article/pii/S2352484722002402.

[22] Jie Liu, Xin-Xing Feng, Yan-Feng Duan, Jun-Hao Liu, Ce Zhang *et al.*, "Using machine learning to aid treatment decision and risk assessment for severe three-vessel coronary artery disease", *Journal of Geriatric Cardiology*, ISSN: 1671-5411, pp. 367-376, Vol. 19, No. 5, May 2022, DOI: 10.11909/j.issn.1671-5411.2022.05.005, Available: http://jgc301.com/article/doi/10.11909/j.issn.1671-5411.2022.05.005.

[23] Leticia de Sousa and Igor D. Melo, "Interval power flow analysis of microgrids with uncertainties: an approach using the second-order Taylor series expansion", *Electrical Engineering*, Print ISSN: 0948-7921, Online ISSN: 1432-0487, pp. 1623-1633, Vol. 104, No. 3, 29th October 2022, Published by Springer Nature, DOI: 10.1007/s00202-021-01427-x, Available: https://link.springer.com/article/10.1007/s00202-021-01427-x.