



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

Effectively Analysis and Predict Students Performance and Other Evaluation

Arya R P, Anuja S B

Department of MCA, Narayanaguru College of Engineering, Anna University.
Assistant Professor, Department of MCA, Narayanguru College of Engineering, Anna University.

ABSTRACT

The development of intelligent technologies gains popularity in the education field. Educational Data Mining (EDM) is a research field of data mining, which focuses on the application of data mining, machine learning and statistical methods. The clustering effect of K-means Algorithm is tested by discriminant analysis. K-means Algorithm improves the reliability of prediction results. The development of intelligent technologies gains popularity in the education field. The rapid growth of educational data indicates traditional processing methods may have limitations and distortion. Therefore, reconstructing the research technology of data mining in the education field has become increasingly prominent. In order to avoid unreasonable evaluation results and monitor the students' future performance in advance, this paper comprehensively uses the relevant theories of clustering, discrimination and convolution neural network to analyze and predict students' academic performance. Firstly, this paper proposes that the clustering-number determination is optimized by using a statistic which has never been used in the algorithm of K-means.

Keywords: Academic performance, clustering analysis, convolutional neural networks, discriminant analysis, educational data mining.

1 Introduction

Predicting students' performance becomes more challenging due to the large volume of data in educational databases. There are two main reasons first, the study on existing prediction methods is still insufficient to identify the most suitable methods for predicting the performance of students. Second, is due to the lack of investigations on the factors affecting students achievements. Therefore, a systematical literature review on predicting student performance by using data mining techniques is proposed to improve students achievements. The high accuracy predication in students performance is useful as it helps to identify the students with low academic achievements at the early stage of academic. Data mining (DM) can discover hidden information in a large amount of unordered data. Educational data mining (EDM) is a research field of data mining, which focuses on the application of data mining, machine learning and statistical methods. In the past few decades, the application of data mining technology in the educational environment has become an active research field. Due to the



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

availability of online datasets and learning systems, it has gained wide popularity in recent years [1]. EDM involves the development and realization of data mining methods, which facilitate the analysis of massive data from various educational backgrounds. Higher education institutions regard students' academic performance as one of the most important criteria. Consequently, predicting learning process and analyzing students' performance are considered to be significant. Tasks in the field of EDM [2]. EDM is a continuously evolving subject which focuses on the improvement of self-learning and adaptive methods to reveal hidden patterns or internal relationship of educational data. In the field of education, heterogeneous data is participating and growing in the paradigm of big data. In order to adaptively extract meaningful information from massive educational data, some specific data mining techniques are needed [3]. Because technologies which are relevant to data mining make massive student information used to study valuable patterns of student learning behavior possible, the application research of EDM develops rapidly. Data mining methods have been applied.

The final exam scores of college students reflect the students' learning effects to some extent, but the evaluation of learning effects cannot only be based on absolute scores. The traditional absolute score has certain limitations in reflecting the learning situation. The reasons are that difficulty of different courses is different, the marking standards of different teachers in the same course are different, and so on. In order to ensure the quality of talents, colleges and universities should not only judge the students by scores, but also analyze the learning effects of students, predict the academic performance of students in the future based on the analyzed results, and then set academic warnings in time. This work will not only help colleges and universities to improve the quality of education, but also help students improve their overall performance, thereby improving the management of educational resources.

The research problem of this paper is to objectively evaluate students' academic achievement from the perspective of clustering, and predict the future achievement based on the existing achievement. To achieve the above-mentioned goals, this study formulates the following two subdivided research problems

- 1) How to determine the clustering number objectively in order to evaluate academic performance?
- 2) How to train unlabeled data with the implementation of constructing prediction model?

Therefore, the main contributions of this paper are

- 1) Improve the traditional K-means algorithm. Using objective and quantitative analysis instead of subjective
- 2) Evaluation to determine the value of k makes the academic performance evaluation results obtained by clustering more convincing. Furthermore, the models and prediction results trained by deep learning algorithm are more reliable.

The deep learning algorithm trains data with category labels which the metadata do not have. Clustering is used to add category labels to the metadata, and then deep learning algorithm is



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

used to train the data, which provides a new idea for deep learning without feature labels. The remaining sections of this paper are organized as follows. Section II reviews the literature that are relevant to the research of analysis and prediction in EDM and the application of convolutional neural network (CNN). Section III demonstrates the research methods and theory. Section IV discusses the experimental design and results, and analyzes the reliability of the prediction. Finally, it summarize the full text, highlight our limitations and introduce the research directions of EDM in the future in Section V.

2 Related Works

Many scholars pointed out the application of data mining technology in the field of education, and highlighted the importance of EDM. Angeli et al. [5] illustrated how educational technicians use data mining to guide and monitor school-based technical integration work. They discuss that the significance of the research lies in the necessity of developing EDM tools which can be used in meaningful ways and users can display results and suggestions. Javier et al. [6] pointed out that EDM combines data mining technology with educational data, and summarize the most commonly used data mining methods: factor analysis, regression, and correlation mining and so on. Romero et al. [7], [8] believed that the goal of EDM is to better understand how students learn, and it is also a method developed to explore the unique data types from the educational environment, which can be defined as the application of data mining technology. In order to optimize the education system, Wang [9] used data mining algorithms to classify and summarize educational data. Starting from the process of EDM to the algorithm process and evaluation model. Through effective evaluation of the development of smart education, a comprehensive evaluation model has been verified the high efficiency of intelligent education and the accuracy of data mining algorithms. In this section, it will review the literature on the application of data mining technology in the analysis and prediction of academic performance.

2.1 Research on Analysis in Edm

Some scholars apply clustering and other data mining techniques to the analysis of academic performance.

James et al. [10] pointed out that currently some scholars are using clustering analysis to analyze student performance and distinguishing students' categories based on their performance. They add a K-means clustering algorithm combined with a deterministic model to analyze student performance. Ani et al. [11] used entropy to study the degree of confidence changes over time, and use student data from eight courses to determine the student category based on the degree of student confidence. Moises et al. [12] used a clustering algorithm to detect six different student groups, analyze the interaction patterns of each category, and find that these patterns would be repeated in the early stages of the course. John et al. [13] used log data to identify the online behavior patterns of a group of students, and determine the student's



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

category by using clustering analysis. Karthikeyan et al. [14] developed a novel approach called hybrid educational data mining model for analyzing the student performance for effectively enhancing the educational quality for students. Crivei et al. [15] investigated the usefulness of unsupervised machine learning methods, particularly principal component analysis and relational association rule mining in analyzing students' academic performance data. Delgado et al. [16] used a new unsupervised clustering technology based on self-organizing mapping which performs accurate and diversified user clustering according to student behavior records. Okoye et al. [17] proposed an educational process and data mining plus machine learning (EPDM ML) model applied to contextually analyze the teachers' performance and recommendations based on data derived from students' evaluation of teaching. Kumar et al. [18] developed a model called Multi-Tier Student Performance Evaluation Model (MTSPEM) using single and ensemble classifiers.

2.2 Research on Prediction in Edm

Some scholars apply CNN and other data mining techniques to the prediction of academic Performance. Agaoglu [20] used four classification techniques to predict instructors' performance according to students' evaluation of courses. Qiu et al. [21] proposed an integrated framework with feature selection (FSPred) to predict the dropout in MOOCs, which includes feature generation, feature selection, and dropout prediction. Akram et al. [22] presented an algorithm called students' academic performance enhancement through homework late/non-submission detection (SAPE) for predicting students' academic performance. Based on the characteristics of MOOCs learning, Wen et al. [23] proposed a new simple feature matrix to maintain the information related to the local correlation of learning behavior, and propose a new CNN model to predict the dropout rate. Lin et al. [24] proposed a continuous facial emotion pattern recognition method based on deep learning to analyze students' learning emotions. This method combines CNNs and long-short memory networks for deep learning to identify and analyze students' continuous facial emotions and predict academic emotions. Farissi et al. [25] proposed a method based on genetic algorithm feature selection technique with classification method in order to predict student academic performance. Turabieh et al. [26] proposed a modified version of Harris Hawks Optimization (HHO) algorithm by controlling the population diversity. The proposed approach is employed as a feature selection algorithm to discover the most valuable features for student performance prediction problem. Ma et al. [27] proposed a new perspective called progressive imitation learning to train a lightweight CNN model by imitating the learning trajectory of the teacher model to construct a prediction model.

2.3 Summary

The ever-increasing amount of educational data makes educators, education-related personnel and even the general public focus on the important field of turning massive disorderly



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

educational data into useful information. Whether it is to predict student performance and dropout rate, or evaluate student academic performance, or evaluate teacher performance, data mining technology is widely used in all aspects. EDM researchers usually use automated methods such as data mining and machine learning to explore educational data, which is confirmed on the basis of the previous literature review. Combining computer technology with educational big data, knowledge of subjects such as mathematics, engineering, pedagogy and psychology can also be applied to EDM.

From the literature review, it can be found that EDM models can be divided into two types: descriptive models and predictive models. Descriptive models are used to describe models and provide reference for decision-making, whereas predictive models are mainly used for data-based prediction. The former is mostly used to evaluate students' academic performance and provide a reference for teaching managers to make decisions, whereas the latter is mostly used to predict students' academic performance, help prevent the risk of dropout, and improve students' academic performance. This paper intends to use a combination of descriptive and predictive models. This study will use clustering to analyze students' academic performance and convolutional neural network to predict students' future academic performance. Table 1 presents a comparison of the study with the work mentioned in recent three years. The differences between our method and existing ones are as follows

3 Modules

3.1 Data Pre-processing

Data pre-processing serves as the foundation for valid data which is an indispensable step in building operational data analysis considering the intrinsic complexity of data quality [31]. Data preprocessing refers to a series of necessary cleaning, integration, transformation and reduction of the original data before data mining, so as to achieve the minimum specifications and standards required by algorithms for knowledge acquisition research.

The method for determining clustering number is Different from others.

1. It use the clustering results as the label of the data to train the data.
2. It combine analysis and prediction.
3. It combine clustering algorithm and Data cleaning

Data cleaning refers to the elimination of incomplete, missing or duplicate data. There are many ways to fill in missing values for attributes, such as ignoring tuples, using a global constant to fill in missing values, using the mean of attributes to fill in missing values, etc. Delete the grade records of the courses with more missing courses, and fill in the grade records of the courses with fewer missing courses. This paper follows the following principles: Delete the score records with empty scores in more than two courses, and if there are still students whose course scores are empty, fill it with the average value of the course. It is understood that a course with a score of 0 is a student's absence from the exam, and the corresponding student's score record



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

is deleted.

3.2 Data Cleaning

Data cleaning refers to the elimination of incomplete, missing or duplicate data. There are many ways to fill in missing values for attributes, such as ignoring tuples, using a global constant to fill in missing values, using the mean of attributes to fill in missing values, etc. Delete the grade records of the courses with more missing courses, and fill in the grade records of the courses with fewer missing courses. This paper follows the following principles: Delete the score records with empty scores in more than two courses, and if there are still students whose course scores are empty, fill it with the average value of the course

3.3 Data Integration

Since some courses are divided into several semesters, merging these courses and taking the average score of several semesters as the score of the course is conducive to reducing the characteristics in the process of subsequent analysis. After merging, datasets A, B, and C have 9, 13, and 13 courses and 13 courses respectively for analysis.

3.3 Data Transformation

The original score data are presented in the form of a percentile system, with no difference of order of magnitude, and no standardized operation is required.

3.4 Data Reduction

In order to ensure the simplicity of the presentation results, irrelevant attributes such as credits and class time are omitted, and only the student number and the corresponding score of each course are retained

Table 1: Comparison of the study with the work mentioned in recent three years.

Reference	Year	Dataset	Method	Objective
Karthikeyan <i>et al.</i> [14]	2020	the benchmark education dataset	naive bayes, J48	analysis
Crivei <i>et al.</i> [15]	2020	a real academic data set	principal component analysis, association rule mining	analysis
Farissi <i>et al.</i> [25]	2020	kaggle repository datasets	genetic algorithm	prediction
Okoye <i>et al.</i> [17]	2021	data collected from students' evaluation of teaching	text mining, <i>k</i> -nearest neighbors	analysis
Kumar <i>et al.</i> [18]	2021	a benchmark student data set	naive bayes, random forest	
Turabieh <i>et al.</i> [2]	2021	a real dataset obtained from UCI	harris hawks optimization	
Nabil <i>et al.</i> [28]	2021	a dataset collected from a public 4-year university	deep neural network	prediction



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

Duhayyim <i>et al.</i> [	2022	a benchmark student performance dataset from UCI	an improved evolutionary algorithm based feature subsets election with neuro-fuzzy classification, chaotic whale optimization algorithm	analysis
Gao <i>et al.</i> [29]	2022	two real-world data sets	a deep cognitive diagnosis framework	prediction
Present Study	2022	three datasets in a university	improved K-means, CNN	analysis and prediction

3.5 Clustering Analysis and K-Means Algorithm

Clustering is one of the most common unsupervised data mining methods. Objects with similar characteristic attributes are placed in a category, and the characteristic attributes of the objects in different categories are different. The purpose of clustering analysis is simply to find a convenient and efficient way to organize data, not to establish rules for classifying future data [32]. Classification is to classify samples of unknown categories into a certain category according to the classification criteria that have been determined in advance, whereas clustering is to find features from the data and then classify them according to the features. The clustering results are closer to the actual situation, making the clustering analysis widely used in various areas. K-means algorithm, also known as a kind of fast clustering method, was proposed by Macqueen in 1967. K-means can maintain good scalability and efficiency when dealing with datasets, with strong local search ability and fast convergence speed. However, due to the sensitivity of K-means to outliers and initial clustering centers, the number of data points in different clusters may vary greatly, and it is often not easy to obtain ideal clustering results [33]. The pseudocode of K-means is provided in Algorithm 1. The procedures of the K-means algorithm: Arbitrarily select k samples from n samples as the initial clustering centers, and the initial clustering center is randomly determined. Assign all other sample to the nearest clustering center. Calculate the clustering center of each cluster, and Euclidean distance is used as the formula for calculating distance.

Algorithm 1 Clustering for Analyzing Academic

Performance.

Require: parameter k , inputs $x_1, x_2, \dots, x_n \in \mathbb{R}^d$

1: initialize clustering centers $c_1, c_2, \dots, c_k$

2: repeat

3: for $i = 1$ to n do

4: label the input x_i as belonging to the nearest cluster,

5:
$$\hat{y}_i = \arg \min_{c \in \{c_1, \dots, c_k\}} \|x_i - c\|$$



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

$i = 1, 2, \dots, k$

$j = 1, 2, \dots, k$

statistic is used to determine the clustering number, and then evaluate and analyze the academic performance.

5: end for

6: for $i = 1$:

do

discrimination and so on. Bayes discriminant method is base prior probability of the population to minimize the

7: compute clustering center c_j as the mean of all inputs

of the j th cluster, $c_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i$

8: end for

9: until no change in clustering centers between subsequent iterations

10: return clustering centers $c_1, c_2, \dots, c_k$

Discriminant Analysis and Bayes Discrimination

Discriminant analysis is a statistical analysis method to determine the category of samples. According to different discriminant criteria, discriminant analysis can be divided into distance discrimination, Fisher discrimination, Bayes discrimination and so on. Bayes discriminant method is based on the prior probability of the population to minimize the average loss of misjudgment [35]. As the clustering number increases, the number of Fisher discrimination functions will increase. Although the distance discrimination method is intuitive and easy to understand, it does not consider the prior probability of each category.

$D_\alpha(x) = \sum_{j=1}^k q_j f_j(x) L(\alpha/j) \quad (\alpha = 1, 2, \dots, k) \quad (2)$ be determined the Bayes criterion

3.6 Deep Learning and CNN

Hinton proposed a deep belief network in 2006. The neural network completed the transition from shallow to deep, and the concept of deep learning was born. It has turned out to be very good at discovering intricate structures in high-dimensional data [36]. Deep learning is a general term for a category of methods developed by machine learning, which refers to the laws in data, especially the classification laws. The filtering operation performed by the feature map is discrete convolution, hence the name of the convolutional neural network. CNN is a deep neural network learning algorithm proposed by LeCun et al. [36], which is a supervised



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

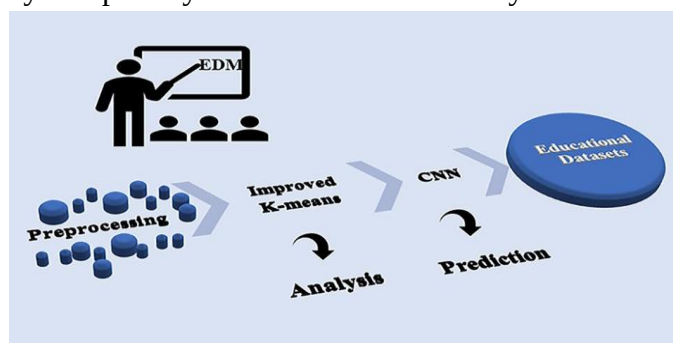
deep model architecture. The characteristics of local connection, weight sharing and pooling operation in Rk is relatively small.

Table 2: Confusion matrix

	Actual Category	
	Prediction Category	
	Positive	Negative
Positive	true positive	false negative
Negative	false positive	true negative

4 Experimental Results and Discussion

To ensure the universality of the analysis results, the scores of make-up exams and retakes were not included. There is a problem of missing marks in elective courses. If the results of elective courses are included in the model, it will seriously affect the performance of the algorithm and model. Therefore, only compulsory course data will be analyzed.



Teaching managers to make decisions and classification algorithm named CNN which is used to predict students' future academic performance and help prevent the risk of dropout. Cross-validation is a model validation technique applied to evaluate how the statistical analysis results are generalized into an independent dataset [39]. It prepare to use two methods for model validation. One is random hold-out method and the other is x-fold cross-validation method.

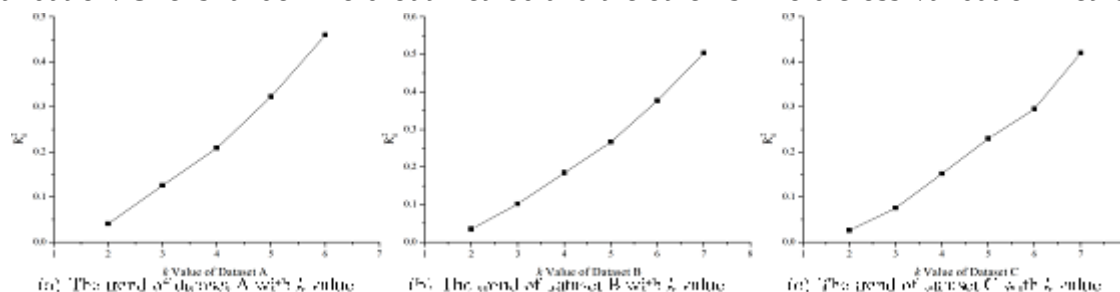
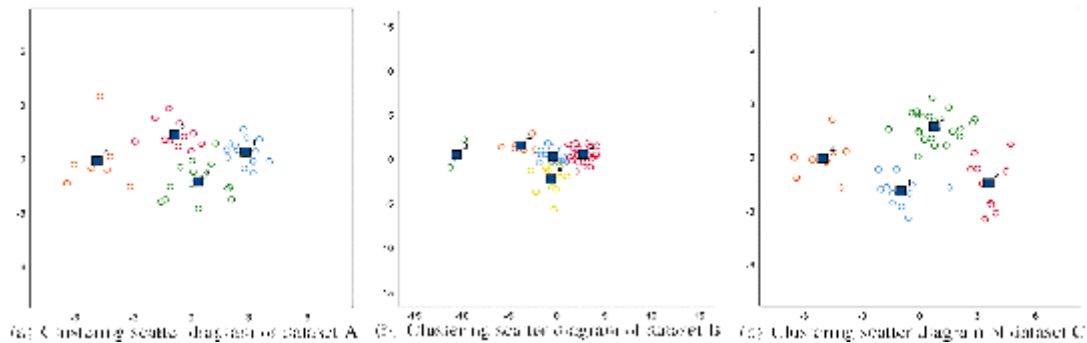


Table 3: The trend of each dataset with clustering number



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation



4.1 Determining K Value

Use the K-means algorithm to cluster the pre-processed data. At this time, the intra-category distance and the inter-category distance corresponding to different clustering numbers can be obtained, and the corresponding statistics can be calculated. The trend of each dataset with clustering number is shown in Fig. 2. According to the rules for determining clustering number and comprehensive consideration of the actual situation, it is more appropriate to cluster the datasets A, B, and C into four, five, and four categories, respectively.

4.2 Clustering Analysis

The clustering scatter diagrams of datasets A, B, and C are shown in Fig. 3. The square box is the clustering center, and the objects in this category surround the clustering center. The overall clustering results of datasets A and C are relatively scattered, and the objects in the category and their corresponding clustering centers are also relatively scattered, whereas the overall clustering results of dataset B and the objects in the category and their corresponding clustering centers are relatively close, so the boundaries among categories are not very obvious. The clustering results of datasets A and C are relatively scattered, and the number of objects in each category is not much different, but the number of objects in each category of dataset B with closer clustering results is relatively large.

4.3 Discriminant Analysis

After determining clustering number, in order to improve the accuracy of the subsequent prediction results, it is necessary to test whether the clustering effect is good or not. Before performing discriminant analysis, it is necessary to check whether the sample is suitable for discriminant analysis. In the equality test of the group mean, the significance of all the characteristics of the three datasets is less than 0.05, indicating that there are significant differences in the means within different categories, so it is reasonable to do discriminant analysis. The clustering results of K-means algorithm are basically the same as those of Bayes discrimination. Each student has a higher posterior probability value in the category to which they belong. The clustering analysis results of datasets A, B, and C are correctly aligned 95.7%,



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

98.4%, and 98.0% of the original grouped cases which were clustered, and the clustering effect was good. The discriminant analysis verifies the validity and reliability of the clustering results.

4 Conclusion

Considering there is a certain degree of irrationality and subjectivity in the results of the school's evaluation, by using the K-means algorithm in unsupervised learning to perform clustering analysis on student performance, and then using the clustering results as the category label of CNN, the paper starts with data mining. It is ultimately found that the model has a more ideal prediction accuracy, which is beneficial to ensure the objective and fair evaluation of students by school. Moreover, it is available to promptly remind students who are in a state of academic warning. When considering data labels, the selection range of label value must be involved, and the selection range of label value is related to clustering number. The K-means algorithm has a well-known defect: the value of k is artificially determined. To improve the algorithm, the paper uses an objective statistic to optimize k -value selection, and replaces subjective evaluation with quantitative analysis, which makes the clustering results more powerful. The persuasiveness also makes the training and prediction results of CNN more reliable, and the effectiveness of the model is naturally guaranteed.

References

1. Tsiakmaki Maria; Georgios Kostopoulos; Sotiris Kotsiantis; Omiros Ragos, Year: 2020, "Implementing AutoML in educational data mining for prediction tasks", Appl. Sci., Vol: 10, no: 1, pp. 90–117.
2. Kausar Samina; Xu Huahu; Iftikhar Hussain; Zhu Wenhao; Misha Zahid, Year: 2018, "Integration of data mining clustering approach in the personalized E-learning system," IEEE Access, Vol: 6, pp. 72724–72734.
3. Buenaño-Fernandez Diego; William Villegas-CH; Sergio Luján-Mora, Year: 2019, "The use of tools of data mining to decision making in engineering education—A systematic mapping study," Comput. Appl. Eng. Educ., Vol: 27, no: 3, pp. 744–758.
4. Angeli Charoula; Sarah K. Howard; Jun Ma; Jie Yang; Paul A. Kirschner, Year: 2017, "Data mining in educational technology classroom research: Can it make a contribution?" Comput. Educ., Vol: 113, pp. 226–242.
5. Bravo-Agapito Javier; Claire Frances Bonilla; Isaac Seoane, Year: 2020, "Data mining in foreign language learning," WIREs Data Mining Knowl. Discov, Vol: 10, no: 1, Art. no: e1287.
6. Romero Cristobal; Sebastian Ventura, Year: 2013, "Data mining in education," Wiley Inter- discipl. Rev., Data Mining Knowl. Discovery, Vol: 3, no: 1, pp. 12–27.



Article Title: Effectively Analysis and Predict Students Performance and Other Evaluation

7. Wang Shengnan, Year: 2019, “Smart data mining algorithm for intelligent education,” J. Intell. Fuzzy Syst., Vol: 37, no: 1, pp. 9–16.
8. Jamesmanoharan, J; S. Hari Ganesh; M. Lovelin Ponn Felciah; A. K. Shafreenbanu, Year: 2014, “Discovering students’ academic performance based on GPA using K-means clustering algorithm,” in Proc. World Congr. Comput. Commun. Technol., Tiruchirappalli, India, pp. 200–202.
9. Aghababayan Ani; Nicholas Lewkow; Ryan S. Baker, Year: 2018, “Enhancing the clustering of student performance using the variation in confidence,” in Proc. Int. Conf. Intell. Tutoring Syst. Cham, Switzerland: Springer, pp. 274–279.s
10. Karthikeyan V. Ganesh; P. Thangaraj; S. Karthik, Year: 2020, “Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation,” Soft Comput., Vol: 24, no: 24, pp. 18477–18487.
11. Crivei Liana Maria; Gabriela Czibula; George Ciubotariu; Mariana Dindelegan, Year: 2020, “Unsupervised learning based mining of academic data sets for students’ performance analysis,” in Proc. IEEE 14th Int. Symp. Appl. Comput. Intell. Informat. (SACI), Timisoara, Romania, pp. 11–16.