International Journal of Scientific & Engineering Research, Volume , Issue ,
ISSN 2229-5518

A Proposed Framework Targeting the Enhancement of Students' Performance in Fayoum University

Amira M. Idrees, Mohamed Hassan Ibrahim

Abstract— One of the most important goals of the higher education field is to provide perfect education level by achieving the highest level of quality that could be achieved. Mining data is one of the successful methods to achieve this goal. Applying prediction techniques can discover the most suitable subjects for the students, determine the odd data in the grades of the students and all other prediction scenarios for the learning process. This paper is oriented towards applying knowledge discovery methods on higher education database in the Management Information Systems (MIS) of Fayoum University by applying a data mining model on the university data. In this research, classification techniques are used to judge the student's grades. As there are many approaches that are used for data classification, therefore, in this paper, we applied one of the most suited techniques for the nature of data. We extracted the knowledge that describes the students' performance which supports deciding the right direction for the student according to his grades. .

Index Terms— Educational Systems; Data Mining; Classification; Students Satisfaction; Students performance; Genetic Algorithm, Knowledge Discovery

1 INTRODUCTION

The performance of the students is the most important target in the education system. It is considered the basic indication for the institutions' quality. This quality is basically depends on the students' achievements [1]. Many researchers have defined the performance of students in different perspectives, The performance of students is measured by the teaching assessment, but most of the researches identified the performance by the students' grades [2].

The grades of the students depends on many criteria including the exams' marks, students' activities, and the course structure. [3]. Therefore, evaluating the students is a vital phase in the education process which targets directly the core observation for the performance of the students. It is essential to highlight that the performance of the students affects mainly the education institution strategic plan [4].

The field of data mining is one of the successful fields in business recommendations [5] specially studying the performance of the students. Different data mining tasks could be applied not only for measuring the performance [6] but for other related targets such as predicting the future performance and classification of the students as well [7]. Focusing on data mining, different machine learning approaches are successfully proposed for these tasks such as Support Vector Machine, Naïve Bayes, Genetic Algorithms, and others [8]. Predicting the performance of the students is effective step which can change the educational map in the institutions which consequently change the teaching perspective [9].

The main purpose for focusing on predicting the performance of the students is to predict vital aspects for the students such as his estimated score range, and most important, his suitable specialty. A success in this target can provide a successful advising task in the educational institutions. Successful advising can have a successful step for directing the student to his best direction such as the suitable department for specialty and the suitable subjects to study. This support reduce the risk for the student performance. This whole view provide a positive impact for different stakeholders in the educational system including the students, the family, the supervisors, the instructors, and the high management team.

The amount of data is one of the main criteria in selecting the appropriate mining approach [8]. In [9], an approach was proposed to deal with a large amount of data which can be identified as big data. Other research in [10] focused on the number of attributes rather than the amount of data. Different classification approaches have been successfully applied in different environments such as music [11] and finance [12]. Although there has been a close focus on mining educational data, however, there has been many obstacles for reaching high accurate results. This is because the lack of required data for the research, the long required time for the experiments and the extensive steps required for data pre-processing [13]. Therefore, applying classification algorithms on real data can become an effective step for the required target in determining the students' performance and consequently predicting the required students' information for raining this performance [14].

Different data resources are used in mining of educational data such as Learning Management Systems (LMS), students' affairs, and pre-university data. The high dimensionality of

size of this data forced the need of intelligent approaches to find novel techniques for exploring data and extracting the required information to enhance the educational level through enhancing the students' performance [15]. This requirement has arisen due to the need for discovering intelligent trends to raise the educational process performance in general and the students which consequently prove the effectiveness of the educational process [16]

This research focuses on the education system targeting to enhance the performance of the students based on directing the students for their suitable department as well as suitable courses. The research applied data mining approaches on Fayoum University students' data. The results of the experiment succeeded to provide a satisfying prediction to the students. The remaining of the research includes the discussion of the literature review, describing the research approach, illustrating the experimental results, and then finally providing the research conclusion and the future work.

2 Previous Works

Many researchers have presented different approaches for predicting the performance of organizations in general [17] and the performance of the students in specific. This section presents some of these researches. [18] proposed an approach for revealing the hidden relation among attributes. The selected features included the grades, place of residence, and place of school. Another research by [19] presented two case studies which targeted to solve the problem of unequal level of students in the same class by applying support vector machine approach. The research claim that the results were satisfactory. Another research in [20] applied two different techniques, they are FP growth and K-mean algorithms

FP Tree and K-means clustering technique is applied in [6] for finding the similarity between urban and rural students programming skills. FP Tree mining is applied to sieve the patterns from the dataset. K-means clustering is used to determine the programming skills of the students. The study clearly indicates that the rural and the urban students differ in their programming skills. The huge proportions of urban students are good in programming skill compared to rural students. It divulges that academicians provide extra training to urban students in the programming subject.

[21] Attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a fourclass dataset.

The research by [21] applied a research in Portugal including 29 attributes for high schools' students. The research focused on two courses and applied four of the most traditional data mining algorithms. The research revealed that decision tree algorithms provide 93% of accuracy.

A research, by [9] presented a review of different data mining algorithms which is applied in the students' data and proved its applicability for this field. Different researches in [22], and [23] applied K-nearest neighbor algorithm over students' data and reached an accuracy from 57% to 62.9%. Naïve Bayes algorithm is also applied which revealed to be simple and have higher speed while it has a shortcoming to have unrelated factors [24] SVM is also proposed in [25] and [26] which achieved stable result with accuracy 86.3%. Moreover, Decision Tree algorithm is also applied in [27] with 85.9% accuracy. The same algorithm is applied in [28] to predict if the student will pass the course or not and achieved a satisfied accuracy

2.1 Review Stage

3 Proposed Methodology

Referring to the presented literature review and the analysis of their results with respect to the proposed approach, it is clear that the performance of the students can be affected by different factors as well as predicted by a set of influencing attributes. In the proposed approach, we focus on a set of attributes targeting to predict the performance of students in addition to predicting the most suitable studying direction for these students.
The primary goal for the proposed approach is to reveal a suitable approach for the required target using the existing available parameters. Therefore, different predicting algorithms are applied on the available data to reach the most suitable direction. Figure 1 illustrates the main steps for the proposed approach while the following sections discuss each phase in details

3.1 Collecting Data

This phase is considered with collecting the required data. Considering the target of the proposed approach, the data includes three categories, they are Personal Data, Tanseeq Data, and University Data. The attributes for each data sources depends on the availability from the sources, however, main attributes will be highlighted. Main attributes should be included in the data sources with adding the additional available attributes that could enrich the source targeting more accurate results.

3.2 Data Preparation

Although data preparation is one of the major basic steps [29], however, it is not highlighted in most of the research. Real data usually have incomplete values due to the inaccurate gathering of the data as well as the inconsistent data among the records. Therefore, a mandatory step for predicting the missing fields is applied in this approach. It is also mandatory to eliminate the inconsistent values and predict the most near and accurate ones. Another valuable step in the data preparation phase is to eliminate unrelated features. This step is considered with removing the features that do not represent the required task.

In this phase, the following steps are applied.

1. Define the missing data cells and records.

We can deal with the missing data by one or more of the following methods:

a) Eliminate the record which has the missing value

b) Predict the missing value by using a suitable method either by using one of the statistical methods such as the mean, the average, etc. or by using a prediction algorithm. A successful algorithm for this task is Bayes algorithm or decision tree.

2. Discover the inconsistent data

One method for discovering the inconsistent data is to apply a clustering algorithm which clearly highlight the outlier cluster. Dealing with outliers should not only be with directly removing them, therefore, in this approach, the first preparation step is also applied to minimize the inconsistent data targeting to minimize the amount of data which will be removed as an inconsistent data. This task is performed with a belief that each record in the provided data worth to be included, therefore, in this approach, we eliminate as few records as possible.

3. Predict the replaceable values for both situation

This step is a crucial part in the data preparation phase, therefore, selecting the suitable algorithm should be highly considered to predict the replaceable values for the selected inconsistent or null ones. High accurate values is required to avoid any degradation in the system performance. From many literatures, defined algorithms are successful in predicting the numerical data [5, 8] while others are successful in predicting the nominal data [30], [7, 15] according to the nature of these algorithms. Other approaches could be applied by provide the higher general value or the more specific value.

4. Eliminate unrelated features

Naturally not all features are required for the task, therefore, this step requires defining the required features and eliminate the unrelated ones. The determined features to be eliminated depends on the stakeholders' view who are responsible for maintaining the performance of the students. Developing a systematic research for revealing the suitability of the existing methods with the required parameters in the current study is performed in this step to avoid eliminating required features and support selecting the most influencing ones.

5. Reduce the required data

One of the data preparation phases is to reduce the amount of data targeting to minimize the processing cost. However, in this approach, we did not apply this step as it is crucial to include all students' records with all variations and the applied environment did not need this step. Moreover, this phase can be applied if required for other applications using sampling techniques [31]

3.3 Data Integration

In this phase, a set of processes is applied to successfully integrate the different sources of data into a multidimensional form. The process includes applying business intelligence techniques to provide the analytical processes with the suitable data modeling [17]. Selecting business intelligence techniques in this phase ensures the applicability of providing an integrated model which is suitable for high performance with successful analysis, simulation, and prediction models. Business intelligence techniques ensures applying the required data mining task over the integrated multidimensional data for developing a high performed data mart which includes the required relations for the students' data that are gathered from the different data sources.

3.4 Students' Progress and Performance Prediction.

This phase is considered the core phase of the proposed approach. This phase considers applying the data mining task to provide the student with his best route for studying, best department that he should apply for and predict his performance in this department. The success in this stage is considered the main target of the presented research. In this phase, one of the most successful machine learning algorithm is applied to predict the required information.

In this phase, based on the study in [30, 5], Genetic Algorithm is applied on the students' data to reveal the most suitable direction of the students, the following is the genetic algorithm steps.

Genetic algorithm incorporate concepts of usual analysis. The final plan behind Genetic algorithm is that there is an ability to deduce more accurate solution by merging between the best solution and other offered solutions [32]. Genetic algorithm is essentially used to suggest the best alternative with the least required data.

In order to generate more than one solution in a determined problem, Genetic algorithm is applied based on a biological perspective in three main phases: first, representing the problem as chromosomes, the proposed plans are evaluated using fitness functions, the selected plan is based on the best form of the suggested life for the identified problem. Each individual plan is evaluated according to two main activities, they are survive and reproduce. New solutions are produced based on the restructuring of the provided attributes. These solutions are then evaluated against the current one to finalize the best solution. To summarize, in [32], a flow chart for genetic algorithm was suggested which is illustrated in fig. 1 While the main steps are represented in fig. 2

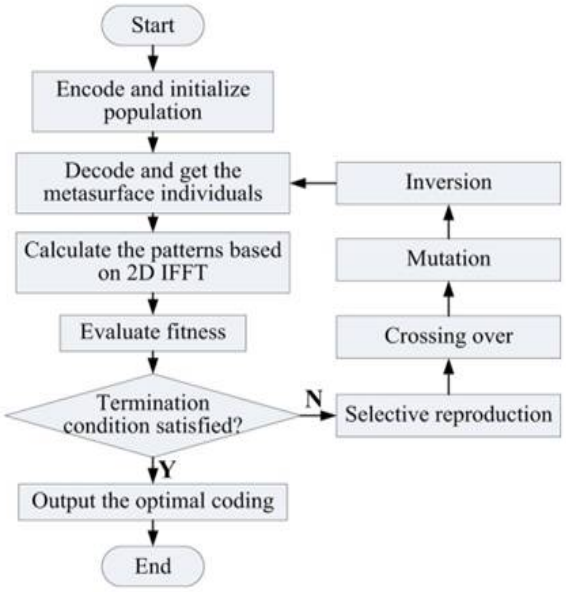

Fig. 1: A Flow Chart for Genetic Algorithm [33]

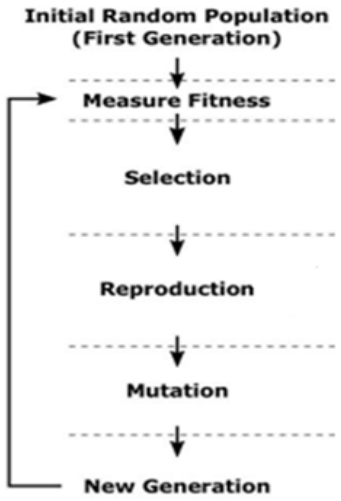

Fig. 2: The main Steps for Genetic Algorithm [33]

4 Experimental Study (Fayoum University)

The case study of the proposed approach is applied in Fayoum University. The data was collected from management information systems department which has all the students' data of the university. A total of 84483 records were collected for the students including a set of attributes. Table 1 presents the number of collected records for each college, while this diversity is represented in fig. 3.

Faculty	No of Records
Faculty of Specific Education	92579
Faculty of Computer Science	3039
Faculty of Medicine	9042
Faculty of Science	62823
Total	**84483**

TABLE 1: NUMBER OF COLLECTED RECORDS FOR EACH COLLEGE

As shown in table 1, the dataset had a variety of the students' background to ensure the applicability of the study. The aim of this research is to provide a guaranteed prediction to the students which provide them with their best path. Following the research approach, the gathered data has been analyzed and the records that required more processing are then determined, table 2 presents the number of processed records for completing data and guarantee consistent students' data. The attributes that are included in the data description are Student ID, Course, Degree, Grade, CGPA, Status, Level, College, Specialty, and Branch.

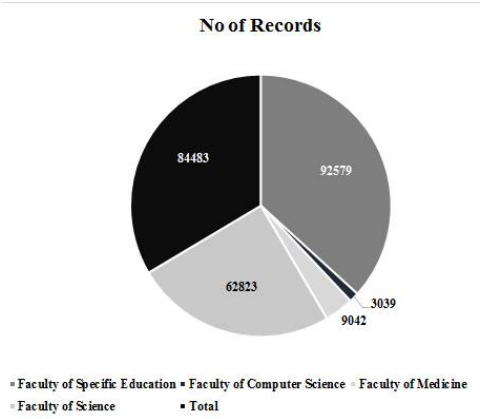

Fig. 3: Diversity of Collected Records for Each College]

Faculty	No of Records
Faculty of Specific Education	5032
Faculty of Computer Science	312
Faculty of Medicine	849
Faculty of Science	1598
Total	**7791**

TABLE 2: NUMBER OF UPDATED RECORDS

According to phase 2 in the proposed approach, inconsistent data are then determined, table 3 presents the number of records that are included in the experiment after removing records with inconsistent data which could not be replaced.

Faculty	No of Records
Faculty of Specific Education	92279
Faculty of Computer Science	3020

Faculty of Medicine	8760
Faculty of Science	62323
Total	**166382**

TABLE 3: NUMBER OF RECORDS INCLUDED IN THE EXPERIMENT

After setting up the main data which is included in the experiment, determining the required attributes for the experiment is then performed, and the following attributes' list are the determined set of attributes for the experiment.

Applying genetic algorithm is performed which revealed to successfully predicting the suitable department to the students' segment under investigation with an average accuracy of 97.29%. Table 4 presents the accuracy deviation of the colleges while this diversity is illustrated in fig. 4. The researchers have investigated the cause of the 2.7% that were unsuccessful in the prediction and the following reasons have been revealed.

Students' data deviation were not consistent in the departments of the colleges, variation in the attributes of the colleges were one of the reasons as the types of attributes varied hugely. However, these reasons are currently investigated along with different algorithms for raising the accuracy percentage.

Faculty	%
Faculty of Specific Education	97.32%
Faculty of Computer Science	97.8%
Faculty of Medicine	98%
Faculty of Science	96.02%
Average	**97.29**

TABLE 4 ACCURACY % FOR THE RESULTS IN EACH COLLEGE

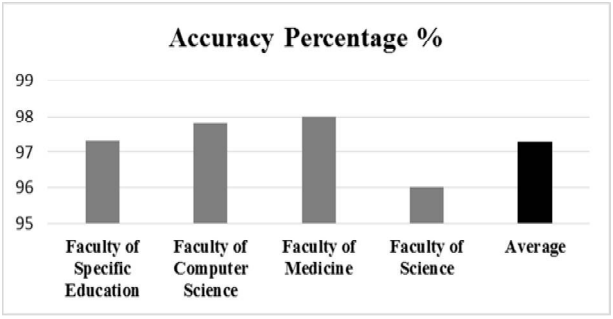

Fig. 4: Accuracy % for the results in each college

4 CONCLUSION

Recently, one of the most important research approaches is predicting the students' performance as well as the students' division which directly lead to raise their performance. This research aims at applying one of the most successful predicting algorithm on Fayoum University data in order to raise the performance of the students in higher education. The proposed approach was successful in recommending the suitable division to the students according to their skills. The accuracy percentage was 97.6%. The 2.4 % deviation has been focused on and the situation has been analyzed which revealed to a requirement for more accurate data and enlarging the attributes' set. One of the main future research is further research on recent data with enlarging the attributes' set and including more colleges in Fayoum University as well as other Egyptian universities to ensure the applicability of the proposed approach in different domains in Egypt.

REFERENCES

[1] M. of Education Malaysia, National higher education strategic plan (2015).

[2] U. bin Mat, N. Buniyamin, P. M. Arsad, R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention", in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, 2013.

[3] Z. Ibrahim, D. Rusli, "Predicting students academic performance: comparing artificial neural network, decision tree and linear regression", in: 21st Annual SAS Malaysia Forum, 5th September, 2007.

[4] Ayman E. Khedr, S. A. Kholeif and Shrouk H. Hessen, "Enhanced Cloud Computing Framework to Improve the Educational Process in Higher Education: A case study of Helwan University in Egypt", International Journal of Computers & Technology (IJCT), Volume 14, No. 6, pp. 5814-5823, April 2015.

[5] Ayman E. Khedr, Amira M. Idrees, Abd El-Fatah Hegazy, and Samir El-Shewy, " A proposed configurable approach for recommendation systems via data mining techniques", Enterprise Information Systems, Published online: 07 Mar 2017.

[6] Ayman E. Khedr, Ahmed I. El Seddawy, Amira M. Idrees, "Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), Vol2, Issue 6, 2014

[7] Aya M. Mostafa, Ayman E. Khedr, A. Abdo, "Advising Approach to Enhance Students' Performance Level in Higher Education Environments", Journal of Computer Sciences, Volume 13, Issue 5, 2017.

[8] Mahmoud Othman, Hesham Hassan, Ramadan Moawad, Amira M. Idrees, "Using NLP Approach for Opinion Types Classifier", Journal of Computers (JCP), Volume 11, September 2016.

[9] C. Romero, S. Ventura, "Educational data mining: A review of the state of the art", Trans. Sys. Man Cyber Part C 40 (6) (2010)

[9] D. M. D. Angeline, "Association rule generation for student performance analysis using apriori algorithm", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1 (1) (2013).

[10] RadhyaSahal, Mohamed H.Khafagy, Fatma A.Omara. "Exploiting coarse-grained reused-based opportunities in Big Data multi-query optimization", Journal of Computational Science, Volume 26, 2018.

[11]Scheuer, O., & McLaren, B. M., "Educational data mining. In Encyclopedia of the sciences of learning", Springer, 2012

[12] Costa, Y. M., Oliveira, L. S., & Silla, C. N., "An evaluation of convolutional neural networks for music classification using spectrograms". Applied Soft Computing, 52 (28), 2017.

[13]Nametala, C. A., Pimenta, A., Pereira, A., & Carrano, E. G., "An automated investment strategy using artificial neural networks and econometric predictors", In Proceedings of the xii brazilian symposium on information systems on brazilian symposium on information systems: Information systems in the cloud computing era-volume 1, 2016

[14] Schmidhuber, J., "Deep learning in neural networks: An overview", Neural networks, 61 (85), 2015

[15] Ayman E. Khedr, Amira M. Idrees, "Enhanced E-Learning System for E-Courses Based On Cloud Computing", Accepted in Journal of Computers (JCP), Volume 12, issue 1, 2017.

[16] Ayman E. Khedr, Sherief Kholeif, and Shrouk H. Hessen "Adoption of cloud computing framework in higher education to enhance educational process" International Journal of Innovative Research in Computer Science and Technology (IJIRCST), Volume 3, Issue 3, pp. 150-156, March 2015

[17] Mostafa Medhat Nazier, Ayman E. Khedr, and Mohamed Haggag, "Business Intelligence and its Role to Enhance Corporate Performance Management", International Journal of Management & Information Technology (IJMIT), Volume 3, No. 3, pp. 8-15, May 2013

[18] M.Ramaswami and R.Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues Vol. 7, Issue 1, No. 1, January 2010.

[19] Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme, "Improving Academic Performance Prediction by Dealing with Class Imbalance", 2009 Ninth International Conference on Intelligent Systems Design and Applications, 2009.

[20] L.Arockiam, S.Charles, I.Carol, P.Bastin Thiyagaraj, S. Yosuva, V. Arulkumar, "Deriving Association between Urban and Rural Students Programming Skills", International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010.

[21] P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008.

[22] Stapel, M., Zheng, Z., & Pinkwart, N., "An ensemble method to predict student performance in an online math learning environment", In Proceedings of the 9th international conference on educational data mining, international educational data mining society, 2016.

[23] Kabakchieva, D., "Predicting student performance by using data mining methods for classification", Cybernetics and Information Technologies, 13 (1), 2013.

[24] Gupte, A., Joshi, S., Gadgul, P., Kadam, A., & Gupte, A., "Comparative study of classification algorithms used in sentiment analysis", International Journal of Computer Science and Information Technologies, 5 (5), 2014

[25] Meyer, D. (2015). Support vector machines the interface to libsvm in package e1071. 2014

[26] Murty, M., & Raghava, R., "Kernel-based SVM." In Support vector machines and perceptrons, 2016

[27] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics, Springer, 2014

[28] Jayaprakash, S. M., Moody, E. W., Laura, E. J., Regan, J. R., & Baron, J. D., "Early alert of academically at-risk students: An open source analytics initiative", Journal of Learning Analytics, 1 (1), 2014

[29] Kaseb, Mostafa & Khafagy, Mohamed & Ali, Ihab & M. Saad, ElSayed. "A Technique for Reducing Storage and Resources in Big Data Replication", Redundant Independent Files (RIF),

[30] Ahmed ElAzab, Amira M. Idrees, Mahmoud A. Mahmoud, Hesham Hefny, "Fake Accounts Detection in Twitter based on Minimum Weighted Feature set", ICDAR 2016: 18th International Conference on Document Analysis and Recognition, January 2016

[31] Hesham Ahmed Hassan, Amira Mohamed Idrees, "Sampling Technique Selection Framework for Knowledge Discovery", INFOS2010 : 2010 7th International Conference on Informatics and Systems, Faculty of Computers and Information, Cairo University, Cairo, Egypt, 2010

[32] Nikita Jain1, Vishal Srivastava, "Data Mining Techniques: A Survey Paper" IJRET: International Journal of Research in Engineering and Technology, 02 (11), Nov-2013

[33] Huanhuan Yang, , Xiangyu Cao, , Fan Yang, , Jun Gao, , Shenheng Xu, , Maokun Li, , Xibi Chen, , Yi Zhao, , Yuejun Zheng, & Sijia Li, "From A programmable metasurface with dynamic polarization, scattering and focusing control", Scientific Reports (6), (2016)