Ermatita Paper 5 1570805890-Dafid

by Ermatita Paper 5

Submission date: 11-Oct-2022 02:37PM (UTC+0700)
Submission ID: 1922393080
File name: Paper1-1570805890-Dafid.pdf (227.97K)
Word count: 3588
Character count: 19830

2022 International Conference on Data Science and Its Applications (ICoDSA)

Determining Appropriate Classification Method Based on Influential Factors for Predicting Students' Academic Success

Dafid
Doctoral Program in Engineering Science; Information System
University of Sriwijaya; Universitas Multi Data Palembang
Palembang, Indonesia
03013622025009@student.unsri.ac.id; dafid@mdp.ac.id

Ermatita
Doctoral Program in Engineering Science
University of Sriwijaya
Palembang, Indonesia
ermatita@unsri.ac.id

***Abstract*— The need for accuracy in a prediction is a non-negotiable thing. One of the determinants of the accuracy of a prediction model is the classification method. Data mining offers various classification methods for predicting. Therefore, determining appropriate classification methods that produce high accuracy prediction model is a must. Several previous studies have shown excellent results based on influential factors for predicting students' academic success. However, the research only focuses on one influential factor category rather than a combination of multiple influential factor categories. It becomes a serious issue since there are influential factors on the dataset that not only have one influential factor category but mostly multiple factor categories. Therefore, the best classification method for a multiple influential factor category has not been known yet. This research analyzes the performance of classification methods based on multiple categories of influential factors. The result will help the researcher find the best combination of factor category and classification method should they used. Among multiple factor category and classification methods have been tested show combination of certain classification method give the best result for certain multiple factor category.**

Keywords*— *prediction, academic success, classification method, factor category, accuracy

I. INTRODUCTION

It can't be denied anymore that students' academic success has become the main concern for all higher education. The need to achieve that goal has enforced the higher education to anticipate the risk of failed student. One of stategy to minimize the risk of failed student is getting information of students' academic performance early by implementing prediction. The students' academic success depends on the influential factors that can be predict by using data mining. Classification method in data mining is one of most popular technique that has been used by several researchers [1]. For this reason, classification method is considered to be the best choice for predicting students' academic success. The influential factors by using classification method enable educationist to raise prediction model accurately. Accurate prediction model has been a challenging task due to the variety of influential factors involved. Generally, influential factors can be group into *Prior Academic Achievement, Student Demographics, Students' Environment, Psychological and Student E-learning Activity* [2]. Several researchers have made significant findings in the prediction of students' academic success [3–11]. Among the algorithms under classification used are Decision Tree (DT), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Naïve Bayes (NB) and Artificial Neural Networks (ANN).

Research done by Jishan [12] using Decision Tree algorithm found that the final cumulative grade point average (CGPA) factor achieve the highest accuracy (91%) than other factors. Meanwhile researchers Kumar [13] using Neural Network algorithm achieve the highest accuracy (98%) on Internal assessments and External assessment factors than others. Osmanbegovic and Suljic [14] have used Naive Bayes algorithms to estimate students' performance. Their research showed that Naive Bayes had achieved the highest accuracy (76%) on CGPA, Student Demographics, High school background, Scholarship, Social network interaction factors than others. Other researchers Mayilvaganan and Kapalnadevi [15] found that Internal assessment, CGPA, Extracurricular activities factor under kNN give a good accuracy (83%) than others. The last algorithm is Support Vector Machine applied by Sembiring [16] gives a better accuracy on Psychometric factors (83%) than others. According to the research result above, it can be concluded that each factors will achieve the highest accuracy if it meet the appropriate classification method. However, the research above only focus on one category factor rather than multiple category factors. No research has been conducted before based on multiple factor category comprehensively. In the fact, the influential factors that available on dataset consist of multiple category factor. Therefore, the multiple category factors still have problem influencing the accuracy of the prediction model. The problem is deciding the best classification method for certain multiple category factors. To overcome the problem, determining appropriate classification method based on certain multiple category factors is applied.

This research aims to get the highest accuracy prediction model for predicting students' academic success under the classification method based on certain multiple category factors. All of the prediction models that have been created are next tested in finding the best classification method using evaluation measurement which is Accuracy.

The next section of this paper is organized as follows: Section II gives the study of literature review. Section III explained the methodology of this research. Section IV discusses the research findings and introduces the study implications. Finally, Section V outlines the conclusion.

978-1-6654-8665-1/22/$31.00 ©2022 IEEE

II. LITERATURE REVIEW

A. Classification Method

Decision Tree method has ability to uncover small or large data struture. Its simpilicity and comprehensibility lead Decision Tree be the most popular technique for prediction. Decision tree transform the data table into a tree model, which determines the attributes from the roots, branches to the decision. The determination of attribute is calculated using gainratio. The value of information gain means how much information is obtained by knowing the value of an attribute while the split information value is used for an attribute that has multiple instances (more than two and varied). The formula is (1):

$$\text{GainRatio(S|A)} = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (1)$$

K-Nearest Neighbor (KNN) is a noise-sensitive classifier and an popular non-parametric classification method which have succesfully implemented in many classification problems. The KNN highly depends on the quality of the training data for its performance. The things affect the accuracy are the noise of data and mislabeled data, outliers, and overlaps regions between the data of different classes or targets lead.

Another popular technique for prediction is Artificial Neural Networks. ANN has the advantage of doing a complete detection in nonlinear relationship between dependent and independent variables. ANN also could detect all possible interaction between predictors variables[17].

Naïve Bayes is used for two class and multiclass classification problem[14]. The formula is (2):

$$\text{P}(h|\text{d}) = \frac{P(d|h)*P(h)}{P(d)} \quad (2)$$

Where:

- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis)

Select the hypothesis with the highest probability after calculating the posterior probability to get the maximum probable hypothesis (MAP). The formula is (3):

$$\text{MAP}(h) = max\ \frac{(P(d|h)*P(h))}{P(d)} \quad (3)$$

P(d) is a normalizing term which allows us to calculate the probability. If there is an even number of instances in each class in the training data, then the probability of each class (e.g. P(h)) will be equal. Again, this would be a constant term in the equation and drop it so that end up with (4):

$$\text{MAP}(h) = \max(P(d|h)) \quad (4)$$

One of supervised learning method for classification is Support Vector Machine. It has advantage in ability to classify the data in small datasets. It also has a good generalization ability and faster than other methods.

B. Influential Factors

One of crucial component to predict students' academic success is the potential influential factors which are driving data to be collected and mined. Actually there are many factors that have been investigated by researchers with respect to their impact on the prediction of students' academic success. Then all of the influential factors can be categorized into *Prior Academic Achievement, Student Demographics, Students' Environment, Psychological and Student E-learning Activity* [2] as shown in Table I.

TABLE I. DATA CATEGORY

Factor Category	Factor Description
Prior Academic Achievement (PA)	Pre-university data: high school background (i.e., high school results),pre-admission data (e.g. admission test results) University-data: semester GPA or CGPA, individual course letter marks, and individual assessment grades
Student Demographics (SD)	Gender, age, race/ethnicity, socioeconomic status (i.e., parents ' education and occupation, place of residence / traveled distance, family size, and family income).
Students' Environment (SE)	Class type, semester duration, type of program
Psychological (PS)	Student interest, behavior of study, stress, anxiety, time of preoccupation, self-regulation, and motivation.
Student E-learning Activity (EL)	Number of logins times, number of tasks, number of tests, assessment activities, number of discussion board entries, number / total time material viewed

The Prior Academic Achievement category consist of Pre-university data and University-data. Pre-university data describes about previous education student result and University-data describes about the current student academic result since entering the university. Generally, *Prior Academic Achievement* category indicates the student academic performance that it's commonly identified as grades [18]. *The Student Demographics* category indicates various characteristics of student and the circumstances in which student are born and grow up that strongly influence their well being and academic success [19]. *The Students' Environment* category indicates various characteristics of the courses that student enrolled [20]. *The Psychological* category indicates the interests and personal behavior of the student [19, 21, 22]. *The Student E-learning Activity* category indicates student's type of learning that consist of full online learning and blended learning [21] as a completion of traditional learning. In pandemic era, online learning be the most alternate learning that are implemented in all over the world.

III. METHODOLOGY

The aim of this research is to determine the best classification method that raised the highest accuracy for spesific combination of influential factor category in predicting students' academic success. This research would be able to answer the folowing question:

Q : What is the best classifier to predict students' academic success for spesific combination of influential factor category?

For the purpose the research objective and to get answer the research question above an educational datasets is taken from valid sources and on the dataset classification method is applied.

A. Proposed Framework

The first step to conduct this research is getting educational datasets. Then, identify from them to decide what categorized of the data. The dataset need to be categorized to determine the apropiate classification method. The data category [2] are *Prior Academic Achievement, Student Demographics, Students' Environment, Psychological and Student E-learning Activity*. Fig 1 describes the framework of this research

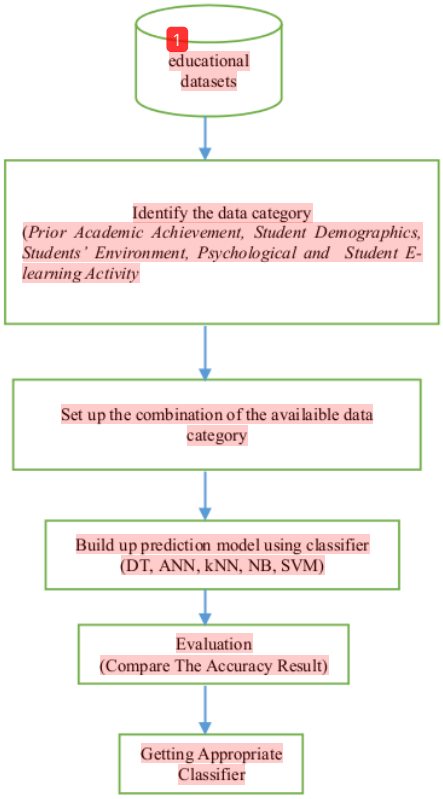

Fig. 1. Proposed Framework

B. Data

This research analyzed data of public dataset from UCI repository of machine learning datasets. The dataset are Open University Learning Analytics Data Set (OULAD). The dataset contains data about courses, students and their interactions with Virtual Learning Environment for seven selected courses and more than 32594 students. According to Table I the attributes on dataset dan its data category can be seen in Table II.

TABLE II. OULAD DATA SET AND DATA CATEGORY

Influential Factor	Factor Description	Factor Category
school	student's school	PA
sex	student's sex	SD
age	student's age	SD
address	student's home address type	SD
famsize	family size	SD
Pstatus	parent's cohabitation status	SD
Medu	mother's education	SD
Fedu	father's education	SD
Mjob	mother's job	SD
Fjob	father's job	SD
reason	reason to choose this university	PS
guardian	student's guardian	SD
traveltime	home to school travel time	SD
studytime	weekly study time	EL
failures	number of past class failures	SE
schoolsup	extra educational support	SE
famsup	family educational support	SE
paid	extra paid classes within the course subject	SE
activities	extra-curricular activities	SE
nursery	attended nursery school	SD
higher	wants to take higher education	PS
internet	Internet access at home	EL
romantic	with a romantic relationship	PS
famrel	quality of family relationships	PS
freetime	free time after school	PS
goout	going out with friends	PS
Dalc	workday alcohol consumption	PS
Walc	weekend alcohol consumption	PS
health	current health status	SD
absences	number of university absences	SD
GPA	Grade Point Average	PA

C. Experimental Setup

For the model generation, this research used the Rapid Miner Studio version 9.9 software package. This software is very powerful to build predictive analytic models because it's a suitable platform for data preparation, machine learning and model deployment as well. This experiment run on Intel Processor i7 10th generation, 12 GB RAM and Win10 operating system.

IV. RESULT

This section reported the performance evaluation of combination of multiple factor category and classifier for predicting academic success. The dataset prepared previously then imported to Rapid Miner Studio software. Next the software started to process dataset through following phases: data analysis, data pre-processing, and then design by applying the classification algorithm, training and testing. In the data analysis phase, the target attribute are determined. In the data analysis step, Rapid Miner divides data into two sets:training(90%) and testing(10%) which type of sampling is stratified sampling due to the label is nominal. In data pre-processing, the influential factors on data set is grouping according to factor category. Training process train the model used classification algorithm which requires criterion options like accuracy. During the testing process, it used two operations: The Apply Model on the test dataset and the Performance operation for measuring the model performance. In order to analyze the performance of these students, a prediction model is created based on classification algorithm which in the end helped to predict which students may passed or failed. The result of various one and multiple factor category are explaind in Table III, IV and V by applying classification method.

A. Experimental Results from Previous Research (One Factor Category)

TABLE III. SUMMARY OF ACCURACY OF VARIOUS MODEL (ONE FACTOR CATEGORY)

Factor Category	Classification Method Algorithms				
	DT	NB	ANN	KNN	SVM
PA	91%	75%	75%	83%	80%
SD	65%	76%	72%	-	
SE	68%	72%	98%	-	80%
PS	65%	-	69%	69%	83%
EL	-	-	-	-	-

B. Experimental Results with Multiple Factor Category

TABLE IV. SUMMARY OF ACCURACY OF VARIOUS MODEL (MULTI FACTOR CATEGORY)

Combination of Multi Factor Category	Classification Method Algorithms				
	DT	NB	ANN	KNN	SVM
PA+SD	91%	75%	92%	85%	83%
PA+SE	93%	78%	94%	82%	84%
PA+PS	90%	76%	91%	84%	80%
PA+EL	94%	75%	92%	80%	80%
SD+SE	90%	80%	95%	87%	82%
SD+PS	93%	82%	93%	86%	81%
SD+EL	96%	84%	96%	86%	84%
SE+PS	90%	79%	92%	85%	87%
SE+EL	93%	80%	94%	85%	84%
PS+EL	92%	80%	92%	84%	95%
PA+SD+SE	92%	75%	92%	85%	80%
PA+SD+PS	90%	84%	95%	87%	88%
PA+SD+EL	90%	80%	91%	85%	83%
PA+SE+PS	92%	78%	94%	87%	89%
PA+SE+EL	94%	76%	91%	84%	80%
PA+PS+EL	91%	73%	92%	85%	85%
SD+SE+PS	93%	78%	96%	82%	84%
SD+SE+EL	90%	81%	91%	80%	80%
SD+PS+EL	92%	77%	97%	80%	80%
SE+PS+EL	90%	80%	95%	87%	82%
PA+SD+SE+PS	94%	82%	93%	86%	84%
PA+SD+SE+EL	95%	86%	97%	86%	80%
PA+SD+PS+EL	96%	75%	96%	85%	87%
SD+SE+PS+EL	93%	80%	92%	93%	94%
PA+SE+PS+EL	94%	76%	96%	84%	80%
PA+SD+SE+PS+EL	95%	75%	98%	83%	87%

The outcome in Tables IV show the variety accuracy value for 26 combination of multiple category under 5 classifier in predicting students' academic success. According to Table IV we conclude the best combination between multiple category under 5 classifier that getting highest acuracy as shown in. Table V.

TABLE V. SUMMARY OF HIGHEST ACCURACY FOR COMBINATION OF MULTIPLE FACTOR CATEGORY AND CLASSIFIER

Combination of Multi Factor Category	The Best Classification Method
PA+SD	ANN
PA+SE	ANN
PA+PS	ANN
PA+EL	DT
SD+SE	ANN
SD+PS	DT,ANN
SD+EL	DT,ANN

Combination of Multi Factor Category	The Best Classification Method
SE+PS	ANN
SE+EL	ANN
PS+EL	SVM
PA+SD+SE	DT,ANN
PA+SD+PS	ANN
PA+SD+EL	ANN
PA+SE+PS	ANN
PA+SE+EL	DT
PA+PS+EL	DT
SD+SE+PS	ANN
SD+SE+EL	ANN
SD+PS+EL	ANN
SE+PS+EL	ANN
PA+SD+SE+PS	DT
PA+SD+SE+EL	ANN
PA+SD+PS+EL	DT,ANN
SD+SE+PS+EL	SVM
PA+SE+PS+EL	ANN
PA+SD+SE+PS+EL	ANN

V. Conclusion

This research successfully applies multiple factor categories (PA, SD, SE, PS, EL) and classifiers (DT, ANN, kNN, NB, SVM in predicting students' academic success. This research successfully gets the highest accuracy if the specific multiple factor category data meet the right classifier. The performance of the prediction model reaches the highest prediction result that can be effectively used to predict students' academic success.

References

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[2] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.

[3] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[4] A. Hellas *et al.*, "Predicting academic performance: A systematic literature review," *Annu. Conf. Innov. Technol. Comput. Sci. Educ. ITiCSE*, pp. 175–199, 2018, doi: 10.1145/3293881.3295783.

[5] M. A. Al-barrak and M. Al-razgan, "Predicting Students Final GPA Using Decision Trees : A Case Study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. July 2016, pp. 528–533, 2016, doi: 10.7763/IJIET.2016.V6.745.

[6] M. Christina, "Predicting Student Performance using Data Mining," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 172–177, 2018, doi: 10.26438/ijcse/v6i10.172177.

[7] J. Feng, "Predicting Students' Academic Performance with Decision Tree and Neural Network," pp. 2004–2019, 2019.

[8] J. Mesarić and D. Šebalj, "Decision trees for predicting the academic success of students," vol. 7, pp. 367–388, 2016, doi: 10.17535/crorr.2016.0025.

[9] M. Sivasakthi, "Classification and Prediction based Data Mining Algorithms to Predict Students ' Introductory programming Performance," no. Icici, pp. 0–4, 2017. doi:10.1109/ICICI.2017.8365371

[10] P. J. M. Estrera, P. E. Natan, B. G. T. Rivera, and F. B. Colarte, "Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School Abstract :," *Philippine*, vol. 3, no. 5, pp. 147–154, 2017. doi:10.29126/23951303

[11] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "On predicting academic performance with process mining in learning analytics," *J. Res. Innov. Teach. Learn.*, vol. 10, no. 2, pp. 160–176, 2017, doi: 10.1108/jrit-09-2017-0022.

[12] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, 2015, doi: 10.1186/s40165-014-0010-2.

[13] S. A. Kumar and M. N. Vijayalakshmi, "Appraising the Significance of Self Regulated Learning in Higher Education Using Neural Networks," *Int. J. Eng. Res. Dev.*, vol. Volume 1, no. Issue 1, pp. 9–15, 2012.

[14] E. Osmanbegovic and M. Suljic, "Data Mining Approach for Predicting Student Performance," 2012.

[15] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the cognitive skill of students in education environment," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, pp. 113–118, 2015, doi: 10.1109/ICCIC.2014.7238346.

[16] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance By an Application of Data Mining Techniques," *Manag. Artif. Intell.*, vol. 6, no. January 2017, pp. 110–114, 2011.

[17] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, no. December 2015, pp. 549–554, 2014, doi: 10.1109/IAdCC.2014.6779384.

[18] C. Anuradha and T. Velmurugan, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance," vol. 8, no. August, 2015, doi: 10.17485/ijst/2015/v8i15/74555.

[19] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 2, p. 26, 2018, doi: 10.9781/ijimai.2018.02.004.

[20] M. H. Mohamed and H. M. Waguih, "Early Prediction of Student Success Using a Data Mining Classification Technique," *Int. J. Sci. Res.*, vol. 6, no. 10, pp. 126–131, 2017, doi: 10.21275/ART20177029.

[21] A. Mueen, "Modeling and Predicting Students ' Academic Performance Using Data Mining Techniques," no. November, 2016, doi: 10.5815/ijmecs.2016.11.05.

[22] N. Putpuek, N. Rojanaprasert, K. Atchariyachanvanich, and T. Thamrongthanyawong, "Comparative Study of Prediction Models for Final GPA Score : A Case Study of Rajabhat Rajanagarindra University," pp. 92–97, 2018. doi:10.1109/ICIS.2018.8466475

Ermatita Paper 5 1570805890-Dafid

ORIGINALITY REPORT

97%	27%	97%	21%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dafid, Ermatita. "Determining Appropriate Classification Method Based on Influential Factors for Predicting Students'Academic Success", 2022 International Conference on Data Science and Its Applications (ICoDSA), 2022 Publication	92%
2	educationaltechnologyjournal.springeropen.com Internet Source	3%
3	Submitted to Telkom University Student Paper	2%
4	www.researchgate.net Internet Source	<1%
5	Eyman Alyahyan, Dilek Düştegör. "Predicting academic success in higher education: literature review and best practices", International Journal of Educational Technology in Higher Education, 2020 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		