European Journal of Research Development and Sustainability (EJRDS)
Available Online at: https://www.scholarzest.com
Vol. 2 No. 11, November 2021,
ISSN: 2660-5570

AN APPROACH METHOD TO PREDICT STUDENTS' EXAM PERFORMANCE USING CLUSTERING METHODS WITH PREDICTION MODEL

Shallaw Mohammed Ali
Department of computer engineering techniques, Al-Qalam university college, Kirkuk, Iraq,
shalaw.eng@alqalam.edu.iq
Noor Jasim Mohammed
Electrical and computer engineering, Altınbaş, Istanbul, Turkey,
noor.jasiimm@gmail.com

Article history:
Received: 30th August 2021
Accepted: 30th September 2021
Published: 25h November 2021

Abstract:
The abilities of predicting human's behavior have increased dramatically in the new era of data mining applications. one of these applications is the attempts of predicting students' performance based on their activities and parental level of study. In this work, we present an approach method of predicting students' exam performance using clustering methods of (Fuzzy c-means, K-means and Hierarchal) combined with artificial neural network model of prediction. The results show that the use of clustering algorithms in the prediction process provides a high quality of prediction from (70% to 95%). This work also involves a comparison between these algorithms, which shows that the highest quality of predication can be obtained by using K-means method.

Keywords: Prediction model; clustering; K-means method; Fuzzy c-means; Hierarchal

1.INTRODUCTION

In the recent era of artificial intelligence and data mining possibilities, the importance and the applications of predicting and analyzing humans' behavior and performance have increased dramatically. One of these applications is the attempt of forecasting students' performance based on their activities and behaviors. in this area of study, researchers investigated different methods and approaches to enhance the prediction accuracy and address the challenges of behavior extracting and performance prediction. Authors tested the use of classification methods such as (K-Nearest Neighbor algorithm, Decision Tree and Bayesian) classifiers in the attempt of predicting the performance of students in exams and last tests. In [1], the authors investigated this approach by using K-NN classifier and Support Vector Machine algorithm. While researchers in [2] ,tested the (*Decision Tree, Bayesian, k-NN, and Rule learners)* to check the ability of extracting patterns in the datasets of students' activities to predict their final academic performance. Amra and Maghari [3] also proposed a prediction model through exploiting two classification algorithms (K-NN and Naïve Bayes). Furthermore, Shaheed et. al [4] presents an approach of performance prediction using Iterative dichotomiser 3 (ID3), C4.5 and Classification and Regression tree (CART). It can be noticed from above research studies that these efforts and attempts focuses on the use of classification methods and neural network prediction model for students' performance forecasting. These includes the targeting students' activities, parental graduation levels, attendances, and prior exam results. However, there were no consideration of using clustering methods in the approaches and models in these studies. Therefore, in this work, we propose a model to predict students' performance through exploiting clustering methods combined with Artificial intelligence classification model.

2. RESEARCH METHODOLOGY

In this work, we propose an approach model to combine the use of K-means, fuzzy c-means and Hierarchal clustering methods with artificial intelligence method of prediction (decision tree) to predict student performance based on their activities and parental status. This work also involves a comparative evaluation study between both clustering methods and the accuracy of the approach using each of these approaches. The following diagram shows the proposed prediction model in this work.

As shown in Figure 1, the methodology of this work starts by gathering dataset of students' activities and characteristics gathered from University of Minho, Portugal (from the website of http://www3.dsi.uminho.pt/pcortez.) Figure 2 shows a description of each of these attributes.

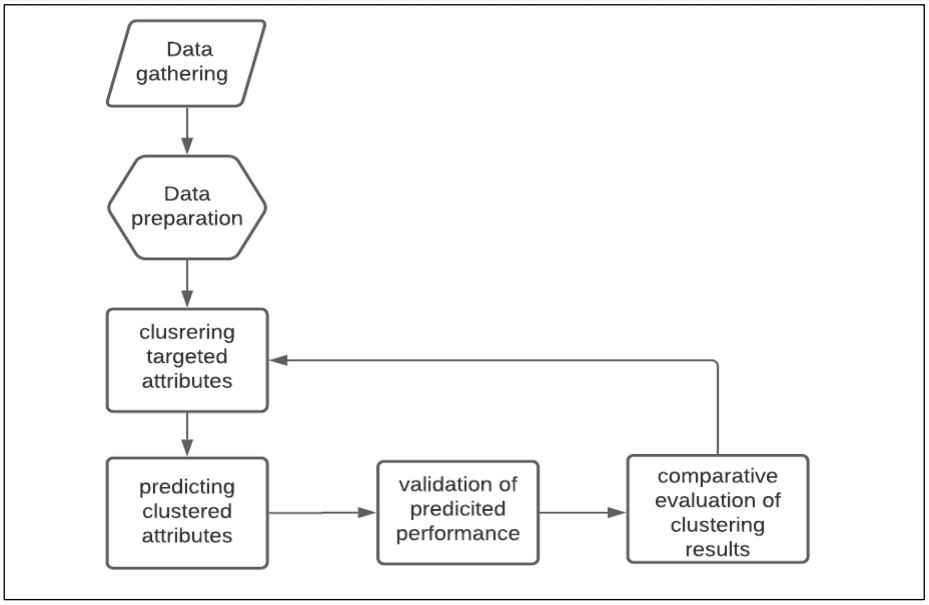

Figure 1: Diagram of proposed Approach model

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Figure 2: Attributes description [5]

Secondly, the preparation of the dataset is conducted by converting the nominal attributes such as (Mjob and Fjob) into numerical format which to be used for clustering process.

Then, the attribute (G3), which represents the final grade of each student in the scale from (0 to 20), is clustered to be used for prediction purpose. Data clustering is considered a data exploration technique that provides the ability to group objects with similar characteristics [6] as *"Clustering is the grouping of similar objects"* [7] . One of the well-known and popular clustering algorithms is K-mean.

The targeted attribute is clustered using two known clustering methods (K-*mean*, Hierarchal and *Fuzzy c-mean)*.

Then the clustered attribute(G3) is used in the prediction process through exploiting *Decision Tree* algorithm for prediction, which is one of the widely used algorithms for classification and prediction. It's an upside-down tree that makes decision through checking the status of the attributes in the datasets.

The implementation of the proposed prediction model conducted using *KNIME* toolkit. KNIME or Konstanz Information Miner is a *modular environment* that provides a user-friendly interface of workflow presentation of data mining tools [8].

In the KNIME we built the implementation environment using Workflow interface with node blocks provided by KNIME for each function algorithm as shown in Figure 3.

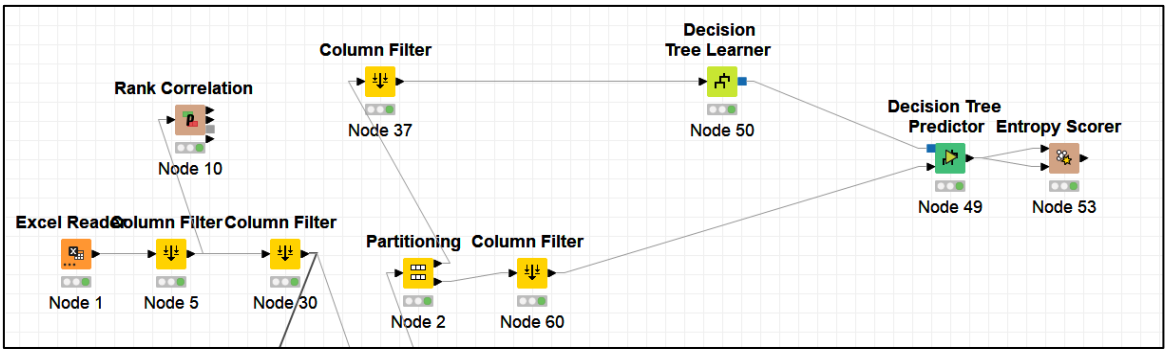

Figure 3: KNIME workflow

In this implementation, we used *Entropy scorer* results comparison.

3. RESULTS AND ANALYSIS:

In this work, we propose an approach model to combine the use of K-means, fuzzy c-means and Hierarchal clustering methods with artificial intelligence method of prediction (decision tree) to predict student performance based on their activities and parental status. This work also involves a comparative evaluation study between both clustering methods and the accuracy of the approach using each of these approaches. The following diagram shows the proposed prediction model in this work.

Table 1: prediction results of using different clustering methods

Clustering method	Entropy	Normalized Entropy	Quality
Fuzzy C-means	0.53	0.265	0.73
K-means	0.054	0.054	0.94
Hierarchal	0.12	0.12	0.88

According to Table 1, the results show that the quality of the prediction model ranges from 73% to 94%. Table 1 also shows that K-means method provides highest prediction quality of 94% in compared to other methods. While the quality of the prediction process using Fuzzy C-means reaches 73% and for Hierarchal method reaches 88%. The results also present Entropy measure, which is a measure of *true randomness in the dataset*. According to our results, K-means methods show the lowest Entropy with 0.054.

It can be inferred from above results, that the use of clustering methods in the process of predicting students' performance through artificial intelligent for classification, shows high prediction results of > 70%. Accordingly, K-means method shows the highest prediction quality of 94%.

4. CONCLUSION AND FUTURE WORK:

In the attempt of predicting students' performance based on their activities, several data mining models and algorithms were studied. Researchers investigated the use of classification methods such (K-NN, Naïve Bayes, Decision Tree and Bayesian) algorithms to implement the prediction process. However, the use of clustering models and was not investigated thoroughly. Therefore, we propose an approach model to predict students' exam performance using clustering methods combined with Artificial intelligent prediction model. In this work, we targeted student performance datasets gathered from University of Minho, Portugal. We conducted the prediction process by using (K-means, Fuzzy c-means and Hierarchical) clustering methods combined with Decision tree algorithm. This involves a comparison evaluation between these methods to detect the one with the highest prediction quality. The results show that the use of clustering methods in the process of students' prediction provides high quality of prediction accuracy. In addition, according to our results, K-means method of clustering show the highest quality of 94%.

For the future work, we intend to test more clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with different prediction models.

REFERENCES:

1. H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed and J. Alhiyafi, "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor," in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Saudi Arabia, 2017.
2. Shingari and D. Kumar, "Predicting Student Performance Using Classification Data Mining Techniques," *International Journal of Computer Sciences and Engineering,* vol. 6, no. 7, 2018.
3. A. Amra and A. Y. A. Maghari, "Students Performance Prediction Using KNN and Naïve Bayesian," in *8th International Conference on Information Technology (ICIT)*.
4. Y. K. Saheed, T. O. Oladele, A. O. Akanni and W. M. Ibrahim, "STUDENT PERFORMANCE PREDICTION BASED ON DATA MINING," *Nigerian Journal of Technology (NIJOTECH),* vol. 37, no. 4, 2018.
5. P. Cortez, "Student Performance Data Set," University of Minho, Portugal, 2014.
6. D. T. Pham, S. S. Dimov and C. D. Nguyen, "Selection of K in K-means clustering," *Manufacturing Engineering Centre,* pp. 103-119, 2004
7. J. A. Hartigan, Clustering Algorithms, 369: JOHN VVILEY & SONS, 1974.
8. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel and B. Wiswedel, "KNIME - the Konstanz information miner: version 2.0 and beyond," *SIGKDD Explorations ,* vol. 11, no. 1, 2009