IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

School Students' Performance Predication Using Data Mining Classification

Hafez Mousa[1], Ashraf Maghari[2]

Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine[1, 2]

Abstract: Educational management information systems generate huge amounts of data which hide a very useful knowledge. The techniques and methods used to discover the knowledge from students data are known as Educational Data Mining (EDM). The main objective of EDM is to improve students and teachers performance. Many researchers analysed students' behaviour to obtain useful knowledge that can help educators in planning for improving students' performance. There are two approaches which can be used to discover knowledge; by statistical methods and by DM techniques such as classification. This paper proposes a students' performance prediction model based on DM classification algorithms (Naïve Bayes, Decision Tree and K-NN). The dataset was collected from a preparatory male schoolin Gaza strip, includes over 1100 records. Obtained results show that Decision Tree gives the best results. Moreover, the results indicates that social case has little impact on the students' performance, while the academic features such as previous year and first term results have more impacts on the performance. These results can be used in improving students' performance by predication their retention early to minimize students' failure.

Keywords: Data Mining DM, Classification, Educational Data Mining EDM, Students' Performance, Naïve Bayes, Decision Tree and K-NN.

I. INTRODUCTION

One of the serious challenges that face education fields is the academic failure, and the education institutes work hard to minimize the failure ratio. According to the Palestinian Ministry of Education, 0.93% of 1192808 students (i.e. over 11000 students) failed in scholastic year 2014/2015 [1]. In the preparatory male school under study, the ratio is high (153 failed of 1136 students, i.e. 13.5%), because this data was collected before the summer treatment program in United Nations Relief and Works Agency for Palestine Refugees in the Near East(UMRWA) schools in Gaza Strip. UNRWA is an international agency which serves five million of Palestinian refugees. UNRWA provides Education service for more than half million students [2]. Recently, UNRWA performed a treatment program during second term to minimize failure ration in the final results.

Early students' performance prediction is one of the effective solutions to minimize students' failure problem. Analysing the stored students' data can help in early predicting students' performance, but the large amount of available data does not allow to perform this task manually or by trivial tools. Thus, the task can be achieved by using Data Mining (DM) techniques [3].DM techniques and methods that used in mining a dataset from educational resources is known as Educational Data Mining (EDM).

Researches objectives in EDM field can be classified into academic or administrative objectives. The administrative oriented is related to resource management and vision of the institutes, while, the academic oriented is interested in the individual behaviour for students and teachers, such as predication and improving of students' performance [4].

Most of researchers, to the best of our knowledge, aimed to improve the students' performance using classification techniques applied to datasets that came from e-learning environment, higher education or academic database.

In this paper, we aim to construct a students' performance prediction model by mining a real students' dataset collected from traditional school educational environment (UNRWA preparatory male school). To achieve this task we compare different type of DM classification techniques (Naïve Bayes, Decision Tree and K-NN), and determine which one is the most appropriate to early predication of students' performance. Our study results can be used to help the educators in their planning to support students with expected low performance.

The rest of the paper arranges as follows: Section 2 introduces a background about EDM and classification for predication. Section 3 summaries the related works. Section 3 provides details of used methodology. Section 4 explains and discusses the results. Finally, the paper is concluded in Section 4.

II. BACKGROUND

DM is very famous field in computer science, which can be defined as "the process of discovering valid, novel, potentially useful, understandable, and actionable patterns in data" [5].DM techniques have been used in many fields

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

such as Economic, Medical, Fraud detection, Engineering, and recently in Education which known as Educational Data Mining (EDM) [6].
EDM is a sub-field in DM, which deals with data that comes from different educational environments, and aims to achieve some educational objectives [7].EDM is defined by Baker as "the Data Mining methods that used to explore and understand data that come from educational environments, and using those methods to better understand students, and the environment of learning, which help in improving students' performance" [4].

Classification is one of the popular DM processes which are used in generating models that classify the dataset cases according to a class (label). The classification sub-process are: training classifier on 70% of dataset –as a popular percentage- to generate the mode, then apply the generated model on a testing dataset (30%) [8]. For EDM, Classification is used for predicting students' performance.

DM Classification includes some classification algorithms such as:

- Naïve Bayes (NB) Classifier, which is a popular classifier to predict data, usually used with independent attributes datasets. NB algorithm need less computation than other algorithms, so it is more quickly but less accuracy. NB classifier requires a small amount of data for training.
- Decision Tree (DT) Classifier, which was known as ID3 (Iterative Dichotomized), then improved into C4.5, based on a greedy approach. DT used as a predictive model by generation of a classification tree, so it is easy to interpret [9].
- K-Nearest Neighbor (K-NN) Classifier, which based on comparing test cases with training cases. K-NN one of the most easy in machine learning algorithms.

III.RELATED WORK

In past efforts, many researchers have used EDM techniques. Their researches can be mainly classified into:

A. Researches focused on EDM techniques and application:
Some researchers are interested in the DM techniques and application which are appropriate to the educational process and data to improve the student's performance and facilitate using DM techniques by educators [7, 8]. Oother researchers compared the DM techniques with other mathematical and statistical methods [10], while others compared between usage and accuracy of DM algorithms using educational datasets [11, 12].

B. Researches focused on the Academic objectives:
Most of EDM researches focused on academic objective directly or indirectly, so we classify the adjectives into sub-domains:

- Students' performance predication:

By early predication of students' performance, the educators can work to deny the failure and drop out problems [13-18].

- Improving students' performance:

Under this objective, we reviewed many researches: some researchers applied single EDM technique using decision tree as a classification methods to predict the students' final marks to reduce failure ratio and did the appropriate treatment at right time [18]. Other researchers used Naïve Bayes and Decision Tree techniques to analyse students' academic behaviour to increase student retention [20]. And other applied collection of EDM technique to obtain different rules and knowledge [21-22].

In this paper, we interest in students' performance predication using EDM techniques (classification). Amrieh, et al [18] proposed a new student's performance prediction model based on DM techniques with new data called student's behavioural features. This mode was created by set of DM classifiers (Artificial Neural Network, Naïve Bayes, and Decision Tree). In addition they applied ensemble methods (Bagging, Boosting, and Random Forest) to improve the performance of classifiers.

Acharya, et al [19] applied collection of Machine Learning algorithms (Decision Trees (DT), Bayesian Networks (BN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM)), on a set of attributes for Computer Science students in Kolkata. The obtained results show that DT (C4.5) and SVM do the best.
Abu Saa [23] used multiple data mining tasks (Naïve Bayes and different algorithms of Decision Tree) to create qualitative predictive models which were efficiently and effectively able to predict the students' achievement from a collected dataset consists of academic and social data. Researcher obtained the best accuracy 40% with CART Decision Tree.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

By studying the previous efforts, we did not found researches –to the best of our knowledge- which study the accuracy of predication model for students' performance using Naïve Bayes, Decision Tree and K-NN classifiers on a dataset for school students (traditional education school environment). There for this paper studies which classifier of (NB, DT and K-NN) generates the best predication model for student's performance using students' dataset from traditional school environment. Furthermore, the paper discovers the impact of various features on students' performance.

IV. METHODOLOGY

This section describes the methodology which includes the phases used for the classification process. The phases are shown in Fig 1 and discussed hereafter.

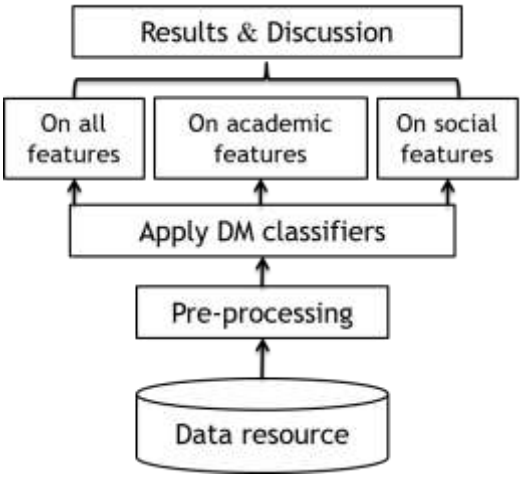

Fig. 1. Methodology phases

A. Phase 1: Business Understanding

The main objective of this paper is to develop a predication model for students' performance by using DM classification, and as a sub-objective to determine which classifier performs better with the collected educational data set.

B. Phase 2: Data Collection

To construct the prediction model, the data was collected for 1036 students as shown in Table 1. The data has been collected from a preparatory school (7th, 8th, 9th grade) in scholastic year (2014-2015) by the Education Management Information System(EMIS)which is used in UNRWA schools as information system.

TABLE 1. DATASET ATTRIBUTES

Attributes	Type	Description
Level	Cat.	7th, 8th, 9th
Orphan	Cat.	no, father, mother, both
SHC	Cat.	yes, no
BirthYear	Number	
Camp	Cat.	yes, no
FatherWork	Cat.	yes, no
FailYears	Number	
PrevYear	Cat.	success, fail
FirstTerm	Cat.	excellent, very good, good, fair,poor, very poor
FinalResult	Cat.	success, fail

To understand the collected data and the relations between attributes we obtained some explanations from users of EMIS and teachers in the school. Some graphs such as Fig 2shows the ratio of failure in the collected dataset (153students failed of 1136). The ratio is high because the dataset was collected before summer treatment program that performed in summer holiday.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

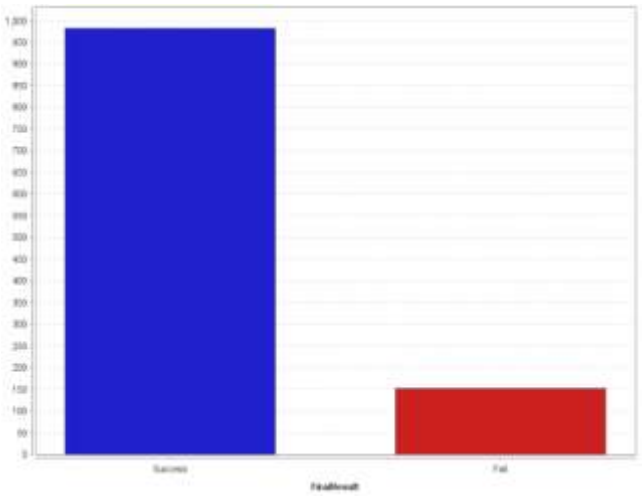

Fig. 2. Ratio of failure to success

C. Phase 3: Data Pre-processing

The data need some of pre-processing to be prepared as input for EDM techniques. For example, the data are needed to be integrated from multiple tables (marks details for term1, 2, students' information and social cases) in one table. Some data need discretization from numerical into categorical such as student results in the first term [Excellent, V. good, Good, Fair, Poor, V. poor] and final result [Success, Fail].

D. Phase 4: Applying DM Classification Algorithms

We performed our experiments by using Naïve Bayes, Decision Tree C4.5 and K-NN classifiers. These three classifiers have been selected because of our dataset which is in medium size, includes (1036 records) and (9 features), and almost polynomial attributes.

RapidMiner was used to conduct the experiments. It is a famous open source DM tool based on Java.

Split-Validator is used to split dataset into 70% for training to generate model, and 30% for testing to apply model.

We ran the experiment on a PC with 4GB RAM, (2.20GB * 5 cores).

E. Phase 5: Results Evaluation

In addition to the accuracy measure, we used other evaluation techniques and methods to evaluate the results obtained from Phase 4 against to the project objectives. The results and their evaluations are included in the next section (Results and Discussion) of this paper, where we used Precision and Recall measures.

V. RESULTS & DISCUSSION

In this section, we explain the obtained results of our applying EDM techniques.

A. Classification results using all features:

Firstly we applied both of NB and DT classifiers on all of data features (academic and social features), and the results were shown in Table 2.

TABLE 2. CLASSIFICATION RESULTS BY USING ALL ATTRIBUTES

Evaluation Measure	Naïve Bayes (NB)	Decision Tree (DT)	K-NN
Accuracy	91.79	92.96	88.86
Recall	63.04	97.83	63.04
Precision	72.50	66.18	58.00
F-Measure	67.44	78.95	60.42

B. Classification academic features:

In the second experiment we selected the academic features (FailYears, PrevYear, FirstTerm), and the results were demonstrated in Table 3.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

TABLE 3. CLASSIFICATION RESULTS BY USING ACADEMIC ATTRIBUTES

Evaluation Measure	Naïve Bayes (NB)	Decision Tree (DT)	K-NN
Accuracy	91.50	92.96	90.91
Recall	58.70	97.83	50.00
Precision	72.97	66.18	74.19
F-Measure	65.01	78.96	59.74

From results in Tables 2 and 3, we notice that DT classifier has the highest accuracy values. This high accuracy can be explained by the methodology of DT in generation a tree and pruning technique, and medium size of used dataset.
Our results are similar the results obtained byAmrieh et al.[18]where they compared between (DT, NB and ANN) on students' dataset that contains academic and behavior features to get that (75.9 DT, 67.1 NB). Moreover,Acharya and Sinha[19] applied five ML algorithms include DT and NB, on students' dataset of computer science course college. Their results show that DT has the highest accuracy of 83.0 [19] which are consistent with our results.

C. Classification results using social features:
In the third experiment we applied both of NB and DT classifiers on data with social features (Orphan, SHC, BirthYear, Camp, FatherWork, FailYears). The results are shown in Table 4.

TABLE 4. CLASSIFICATION RESULTS BY USING SOCIAL ATTRIBUTES

Evaluation Measure	Naïve Bayes (NB)	Decision Tree (DT)	K-NN
Accuracy	86.80	87.10	82.99
Recall	17.39	13.04	34.78
Precision	53.33	60.00	36.36
F-Measure	26.23	21.43	35.56

From Table 4, we notice that K-NN classifier is suitable for the data with few changes, because it just compare training with testingcases. In addition, we found that the social case has small impact on the students' performance; this can be explained by the small number of these cases in the dataset, and the students' environment.

VI. CONCLUSION

This paper has presented mining of real dataset for students, collected from preparatory male school in 2014/2015 scholastic year, using DM classification techniques to predict the performance of students. We applied three classifiers (Naïve Bayes, Decision Tree and K-NN) and found that DT classifier gives the best results when used with students' data (social an academic attributes). In addition, we found that the social case has little impact on the students' performance, while the great impact come from the academic features such as previous year and first term results.

The results may help the educators to minimize the failure ratio, by determining students that may fail, and treat them early.
For future work, we may use more EDM classification classifiers, and additional data about students' social case and academic situation.

REFERENCES

[1] http://www.moehe.gov.ps "Education Statistical Yearbook For Scholastic year 2015-2016".
[2] https://www.unrwa.org "United Nations Relief and Works Agency for Palestine Refugees (UNRWA)".
[3] Shahiri, A. M., & Husain, W. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414-422.
[4] Baker, R., Data mining for education. International encyclopedia of education, 2010. 7: p. 112-118.
[5] Fayyad, U.M., et al., Advances in knowledge discovery and data mining. 1996.
[6] Yadav, S.K. and S. Pal, Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832, 2012.
[7] Jacob, J., et al. Educational Data Mining techniques and their applications. in Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on. 2015. IEEE.
[8] Ahmed, A.B.E.D. and I.S. Elaraby, Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2014. 2(2): p. 43-47.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

[9] Rokach, L., & Maimon, O. (2014). Data mining with decision trees: theory and applications. World scientific.
[10] Sukhija, K., M. Jindal, and N. Aggarwal. The recent state of educational data mining: A survey and future visions. in MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on. 2015. IEEE.
[11] Prabha, S.L. and D.A.M. Shanavas, Educational data mining applications. Operations Research and Applications: An International Journal (ORAJ), 2014. 1(1).
[12] Ramaswami, M. and R. Bhaskaran, A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv:1002.1144, 2010.
[13] Ramesh, V., P. Parkavi, and K. Ramar, Predicting student performance: a statistical and data mining approach. International journal of computer applications, 2013. 63(8).
[14] Osmanbegović, E. and M. Suljić, Data mining approach for predicting student performance. Economic Review, 2012. 10(1).
[15] Romero, C., et al. Data Mining Algorithms to Classify Students. in EDM. 2008.
[16] Dekker, G.W., M. Pechenizkiy, and J.M. Vleeshouwers, Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining, 2009.
[17] Marquez-Vera, C., C. Romero, and S. Ventura. Predicting School Failure Using Data Mining. in EDM. 2011.
[18] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.
[19] Acharya, A., & Sinha, D. (2014). Early Prediction of Students Performance using Machine Learning Techniques. International Journal of Computer Applications, 107(1).
[20] Nasiri, M., B. Minaei, and F. Vafaei. Predicting GPA and academic dismissal in LMS using educational data mining: A case mining. in E-Learning and E-Teaching (ICELET), 2012 Third International Conference on. 2012. IEEE.
[21] El-Halees, A., Mining students data to analyze e-Learning behavior: A Case Study. Department of Computer Science, Islamic University of Gaza PO Box, 2009. 108.
[22] Baradwaj, B.K. and S. Pal, Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417, 2012.
[23] Abu Saa, A. (2016). Educational Data Mining & Students' Performance Prediction. International Journal Of Advanced Computer Science And Applications, 7(5), 212-220.