Early Warning System for Academic using Data Mining

Leo Willyanto Santoso
Informatics Department
Petra Christian University
Surabaya, Indonesia
leow@petra.ac.id

***Abstract*—Nowadays, student academic data in universities are very huge. However, the opportunity to manage the data is a knowledge that cannot be overlooked. Educational data mining is a current research field which uses data mining algorithms to transform large volumes of academic data into valuable knowledge capable of improving the educational processes and decisions. This research makes use of a set of three models. The first two models used the data obtained in the first year (first semester and second semester), to predict the academic success of the enrolled students, while the third model used the information available at the end of the first year to predict the academic performances of the students at the end of their study. At the same time, this work also intends to identify the factors that are most critical to these models. The results of this research paved way for the head of the school to identify students in need of more pedagogical support, as well as students with high probability of excelling in their studies. It could also allow them to focus their attention on the critical aspects, by implementing mechanisms that tackles students' difficulties.**

Keywords—educational, data mining, student academic

I. INTRODUCTION

Education plays an essential role for developing countries. It is the key to eradicating poverty, changing people, communities, and nations. The information system is one of the most recent education technology adopted by institutions of higher learning. This is the reason why many higher institutions are investing a lot of money to improve their academic information system [1]. Data warehouse, is another very common technology that supports data analysis and reports for academic institutions [2].

The amount of information stored in these systems makes them valuable, thereby, leading to the improvement of quality learning. Identifying students with the inability to achieve academic success early, especially within the first year of their academic study, enables such students to be supported.

This research, is therefore, aimed at creating models that can accurately identify students unable to achieve academic excellence during their first year, as well as those capable to having long term academic performances at the end of their studies.

The objectives of this research are as follows: (1). The designed system can easily classify students into groups: those who excel in their studies and those who do not. The exploration of these students' groups, gives an insight on the factors that determines the students' performances. (2) The proposed decision support system provides solutions on how to improve the quality of education by using the cutting edge technology.

The rest of the paper was organized as follows. Section 2 describes the existing literatures on the topic and some related works. Section 3 describes the methodology utilized in this research work. Section 4 provides the experimental evaluation and analysis of the results. Finally, Section 5 is the conclusion of this research.

II. LITERATURE REVIEW

In this section we reviewed the existing literatures on the educational data mining and the academic successes in the institution.

A. Educational Data Mining (EDM)

EDM deals with the use of data mining techniques by, tapping into the data stored therein, in order to extract meaningful information that can support the decision-making processes by enabling a better understanding of the students and their learning environments [3]. EDM easily identifies those factors responsible for students to either graduate or not graduate [4]. However, the results are dependent on the selected dataset.

Student model is defined as the representation of a their characteristics, state, intelligent quotient, motivation, meta-cognition and behaviors [3]. This model allows the educational software systems to adapt to the responses of students. Baker et al. focused on identifying frustrated [5].

Personalized learning environments are systems flexible to students' characteristics. They are closely related to the recommendation systems, and allows students achieve their educational goals [6].

The resource management systems can be improved by integrating the Data Mining tools and all works in the EDM. The essence is to allow the average user to be able to make use of such tools [7]. Pedagogical support revolves around identifying the most effective type of support for a given situation and group of students. Beck and Mostow associated a student's performance to the type of pedagogical support received [8].

Educational theories deals with its empirical analysis and phenomenon. In order to enable a deeper comprehension of the key aspects of these theories, Gong et al stated that there was a relationship between an individuals' self-discipline and the number of mistakes made by that person [9].

Prediction is defined as the determination of the value of an unknown variable using the values of known variables. The known variables are called predictors. Problems associated with prediction can be either be classified as unknown variable belonging to several pre-established classes, or as a regression whose objective is to predict the value of a continuous numerical variable [10, 11]. There are

978-1-5386-7167-2/18/$31.00 ©2018 IEEE

several other algorithms predictions that can be used to predict the students' performances in their work, such as decision trees and Bayesian classifiers. [12, 13] detection of outliers in the EDM, is an inconsistent process. This helps to identify students who have slow learning processes and those students who are gifted. [14].

Students could be grouped according to their educational history and socio-demographic characteristics [15]. Clustering algorithms can sometimes be unclear, thereby, causing a variable to belong to more than one algorithm.

Relationship Mining is the identification of relationships between variables in large data sets [16]. This technique can identify the effective factors on responsible for a student to retake a course [17, 18].

B. Data Mining Technique in Education

Different data mining techniques can be used to support the extraction of relevant information, although this could be dependent on the aim of the study. According to the objectives of this research, our analyses were focused on the classification and clustering techniques. In both cases, there was no consensus regarding the most appropriate technique to be used. According to the classification techniques, the performances of several algorithms were compared and the one with the best performance was found to be the IB1, the Nearest Neighbour algorithm [19]. On the other hand, Decision Trees and Bayesian Networks were used to predict the students' GPA at different points of their academic paths [12]. The Decision Trees consistently outperformed the Bayesian Networks in this task. Different algorithms were compared, and the Random Forest and Support Vector Machines were proven to have the best performances, with Naive Bayesian Networks generally surpassing the Decision Trees most especially when dealing with this imbalanced or discrete datasets [20].

There is a relationship between the classification problems and the set of predictors that produce the most accurate model. Oskuei and Askari focused on the gains of the performances by using sex, parents' level of education and welfare [21]. Pal and Saurabh Pal utilize a broader range of attributes, with certain aspects such as the admission type and the locations of both the students' residence and the college taken into consideration [19]. Meanwhile, Asif, Merceron and Pathan focused on the grades students obtained in certain courses [22].

III. Design and Methodologies

The purpose of this research was to develop three different prediction models that address the student's academic performances. The first model deals with predicting the academic success in their first semester with the data available at the time of enrollment, while the second model uses the data available at the end of their first semester in order to predict his/her success at the end of the first academic year. The success of these models were defined according to the students who were able to complete at least 20 credits in a semester with a GPA of 3 or 4. For a student to get a total of 20 credits, he/she must attend lectures for a total of 20 hours every week. For the third model, a multinomial classifier was preferred. This model focuses on predicting the overall academic success of the student based on the data available at the end of the first year, hence knowledge of the degree of success was considered important. For measuring this performance, the following formula was used:

$$Performance = \frac{\sum(G * C)}{\sum CE} \qquad (1)$$

G represents the final grades the student obtained on their courses, C the number of credits those courses are worth and CE the total number of credits that they enrolled in. This allows us to take into account number of courses they passed and the grades they obtained in each of the courses.

The performance values gotten from the dataset were then submitted to a K-Means clustering algorithm, and this allowed us to congregate them into five groups. By analyzing these groups, we were able to define the performance value for five students. According to the information that was used to develop these models, the attributes consisted of academic information, such as enrollment grades, national exams taken, amount of credits completed on a given semester and the averages obtained, as well as demographic data, like a student's sex, their parents' education level and jobs, and whether he or she is a beneficiary of a scholarship.

In this research, four algorithms were chosen and used to build the three models, namely C4.5, Random Forest, Naive Bayes and Support Vector Machine. These algorithms were chosen due to their widespread use in data mining.

The algorithms themselves were encapsulated in a process that applied x-fold validation with 10-folds, meaning that the data was divided into 10 blocks; the model trained 9 of the blocks and carried out evaluation with the other one. The process was repeated 10 times, once for each of the different blocks. In the end, the average performance was used. This reduces the impact the selection of data for training and test sets has on the performance of the model.

For the performance itself, three different measures were used in the first two models: accuracy, area under curve (AUC) and specificity. in the first year, priority was given to identifying cases of unsuccessful students rather than successful ones, as these are the ones that needs the institutions support, specificity is more relevant in those cases than sensitivity, which prioritizes the identification of successful examples. The third model, deals with accuracy, sensitivity and precision.

IV. Discussion and Analysis

In order to understand the results obtained, it is important to understand the parameters used for them. In this regard, Random Forest used 100 trees with a K equal to approximately the square root of the total number of attributes, while C4.5 was developed as an unprunned tree.

For the SVM, a complexity constant of 1 was used, with a tolerance parameter of 0.001. Logistic models were also fitted into the SVM outputs, and this allowed for the calculation of the relevant AUC values. The Weka's implementation of the SMO changed the output values to the extremes without making use of this parameter. These parameters were used in all the three developed models.

The first model that was developed used the information available at the start of the student's first year to predict the

students' success at the end of their first semester. The features used in determining the model were: parent's education background , parent's job, enrollment option, enrollment stage, degree that the student enrolled in, high school average grade, enrollment exams average grade, enrollment average grade and enrollment exams.

As aforementioned, the data used for these models also underwent outlier detection and removal, with 2459 instances left. It's important to note, however, that this dataset was then subjected to oversampling, resulting in a total of 3145 instances.

The performance of the four algorithms for the 1st model can be seen in Table I.

TABLE I. PERFORMANCE OF FIRST MODEL

Algorithm	Accuracy	Specificity	AUC
RandomForest (I=100, K=4)	80.93%	94.01%	94.30%
J48	83.90%	85.60%	88.74%
Naive Bayes	78.12%	79.67%	85.23%
SVM	83.70%	88.11%	90.12%

From Table I above, we can see that the Random Forest and the SVM produces the best AUC, with the Random Forest having a very poor accuracy. Meanwhile, J48 has the highest accuracy and provides us with all the information regarding the attributes with the higher academic success. AUC curve for the 1st model can be seen in Figure 1.

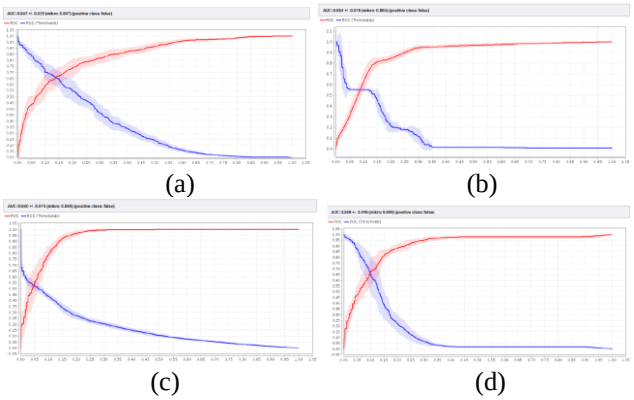

Fig. 1. AUC Curves for First Model . (a). Naive Bayes (b) J48 (c) RandomForest (d) SVM

The second model was developed using the information available at the end of the first semester of the student's first year. The attributes that remained in the model were: the parent's education background, parent's job, enrollment stage, degree that the student enrolled in, high school average grade, enrollment exams average grade, enrollment average grade, enrollment exams, number of college exams for approval, number of college exams for grade improvement, average grade on the first semester, and number of credits completed on the first semester.

Furthermore, the exception to the above list lies within the Random Forest algorithm, with the following attributes: school type, marital status, enrollment year, enrollment option and sex.

This data also underwent detection and removal, with 15 outliers removed, leaving us with a total of 2454 instances. After the resulting dataset underwent oversampling, we were left with a total of 3136 instances.

The performance of the four algorithms for the 2nd model can be seen in Table II. Moreover, AUC curve for the 2nd model can be seen in Figure 2.

TABLE II. PERFORMANCE OF SECOND MODEL

Algorithm	Accuracy	Specificity	AUC
RandomForest (I=100, K=5)	90.11%	94.39%	96.90%
J48	85.40%	94.19%	95.17%
Naive Bayes	85.12%	83.27%	89.57%
SVM	89.10%	91.03%	95.62%

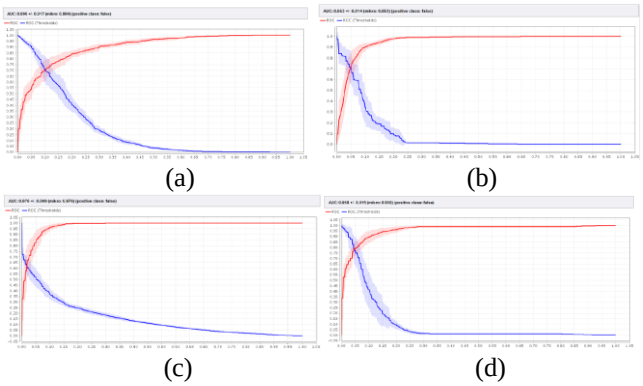

Fig. 2. AUC Curves for Second Model . (a). Naive Bayes (b) J48 (c) RandomForest (d) SVM

In this model, Random Forest again produces the top results, while Naive Bayes remains the worst performing algorithm.

Lastly, the third model that was developed used the information available at the end of the student's first year to predict their success at the end of their degree year. This model also underwent feature selection, with the resulting attributes being: degree that the student enrolled in, high school average grade, enrollment average grade, enrollment exams, number of college exams for approval, average grade on the first semester, number of credits completed on the first semester, average grade on the second semester and number of credits completed on the second semester.

Outlier detection and removal were also performed on this model, with a total of 2459 left. This result is contrary to what was obtained in the previous two models.

Once again, the performance of the four algorithms for the 3rd model can be seen in Table III.

TABLE III. PERFORMANCE OF THIRD MODEL

Algorithm	Accuracy	Specificity	AUC
RandomForest (I=100, K=5)	96.41%	95.69%	96.80%
J48	91.60%	91.44%	91.72%
Naive Bayes	75.12%	73.71%	73.87%
SVM	92.10%	91.03%	91.82%

Random Forest continues to produce some of the best results out of the four algorithms, with a very high performance. J48 once again provides us with information regarding the attributes which have a higher weight in the academic success of students.

V. CONCLUSIONS

In order to reduce the time required to complete a degree, universities needs to be able to identify successful and unsuccessful students at the beginning of their academic career. For this purpose, three prediction models were developed: one that predicts success in the first semester with the data available at the time of enrollment, the second model predicts success in the second semester with the data that was available at the end of the first semester, and the third model predicted the overall academic success with the information available at the end of the first year.

With regards to all three models, the result of the work presented here is extremely positive. This analysis showed that the average entrance exams were attributed to the success of the students during their first semester. In the remaining two models, however, the information about the first and second semesters makes use of the average enrollment. Nonetheless, it is important to note that the information about the second semester can also replace that of the first semester, the entrance exams continue to have a big impact on all the three models. Due to lack of information, however, this analysis did not extend itself to which courses had the most predictive impact.

ACKNOWLEDGMENT

This research was supported by The Ministry of Research, Technology and Higher Education of the Republic of Indonesia. Research Grant Scheme (No: 002/SP2H/LT/K7/KM/2017).

REFERENCES

[1] L.W. Santoso and Yulia, "Analysis of the impact of information technology investments - a survey of Indonesian universities," ARPN JEAS., vol. 9, no. 12, pp. 2404-2410, Dec, 2014.

[2] L.W. Santoso and Yulia, "Data warehouse with big Data technology for higher education," Procedia Computer Science, vol. 124, no. 1, pp. 93-99, 2017.

[3] R.S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," JEDM-Journal of Educational Data Mining, 1(1), pp. 3–17, 2009.

[4] P. Strecht, JM Moreira, and C. Soares, "Educational data mining: preliminary results at university of porto", 2014.

[5] R.S. Baker, A.T. Corbett, and A.Z. Wagner, "Human classification of low-fidelity replays of student actions," In Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems, pp. 29–36, 2006.

[6] R.A. Huebner, "A survey of educational data mining research," 2013.

[7] E. García, C. Romero, S. Ventura, and C. de Castro, "A collaborative educational association rule mining tool," The Internet and Higher Education, 14(2), pp. 77–88, 2011.

[8] J.E. Beck and J. Mostow, "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students," In Intelligent tutoring systems, pp. 353–362, 2008.

[9] Y. Gong, D. Rai, J.E. Beck, and N.T. Heffernan, "Does self-discipline impact students' knowledge and learning?," International Working Group on Educational Data Mining, 2009.

[10] S.B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques", 2007.

[11] L.W. Santoso and Yulia, "Predicting student performance using data mining," In the Proceedings of 5th International Conference on Communication and Computer Engineering (ICOCOE), 2018.

[12] N.T. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," In Frontiers in education conference - global engineering: Knowledge without borders, opportunities without passports, Oct 2007.

[13] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," Cybernetics and Information Technologies, 13(1), pp. 61–72, 2013.

[14] V. Hodge and J. Austin, "A survey of outlier detection methodologies," Artificial Intelligence Review, 22(2), pp. 85-126, 2004.

[15] J. Ranjan and K. Malik, "Effective educational process: a data-mining approach," Vine, 37(4), pp. 502–515, 2007.

[16] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," SIGMOD Rec., 22(2), pp. 207–216, June 1993.

[17] N. Hajizadeh and M. Ahmadzadeh, "Analysis of factors that affect students' academic performance-data mining approach," 2014.

[18] L.W. Santoso and Yulia, "The analysis of student performance using data mining," In the Proceedings of 3rd International Conference on Computer, Communication and Computational Sciences (IC4S), 2018.

[19] A.K. Pal and S. Pal, "Data mining techniques in EDM for predicting the performance of students," International Journal of Computer and Information Technology, 2013.

[20] V.T.N. Chau and N.H. Phung, "Imbalanced educational data classification: An effective approach with resampling and random forest," In Computing and communication technologies, research, innovation, and vision for the future (RIVF), 2013 IEEE RIVF International Conference on, pp. 135-140, Nov 2013.

[21] R.J. Oskouei and M. Askari, "Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies)," Computer Engineering and Applications Journal, 3(2), pp. 79–88, 2014.

[22] R. Asif, A. Merceron, and M.K. Pathan, "Predicting student performance at degree level: a case study," International Journal of Intelligent Systems and Applications, 2015.