AKADÉMIAI KIADÓ

UNIVERSITY of DEBRECEN

International Review of Applied Sciences and Engineering

12 (2021) 3, 269–277

DOI:
10.1556/1848.2021.00269

ORIGINAL RESEARCH PAPER

Extract hidden patterns in students' academic information to improve the curriculum by using data mining

Saeide Amerioon[1], Mohammad Mehdi Hosseini[1*] and Mahshid Moradi[2]

[1] Department of Computer Engineering, Shahrood Branch, Islamic Azad University, Shahrood, Iran

[2] Department of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

Received: February 27, 2021 • Accepted: May 6, 2021
Published online: July 2, 2021

ABSTRACT

Educational data mining is an emerging exquisite field that has been successfully implemented in higher education. One of the best ways to increase the efficiency of this emerging phenomenon is to select efficient professors and effective teaching methods. This study is intended to show academic success factors to have better management in student curriculum, contextualizing the progress and to prevent unfavorable conditions for students. In this research, students of Shahrood University of Technology were studied. Initially, 3,765 samples of students' educational background were considered. Then, pre-processing was performed to make the data normalized. Next, several predictive models were developed using a supervised data mining approach. Next, five algorithms by the best result were selected. Comparing the results of algorithms applied to data, the two algorithms, radial basis function network and the Naïve Bayes with respectively value F-measure 0.929 and 0.942 showed more accurate results. Finally, the most effective feature was selected, the attributes 'maximum semester' and 'the total number of units dropped' were ranked an the most important, and the three attributes of 'the basic courses average', 'the number of units of main courses' and 'the total average', were placed in the next ranks.

KEYWORDS

classification, data mining, educational data mining, feature extraction

1. INTRODUCTION

It is essential to predict the factors of academic achievement in order to identify talents, Provide Opportunities for successful individuals to grow, prevent unfavorable conditions for unsuccessful individuals and guide them towards better conditions. Therefore, discovering and studying the affective variables in higher education lead to better understanding and factors affecting them. Studying these variables is one of the main topics of research in education system in different countries. Such assessment can guarantee the efficiency and effectiveness of the scientific and practical resources and individual skills, which are the end-product of education field and increase the chances of students' success. Using data mining, we can extract appropriate patterns and improve planning curricula based on the data and provide appropriate solutions regarding which field of study students will be most successful in.

Due to the growing interest in e-learning as an important teaching and learning process, Rodrigues, Isotani & Zárate [1] mention the necessity for some mechanisms to evaluate educational effectiveness. Their review describes the e-learning scenario and the main topics of the educational aspect, which is a hidden problem. In their proposed approach, they show that research in educational data mining has expanded into several areas. Their review provided insights and identified trends and potential research directions. Fernandes et al. [2],

*Corresponding author.
E-mail: hosseini_mm@yahoo.com; hosseini_mm@shahroodut.ac.ir

AKJournals

in their proposed approach, showed that although 'grades' and 'absences' are most relevant to predict end-year output for student performance. Also, based on their demographic analysis, they showed (that) the three variables of 'neighborhood', 'school' and 'age' are potential indicators of students' academic success or failure.

In their proposed approach, Valles et al. [3] obtained and extracted valuable data to determine the different architectural needs of urban space, but concluded that accessing, organizing, analyzing, and visualizing this data can be challenging. The main purpose of their approach is to define a strategy and present a visualization of the results so that they can be attractive and instructive for both the students and the professionals. They also suggested that the analysis might produce similar results in other urban areas.

Costa et al. [4] also proposed a method in which they showed that the techniques analyzed in their study can identify students who are likely to fail, and they improved the effectiveness of some of these techniques after data processing or algorithms fine-tuning. The algorithms and the support-vector-machine technique significantly outperformed the other techniques. In another study, focusing on a small number of lessons and explicit performance indicators as 'good' or 'poor', Asif et al. [5] showed that this may provide timely warning and support for low-scoring students, and recommendations and opportunities for high-performing students. Yaacob et al. [6] suggested predictive models using classification algorithm to predict student's performance at a selected university in Malaysia. Several predictive modeling techniques were used to predict student's performance whether excellent or non-excellent. Finally, four supervised data mining algorithms were performed on the data to predict student performance. Also, the result of their experiments has shown that the model was accurate and comprehensible. Olsen, Aleven & Rummel [7] used the Additive Factors Model (AFM) in their model, as a standard logistic regression model for modeling individual learning, which is in line with knowledge component models and tutor log data. Their extended model predicts students' performance regarding problem solving collaboratively. Regarding the students working collaboratively, the model also has shown the effective study partners may improve the learning of others in the class. They found that both collaborative attribute (helping and being helped) improve the model fit. Abubakari et al. [8] suggested a neural network algorithm to extract knowledge patterns. Their dataset consisted of 480 instances with 16 attributes for each student. First, Adam model optimizer was used and in the second step, they applied the stochastic gradient descent optimizer and dropout technique, which increased the accuracy to more than 75%.

Chaudhury et al. [9] developed a model for predicting students' performance and thereby identified the students who might underperform in examinations. Demographic and academic information of students was considered in this study. They performed analysis on different attributes using feature subset selection algorithms. Finally, student data feature set (SDFS) was proposed. Their model can be utilized to identify the academically weak students so that appropriate preventive action can be taken to avoid failures. Chrisnanto et al. [10] determined a student's performance in terms of context at a particular time during their education using data mining techniques. They applied integrated data mining model in which the techniques used include the classification (ID3, SVM), clustering (k-Means, k-Medoids), association rules (Apriori) and anomaly detection (DBSCAN). The results of this study indicated the use of several techniques in data mining together could maximize the ability to analyze academic performance. The data mining techniques used will be integrated into one data mining system for academic purposes. Therefore, the main objective of this paper is to develop predictive models using classification algorithms to predict student's performance at Shahrood University of Technology in IRAN. The developed prediction model can be used to identify the most important attributes in the data. The paper is organized into four main sections. In the first section, it is described how to extract students' success factors in two general stages. The proposed method is discussed in the second part, and the experiment results are described in the third section. In the final section, the conclusion of our proposed work is drawn.

2. SUGGESTED METHOD

The prerequisite knowledge or understanding of the background of the domain is required in the first step. The problem can be effectively solved only with good understanding of the problem. Knowing the factors affecting academic progress can improve educational planning and increase the domain of teaching and learning. At this point the goal of the study is to develop a new model that predicts whether a student will attain rules to determine success at the end of the study period. The modeling in this study is conducted in four steps. Figure 1 shows the outline of the proposed method.

In first step, the data were initially processed, followed by 4 stages of data cleaning, data reduction, non-applicable data, and data transformation. After data pre-processing and normalization, in the second step, by reducing the size of the data and sampling them appropriately, 3,765 samples were selected from 4,240 data with 27 attributes and an overview table consisting of a graduate student's information was obtained. In the third step, the five preferred classification algorithms were selected based on the output results in three data modes (equal academic discipline, equal status class, equal academic discipline and status class). Next, best attributes out of 27 attributes based on F-score were chosen. In the fourth step, prioritization of the attributes, reducing the features was performed for three modes and the overall result was obtained. Since the data are analyzed from different technical and engineering academic disciplines of Shahrood University of Technology, a procedure was proposed to create various databases. The aim was to establish similar conditions for different samples and different data analyses for different academic disciplines.

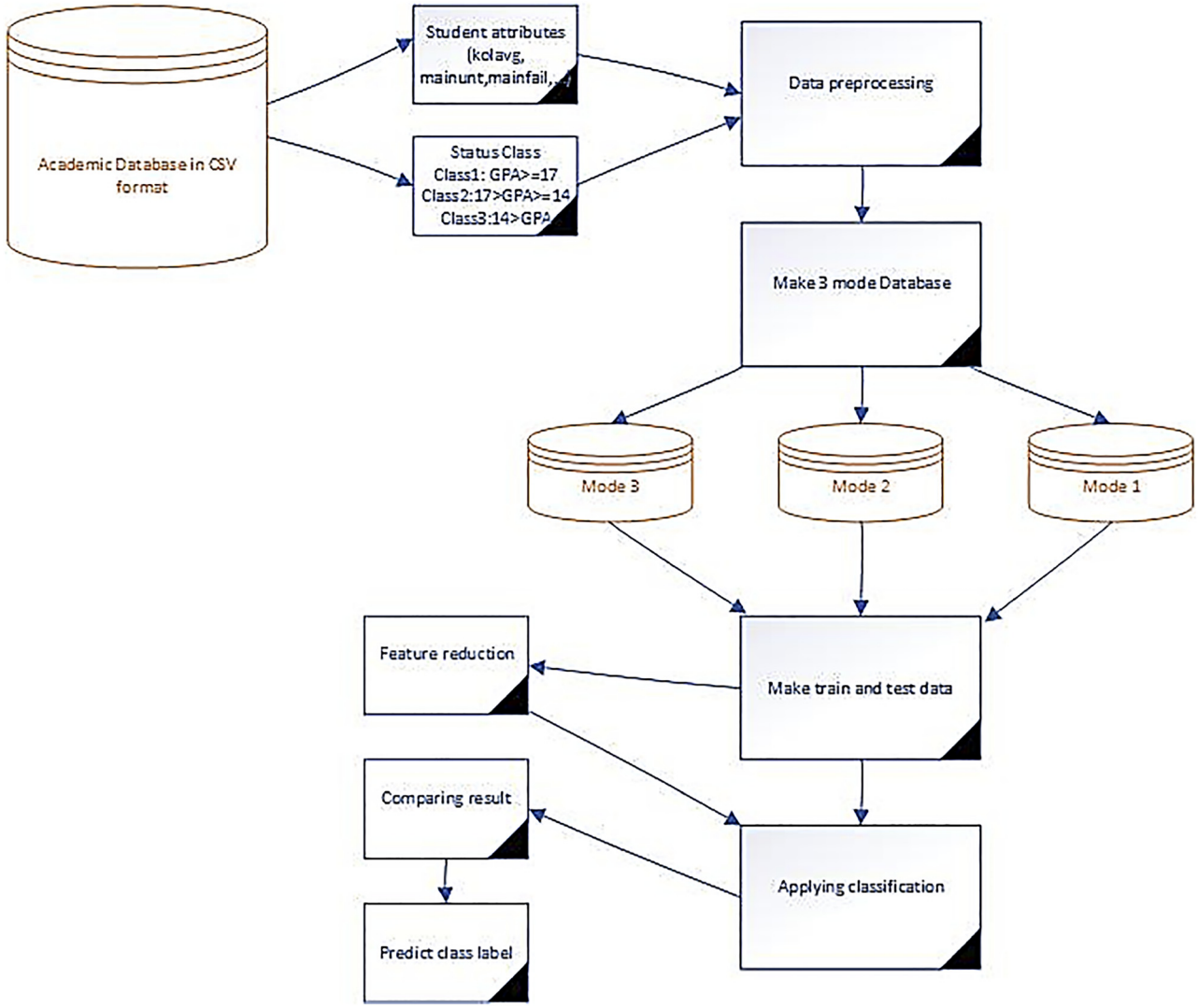

Fig. 1. The structure of the proposed method

Since the status class was considered an external factor in data analysis, the extent of this factor's interference in information analysis should be considered. This step is critical to extract the main causes. Therefore, the information analysis of the proposed method was investigated in three stages.

2.1. Pre-processing

The best algorithm is useless unless your data are ready for the implementation. Data preprocessing is an integral step in data mining as the quality of data and the new insight that can be obtained from it are highly influential on the performance of models. This step contains several preprocessing tasks, including data cleaning, data reduction, non-applicable data and data transformation.

- Data cleaning

Data preprocessing is considered as the most boring task, yet important in data mining modeling. It is an important task in data engineering which is carried out by a group of people since it concerns various data operations. The tasks require the basic knowledge of exploratory data analysis and statistical knowledge. The most important functions of this section are estimating the missing values in the database, eliminating noise disturbance in the data, removing outdated and unrelated data, and eliminating inconsistencies in the data. The important point is that the better this step of data mining is performed; the output of algorithms and data mining techniques could have been better. In data clearing section, the inaccurate or invaluable data were replaced with zero.

- Data reduction

One of the most important points in data mining is that it may not always provide all the data is needed and only part of the data that is needed has to be processed. Using data reduction, 3,765 samples were separated from 4,240 data.

- Non-applicable Data

Here, data usefulness is also confirmed in terms of meeting the desired goal. In the context of this research,

existing values that are reasonably impossible or not specifically related to the domain are deleted from the data.

- Data transformation

In many cases, data may be stored in different files and sources, and need to be integrated before implementing data mining techniques. After adding the file to the statistical software, it is seen that the type of all variables is not the same. To integrate the variables, using the proper filtering, all data were converted to nominal form. using a special filter and selecting an efficient column, which in this method is a "status class", different categories were examined.

2.2. Classification algorithms

Various educational data mining methods, obtained from the literature review in this study, were considered. A comparative study of prediction models was conducted aimed to predict student performance. In classification algorithms, the classification model is constructed by the training data set and evaluated by the experimental data set. In this method, the ratio of the data set for the two training data sets and the experimental data sets depends on the analysis of the analyst, usually two-thirds for training and one-third for testing. So, 70% of data is considered to train the model and 30% of data was allocated to test the model. The main advantage of this method is the simplicity and high speed of the evaluation operation [11]. The analysis is adopted using four classifiers:

2.2.1. Decision tree (C4. 5 algorithm). The C4. 5 algorithm is used in Data Mining as a Decision Tree Classifier which is commonly used in various research and operations. Decision tree analysis is particularly used to identify strategies, which are likely to be targeted in the problem [12]. Decision tree is a decision support tool that uses trees for making models. The decision tree models developed in this study are based on splitting criteria of Information Gain. The information gain defines purity. It means, if the leaf is highly impure, then it is complicated. The higher value information Gain indicates the preference of feature for discrimination of class value.

2.2.2. Naïve Bayes algorithm. The simple Bayesian algorithm, or Naïve Bayes, is a data mining method based on the theory of probability. In Naive Bayes structure, no other relationship is allowed and all attributes are assumed to be independent of each other. It works based on the Bayes Theorem with the independence assumptions between predictors. It is a statistical classifier which represent visually as a graph structure. Naïve Bayes algorithm presumes that the effect that features do on a given class is independent of the values of other features [13].

2.2.3. AdaBoost algorithm. The AdaBoost technique follows a decision tree model with a depth equal to one. AdaBoost is nothing but a forest of stumps rather than trees. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. AdaBoost, short for "Adaptive Boosting", is the first practical boosting algorithm. It makes a number of decision trees during the training period of data. As the first model is made, the record which is incorrectly classified during the first model is given more priority. These records are sent as input for the second model. The process will go on until specifying the number of base learners wants to create. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. Boosting algorithms are enough for most of the predictive learning tasks. They are powerful, flexible, and can be interpreted nicely with some tricks. AdaBoost algorithm also works on the same principle as boosting, but there is a slight difference in working [14].

2.2.4. K-Means clustering. K-Means Clustering is one of the famous unsupervised learning algorithms that is used to solve the clustering problems in data mining [15]. This method groups the unlabeled dataset into different clusters. K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct clusters where each data point belongs to only one group. This method tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different. This algorithm aims at minimizing an objective function known as squared error function given by Eq. (1):

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{C_i} \left(\left\| x_i - v_j \right\| \right)^2 \quad (1)$$

$\left\| x_i - v_j \right\|$ is the Euclidean distance between x_i and v_j. 'C_i' and 'C' are respectively the number of data points in i^{th} cluster and the number of cluster centers.

2.2.5. Radial Basis Function Network (RBF Network). RBF network are a commonly used type of Neural Network for approximation problem. These networks are used in approximation, time series prediction, classification and control functions. A RBF network, in its simplest form, is a three-layer feedforward neural network, namely the input layer, the hidden layer and the output layer. The RBF network requires less time to reach the end of training compared to Multi-layer perceptron. The output of this network is a linear combination of radial basis functions for input parameters and neurons [16].

2.3. Feature selection

Feature selection is applied before classification. Feature selection is one of the crucial steps in unsupervised learning. As the name indicated, feature selection removes unnecessary attributes from the dataset aimed to extract useful and meaningful information. There are several feature selection approaches, it falls into three main categories: filter, wrapper, and hybrid or embed methods [17]. The Bestfirst evaluator used in correlation feature selection (Cfs) and the search method. In the evaluator, Cfs selects a set of attributes that is highly correlated with the class attribute, but with a low intrinsic correlation by examining the predictive ability of each attribute in particular and the degree of redundancy

between them. The Bestfirst search method also performs a greedy hill climbing method using the rollback method. This method can search for the empty set of attributes forward, or search the entire set backwards, or start searching from one midpoint in either direction.

3. MATERIALS AND METHODS

This section reports the performance and gives a comparative result of the five proposed classifiers with optimal models in three-mode on our dataset. The experiment was carried out in three phases. At first, an equal number of academic disciplines were assumed. In the second step, an equal number of status classes were assigned, and in the third stage, equal academic discipline and the status class was checked out. The purpose of this work was to establish equal conditions for different samples and different data analysis for different academic disciplines. The present study has confirmed the effectiveness of models and the generalization of datasets by multiple datasets. To evaluate and compare the performance of prediction models, F-score and RMSE are measured. The higher F-score, the better the model. In contrast to RMSE, the smaller value, the better model.

3.1. Database description

The original/real dataset is obtained from students' information of Shahrood University of Technology in IRAN. Initially, the data were processed by applying queries to the tables and calculating the required values in Mysql. Finally, the data were prepared in Excel format. The data consist of students' personal information, semester, gender, student total average, number of courses (general, basic, major), course average (general, basic, major), course information, student information, college information, and teacher information and another information, each in separate sheets to predict student's success or failure upon graduation. The proposed model can help extracting patterns of students' education information and performance levels so that the curriculum will be implemented. Due to privacy, personal information is removed, and there are 27 input features to learning in this study. The target of modeling is discretized based on the score. To extract the information of the students, a table should be created for the general information of the students in such a way that its rows are students and its columns include important information of the graduates.

Then, three tables of student information, course information and term information were linked in two steps (first student table and course information table together and then its results with student's term information table). The result of these three tables was called the student table. This table provides important general information about the graduate student.

After obtaining the table and analyzing its information, the result shows not all the columns obtained were useful for the analysis of graduate student information. Therefore, the acceptable columns were stored in another table. For ease of analysis, academic disciplines information was also separated in StudenTotal. The separated columns at this stage are: Id, sysno, facno, grpno, depno, entrtrm, gen, stdstat, stdlasttrm, entrtrm, stdstat. At this stage some information such as total average, total unit number, total number of units failed, total number of units, main units average, total number of main units, total number of basic units, total number of general units passed and student's region were extracted and added to the student Total table. Next, a table was created which contains acceptable general information for graduation, as called "Total table", in which the rows show the student information and the columns show the following information: Id, sysno, facno, grpno, depno, entrtrm, gender, stdstat, stdlasttrm, totavg, totunt, totuntpas, totuntfail, mainunt, mainavg, genavg, genunt, baseavg, baseunt, maxtrm, reentrance (Table 1).

3.1.1. Building various database. After building the database, the data output was stored in 3 modes of 3,765. Each of these 3 modes includes 3 different experiments performed under the same conditions.

- **Mode 1** for achieving 3,765 rows because the data were composed of six engineering academic disciplines and were selected in equal numbers from each academic

Table 1. General fields in the database

	Description	Field
1	Student ID	StdId
2	Education system	SysNo
3	Faculty number	FacNo
4	Group number	GrpNo
5	Department number	DeptNo
6	Entry term	EntrTerm
7	Gender	Gen
8	Student status	StdStat
9	Student in the last term	StdLasTerm
10	Total average	TotAvrg
11	Total number of units	TotUnt
12	Total number of units passed	TotUntPass
13	Total number of units failed	TotUntFail
14	Number of Main units	NoManUnt
15	Main units Average	ManUntAvrg
16	Number of General units	NoGenUnt
17	General Units Average	GenUntAvrg
18	Number of Basic units	NoBasUnt
19	Basic units Average	BasUnitAvrg
20	Conditional status or not	Cond
21	Maximum of terms	MaxTerm
22	Region	Regn
23	Course type	CrsType
24	Student Status (studying/graduate/ dropout/expulsion)	StdStat
25	Number of main units failed	ManUntFail
26	Number of general units failed	GenUntfail
27	Number of basic units failed	BasUntFail
28	Status class	ClsStat
29	Score	Scor

discipline, it resulted in selecting approximately 6,275 rows from each academic discipline, in which the disciplines changed according to change in the faculty value.

- **Mode 2** the criterion of equality is the status class, where for each case, 1, 2, 3, there were 1,255 students in different academic disciplines. The status classes were kept equal, and finally there were 3,765 rows.
- **Mode 3** in this mode, like in both of the above modes, namely both the academic discipline and the status classes in 1, 2 and 3 were in the equal selected condition. Since there are 6 different academic disciplines and three status classes, 209 students were necessary.

For analyzing information, each 3,765-item table requires a class that can be used to analyze the rest of the attributes. To calculate the status class, the information was divided into three statuses 1, 2 and 3. Status 1 represents the best with average score of 17–20, status 2 with an average score of 14–16.99, and status 3 with an average score below 13.99. At the end of this section, an output table was obtained with a maximum of 3,765 rows and 27 main columns.

3.2. Performance evaluation metrics description

The proposed model needs to be evaluated before the final selection of models. In this study, two standard model evaluation metrics are utilized. These two metrics are F-measure and Root Mean Square Error (RMSE). The data set is randomly split into 10 subsets of training and testing size for 10-fold cross validation to evaluate the classifier. Then, the performance measures were calculated. F-score obtained based on the value of precision, which is the number of objects in positive classes that are correctly classified and recalled or a fraction of relevant instances that were retrieved. A combination of precision and recall criteria can be used to evaluate the efficiency of recovery called the F-measure:

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{2}$$

where P is precision and R is recall. Therefore, the F criterion (2) is considered as a criterion to compare results [18]. RMSE is utilized for the prediction error. RMSE can be computed as follows (3):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - e_i')} \tag{3}$$

where e_i is the actual performance level and e_i' is predicted performance level.

4. RESULT AND ANALYSIS

In this paper, the classification was performed using the supervised data mining predictive model based on k-NN, Decision Tree (C4.5 algorithm), naïve Bayes, AdaBoost and RBF Network Model. 10-fold cross validation was performed to validate each classifier. The average performance on test dataset can be computed to determine the accuracy of the proposed model. Figure 2 shows the result of RBF network classifier at different k values. The same work can be done for other classifier to select best k value with the highest accuracy. Based on the obtained result, the best k value that can achieve the highest accuracy is eight.

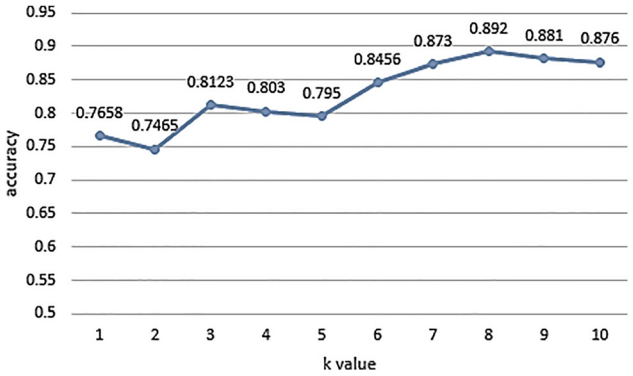

Fig. 2. A sample of the accuracy level RBF network for different K-value in Mode 1

4.1. Experiment result in 3 mode of dataset

In this step, each classifier is applied on data and their accuracy is calculated. However, comparing the performance of five classifiers on datasets is demonstrated in Table 2. Referring to Table 2, the f-measure and RMSE for each classifier is shown. The naïve Bayes offers a higher weighted classification rate based on F-measure with 0.883 and RMSE with 0.139. In addition, Naive Bayes algorithm has obtained better results due to being close to 1. In the next two cases, the procedure will be the same.

The results presented in Tables 3 and 4 demonstrate the performance of the five classification-datasets in Mode 2 and Mode 3. The performance of the two modes leads to different results.

In the second case, according to Table 3, the weighted F-Measure of the three algorithms C4.5, Naive Bayes and RBF networks with respective values of 0.866, 0.871, and 0.846 achieved better results due to being close to 1. Finally, in Mode 3, according to the obtained result, the F-Measure values of the three algorithms C4.5, Naive Bayes and RBF network with respective values of 0.883, 0.911 and 0.906 have obtained better results due to being close to the value of 1.

Table 2. Output table of learning algorithm and F-Measure on Mode 1

Class	F-score			RMSE		
	1	2	3	1	2	3
RBF Network	0.892	0.861	0.902	0.18	0.11	0.128
K-Means	0.634	0.733	0.822	0.43	0.23	0.163
AdaBoost	0.593	0.836	0.842	0.52	0.32	0.22
Naïve Bayes	0.828	0.901	0.910	0.24	0.14	0.14
C4.5	0.621	0.710	0.731	0.29	0.24	0.292

Table 3. Output table of learning algorithm and F-Measure on Mode 2

Class	F-score			RMSE		
	1	2	3	1	2	3
RBF Network	0.743	0.884	0.913	0.15	0.181	0.117
K-Means	0.801	0.711	0.645	0.35	0.27	0.303
AdaBoost	0.702	0.662	0.775	0.47	0.52	0.27
Naive Bayes	0.817	0.876	0.922	0.14	0.116	0.105
C4.5	0.840	0.863	0.895	0.20	0.23	0.192

Table 4. Output table of learning algorithm and F-Measure on Mode 3

Class	F-score			RMSE		
	1	2	3	1	2	3
RBF Network	0.915	0.922	0.883	0.105	0.281	0.147
K-Means	0.735	0.685	0.787	0.225	0.27	0.323
AdaBoost	0.721	0.785	0.701	0.37	0.331	0.251
Naïve Bayes	0.89	0.917	0.927	0.124	0.109	0.129
C4.5	0.882	0.892	0.877	0.191	0.183	0.167

4.2. Feature selection result

Features play a vital role in Machine Learning to generate a good model. Similarly, the identification of a good feature is considered as a drawback in educational data mining. This section summarizes the result of feature selection method utilized for selecting the superior set. After applying Bestfirst algorithm in this method, the result of each mode is indicated as a below:

- First mode results: mainfail, maxtrm, Basefail, Crstype, Mainunt, Stdlasttrm and totuntfail
- Second mode results: totavg, maxtrm, Mashtag, Mainunt, Basefail, Rentrance and totuntfail
- Third mode results: maxtrm, baseavg, Rentrance, Failkolunt, Crstype, Stdlasttrm and totuntfail

According to the results, 14 out of 27 attributes play an important role in determining students' success and improving these factors will have a significant impact on their educational status. Based on the results above, we found out that 14 of the 27 attributes are important. To figure out the importance of each attribute, a table with 3,765 rows and 14 columns was created, the value of which is a combination of the same data and the preceding

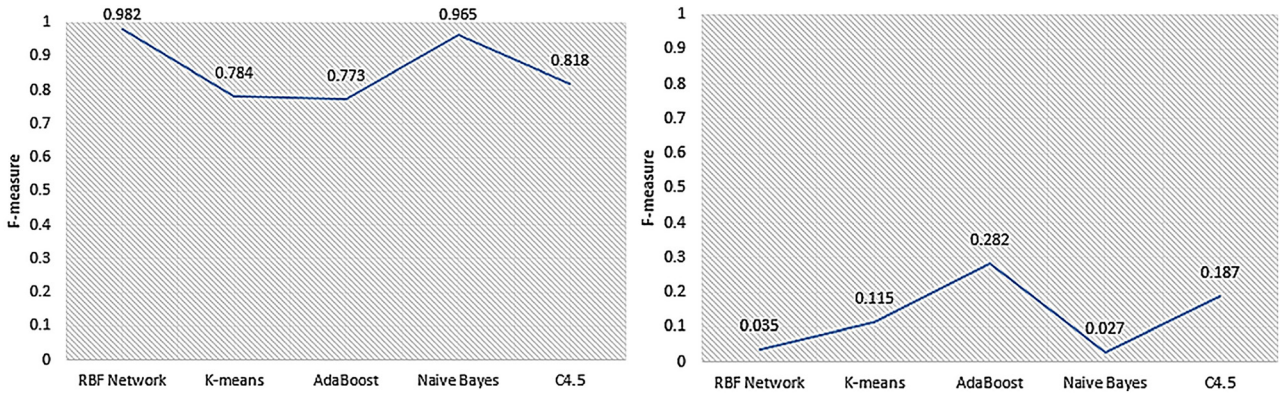

Fig. 3. F-score and RMSE comparison in Mode 1

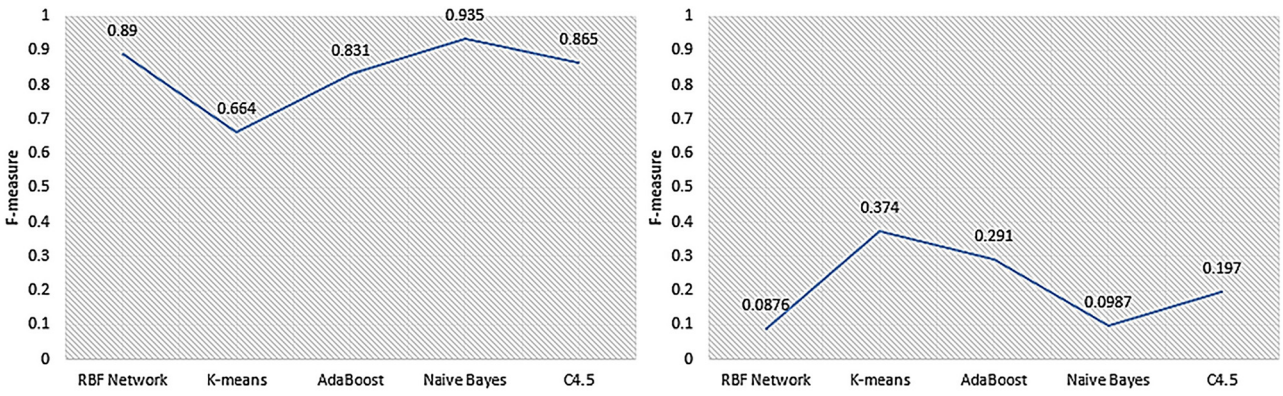

Fig. 4. F-score and RMSE comparison in Mode 2

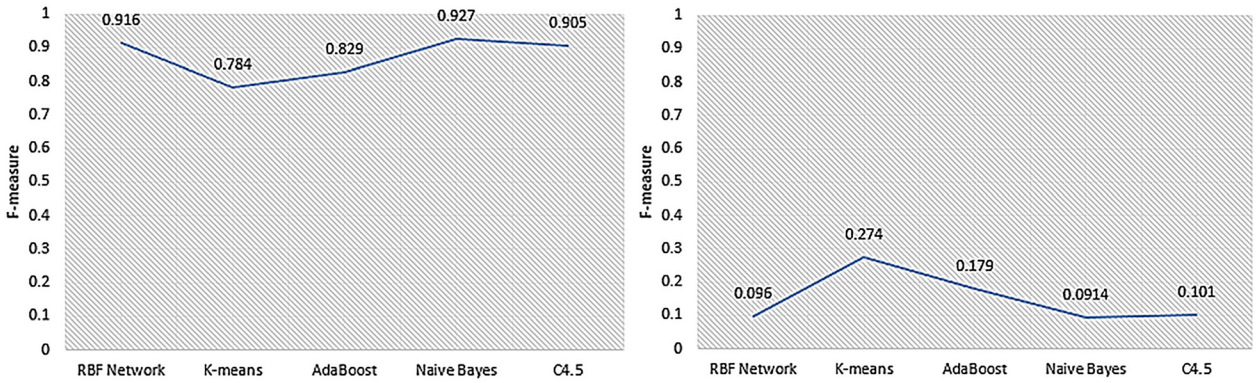

Fig. 5. F-score and RMSE comparison in Mode 3

columns. Then, through re-analyzing and re-testing the values, the priority of the features is clear.

4.3. Experiment result with feature reduction

As the last step, after best feature selection, the feature matrix for 3,765 records was created again. The test was completed. The performance of models on the test was computed to determine the accuracy of the model developed. The result in first mode shows the F-score for Naïve Bayes and RBF network algorithms is higher than the other 3 classifiers. Figure 3 demonstrates this comparison.

In the second mode, the F-score for 3 algorithms C4.5, Naïve Bayes and RBF network is higher than the other 2 algorithms. Figure 4 illustrates the comparison between performance metrics for five classifiers.

In the third mode, the F-scores for the 3 algorithms, NaiveBayes, RBF network and C4.5 algorithm are higher than the other 2 algorithms. According to the diagrams, the 2 NaiveBayes and RBF network algorithms have presented acceptable results in three modes (Fig. 5).

Finally, the analysis of data demonstrated that 14 features are more effective than other features. Table 4 presents a list of these features.

5. CONCLUSION

The proposed method aimed to demonstrate one of the most basic data mining capabilities, namely the discovery of new latent patterns for predictive analysis, and an analytical approach to predicting the success or failure of students at Shahrood University of Technology. The 3,765 samples out of the 4,240 data sets were selected after preprocessing step. The data was split into 70% as training and 30% as a testing set. After applying the mentioned algorithms, the order of importance of attributes and more valuable algorithms were found. Our findings confirm the effectiveness of the prediction model and usability of these features. Based on a descriptive report, the developed model is reported to be useful for educational collaborators and related people at Shahrood University of Technology. Therefore, related individuals can adapt their learning methodology and set up the policy to improve academic performance. For better results, it is recommended to use parameters such as the quality of high school education, high school grade average, type of tuition, studying locally, and others. These parameters can be effective in the model obtained, but these data were ignored due to the lack of complete data in the database. Using these parameters in order to predict student's success factors and improve curriculum planning can be a topic for future research.

ACKNOWLEDGMENT

This work would not have been possible without the constant support of my colleagues at Shahrood University of Technology. I appreciate the generosity of the Shahrood University of Technology to share the data.

REFERENCES

[1] M. W. Rodrigues, S. Isotani, and L. E. Zárate, "Educational data mining: A review of evaluation process in the e-learning," *Telematics Inform.*, vol. 35, no. 6, pp. 1701–17, 2018.

[2] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Business Res.*, vol. 94, pp. 335–43, 2019.

[3] F. Valls, E. Redondo, D. Fonseca, R. Torres-Kompen, S. Villagrasa, and N. Martí, "Urban data and urban design: A data mining approach to architecture education," *Telematics Inform.*, vol. 35, no. 4, pp. 1039–52, 2018.

[4] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Hum. Behav.*, vol. 73, pp. 247–56, 2017.

[5] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177–94, 2017.

[6] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student

performance," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, pp. 1584–92, 2019.

[7] J. K. Olsen, V. Aleven, and N. Rummel, *Predicting Student Performance in a Collaborative Learning Environment*. International Educational Data Mining Society, 2015.

[8] M. S. Abubakaria, F. Arifin, and G. G. Hungilo, "Predicting students' academic performance in educational data mining based on deep learning using TensorFlow," *Int. J. Educ. Manage. Eng. (IJEME)*, vol. 10, no. 6, pp. 27–33, 2020.

[9] P. Chaudhury and H. K. Tripathy, "An empirical study on attribute selection of student performance prediction model," *Int. J. Learn. Technol.*, vol. 12, no. 3, pp. 241–52, 2017.

[10] Y. H. Chrisnanto and G. Abdullah, "The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI)," *Matrix: J. Manaj. Teknol. dan Inform.*, vol. 11, no. 1, pp. 26–35, 2021.

[11] S. Smusz, R. Kurczab, and A. J. Bojarski, "A multidimensional analysis of machine learning methods performance in the classification of bioactive compounds," *Chemom. Intell. Lab. Syst.*, vol. 128, pp. 89–100, 2013.

[12] A. K. Yadav and S. S. Chandel, "Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model," *Renew. Energy*, vol. 75, pp. 675–93, 2015.

[13] N. Sun, B. Sun, J. D. Lin, and M. Y. C. Wu, "Lossless pruned Naive Bayes for big data classifications," *Big Data Res.*, vol. 14, pp. 27–36, 2018.

[14] H. Majidpour and F. Soleimanian Gharehchopogh, "An improved flower pollination algorithm with adaboost algorithm for feature selection in text documents classification," *J. Adv. Comput. Res.*, vol. 9, no. 1, pp. 29–40, 2018.

[15] A. F. Jahwar and A. M. Abdulazeez, "Meta-heuristic algorithms for K-means clustering: A review," *PalArch's J. Archaeol. Egypt/Egyptol.*, vol. 17, no. 7, pp. 12002–20, 2020.

[16] A. A. Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques," *Technol. Knowl. Learn.*, vol. 24, no. 4, pp. 567–98, 2019.

[17] X. Huang, L. Wu, and Y. Ye, "A review on dimensionality reduction techniques," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 10, p. 1950017, 2019.

[18] L. Bautista-Gomez, A. Benoit, A. Cavelan, S. K. Raina, Y. Robert, and H. Sun, "Coping with recall and precision of soft error detectors," *J. Parallel Distrib. Comput.*, vol. 98, pp. 8–24, 2016.