Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities

Ching-Chieh Kiu
School of Computing and IT
Faculty of Built Environment, Engineering, Technology & Design
Taylor's University
Subang Jaya, Malaysia
chingchieh.kiu@taylors.edu.my

***Abstract*—Educational data mining techniques are widely used in academic prediction on student performance in classroom education. However most of the existing researches were studied and evaluated student coursework performance against the passing grade of the exam. In this paper, we performed analysis to identify the significant and impact of student background, student social activities and student coursework achievement in predicting student academic performance. Supervised educational data mining techniques, namely Naïve Bayesian, Multilayer Perceptron, Decision Tree J48 and Random Forest were used in predicting mathematic performance in secondary school. The prediction was performed on 2-level classification and 5-level classification on final grade. The experimental results have shown that student background and student social activities were significant in predicting student performance on 2-level classification. The model can be used for early predicting student performance to help in improving student performance on the subject.**

Keywords—Student Performance, Educational Data Mining, Decision Tree, Naïve Bayesian, Neural Network

I. INTRODUCTION

Educational Data Mining (EDM) has been actively applied to improve student performance in in education systems. Early prediction and analysis of at-risk student identification in classroom education may be helpful for both students and teachers. Teachers can have sufficient time to perform education interventions to improve students' performance [1]. The cycle of applying data mining in educational systems is depicted in Figure 1 [2]. The discovered knowledge may be helpful for teacher to use them to improve student performance, meanwhile the discovered recommendations may be helpful for student to use them to improve their performance in their studied subjects. In additional, applying educational data mining techniques on the education system can help in categorize academic records such as student details, learning pattern, activities, and performance in their classroom education [3].

Existing literature works primarily focus on the student coursework performance, teaching quality and learning activities in predicting student performance. However, student performance might be impacted by other factors such as study habits, attendance of school, social activities, student family background and others. Understanding the impact of these factors might be able to improve student performance in a subject as early as possible.

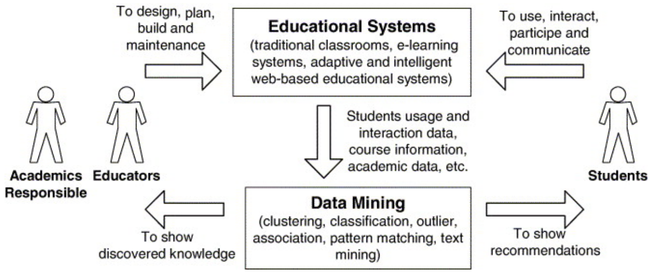


Fig. 1. Education data mining process in education systems [2]

In this paper, we aim to identify and analyse the impact of student background attributes and student social activities attributes on student performance. Supervised educational data mining namely, Naïve Bayesian, Multilayer Perceptron, decision tree J48 and random forest are applied to build prediction model. The significant and impact of student background and social activities attributes can be visualized and defined from the decision tree structure generated by the models.

The paper is presented as follows. Related works are discussed in next section. Following with Section III explains methodology and dataset. Meanwhile, Section IV discusses experimental result. Lastly, the paper is concluded with future works.

II. RELATED WORKS

EDM is used to identify students' learning patterns and process. They can be used to predict their performance to identify at-risk students at early stage. Prediction can be done based on their learning activities, coursework grades and learning outcome.

Educational big data and learning analytics approaches were applied in blended Calculus course by Lu et. al [1] for early prediction of students' academic performance. Principle component regression was used to predict students' final grade performance. Seven critical factors had been identified, whereas they consisted of three traditional factors and four online factors that impacted students' academic performance.

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

TABLE I. STUDENT BACKGROUND ATTRIBUTES

Student Background			
Attribute	**Description**	**Type**	**Value**
sex	gender of student		male | female
school	school of student		Mousinho da Silveira | Gabriel Pereira
address	type of student's home address		rural | urban
Pstatus	cohabitation status of parent	binary	living together | apart
famsize	size of family		≤ 3 | > 3
schoolsup	extra educational school support		yes | no
famsup	educational support from family		yes | no
Mjob	job of mother		- at home - civil services - teacher - health care related - other
Fjob	job of father		
reason	reason to choose this school	nominal	- close to home - school reputation - course preference - other
guardian	guardian of student		- father - mother - other
Medu	education of mother		0 # none 1 # primary education 2 # 5th to 9th grade 3 # secondary education 4 # higher education
Fedu	education of father		
famrel	quality of family relationships		very bad (1) to excellent (5)
age	age of student		15 - 22
traveltime	travel time from home to school	numeric	1 # < 15 min 2 # 15 to 30 min 3 # 30 min. to 1 hour 4 # > 1 hour
studytime	weekly study time		1 # < 2 hours 2 # 2 to 5 hours 3 # 5 to 10 hours 4 # > 10 hours
failures	number of failures in past class		n if 1 ≤ n < 3, else 4

TABLE II. STUDENT SOCIAL ACTIVITIES ATTRIBUTES

Student Social Activities			
Attribute	**Description**	**Type**	**Value**
activities	extra-curricular		
higher	plans for higher education		
internet	home internet access		
nursery	nursery school attended	binary	yes | no
paidclass	extra paid classes		
romantic	in romantic relationship		
absences	absences from school		
health	status of current health		
freetime	free time after school	numeric	very low (1) to very high (5)
goout	outing with friends		
alc	consume alcohol in weekday		
Walc	consume alcohol in weekend		0 - 93

TABLE III. STUDENT PERIOD RESULTS ATTRIBUTES

Student Coursework Result			
Attribute	**Description**	**Type**	**Value**
GI	1st grade period	numeric	0 - 20
G2	2nd grade period		

Suhem Parack et. al [2] applied data mining in education for student grouping and profiling to predict student performance. Apriori algorithm was used to discover co-relations among set of items, then student grouping was evaluated using K-means clustering by assigning a set of observations into subsets.

Romero et. al [4] predicted student performance based on the data collected from online discussion forum. The data were separated into data subsets on a weekly basis. Several data-mining methods had been applied on predicting accuracy of each data subset. Sequential minimal optimization classification algorithm was used to predict student interaction before a midterm exam for predicting student performance.

Decision tree classifier was used by [5] to develop an early warning system to identify at-risk student. A data consisted of 300 students with 13 online attributes was used to build a prediction model. The model achieved 95% accuracy based on 1–4 weeks of data from a skewed data set in predicting whether students would pass or fail.

Naive Bayes classifier was applied on the data collected during freshman year to predict students' grades in their final year [6]. Meanwhile, [7] used regression to predict students' grades. Their algorithm achieved 76% accuracy in prediction.

Sentiment analysis was used by Yu et. al [8] to identify affective information to improve predictive accuracy for the early identification of students who are likely to fail in a subject.

III. METHODOLOGY AND DATASET

In this experiment, the real dataset [9] consisted of 395 instances with 33 attributes that described performance in Mathematics subjects is used. The attributes of the dataset are divided to three subsets:

1) students background with 18 attributes (Table I)
2) student social activities with 12 attributes (Table II)
3) student coursework results with 2 attributes (Table III)

These subsets attributes will be used to predict final grade (G3). G3 is a numeric datatype with range of 1 – 10 used to measure student performance on their final grade. The subset attributes will be evaluated under two models:

- 2-level classification (pass / fail)
- 5-level classification (A / B / C / D / F) (Table IV)

TABLE IV. 5-LEVEL CLASSIFICATION ON G1, G2 AND G3 RESULTS

Mark	16-20	14-15	12-13	10-11	0-9
Grade	A	B	C	D	E

Data conversion and normalization have been applied on following attributes prior to the evaluation of prediction models.

- Age attribute (student background subset) is converted to nominal.
- Absences attribute (student social activities) is normalized to categorial value as depicted in Table V.

- 1st period grade (G1) and 2nd period grade (G2) attributes (student coursework results) are converted to 2-level classification and 5-level classification to predict G3 at 2-level classification and 5-level classification models.

TABLE V. 4-LEVEL NORMALIZATION ON ABSENT DAYS

Absent Day	0-10	11-20	21-50	50-100
Absent	Low	Normal	High	Very High

IV. EXPERIMENTAL RESULTS

Weka data mining tool [10] is used to perform analysis on the dataset. Four supervised educational data mining techniques, namely Naïve Bayesian, Multilayer Perceptron, Decision Tree J48 and Random Forest. The evaluation has been performed on the three subset attributes on 2-level classification and 5-level classification models as shown in Figure 2. The experimental analysis also performed on all attributes which is referred as all subsets dataset.

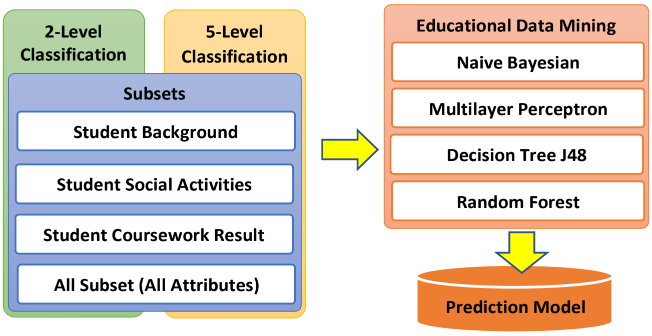


Fig. 2. Experimental Models in Educational Data Mining [1]

As depicted in Table VI (Figure 3) and Table VII (Figure 4), decision tree J48 outperformed other educational data mining algorithms in 2-level classification and 5-level classification on all subsets dataset except student social activities subsets in 2-level classification. Meanwhile, Naïve Bayes outperformed other in evaluating student social activities subsets in 2-level classification and 5-level classification models.

This experimental result also has shown that student coursework results are significant attributes in predicting student performance in mathematic final grade as it has highest precision accuracy 0.924 in 2-level classification and 0.791 in 5-level classification. In overall, accuracy of algorithms in 2-level classification are out performed models in 5-level classification. As shown in Table VI (Figure 3), the models accuracy are > 0.5, this indicated that student background and student social activities are viable to be used to perform early analysis and prediction of at-risk student to determine whether it pass or fail the subject.

An explanatory analysis was performed on 2-level classification model. Decision trees are generated to identify the relevant attributes that might direct impact students' performance. As depicted in Figure 5, following attributes are significant to predict student performance:

- In student background subset, *failures, schoolsup, age, Mjob, school, famsize* and *reason* are significant attributes
- In student social activities, *higher, absences days, activities, Walc, romantic* and *health* are significant attributes
- In sudent coursework subset, *G1* and *G2* are significant attributes
- In all subsets, *G1 and G2* are significant attributes.

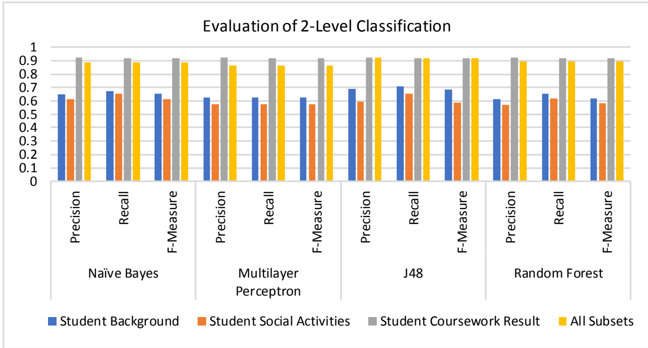


Fig. 3. Accuracy comparision of 2-level classification

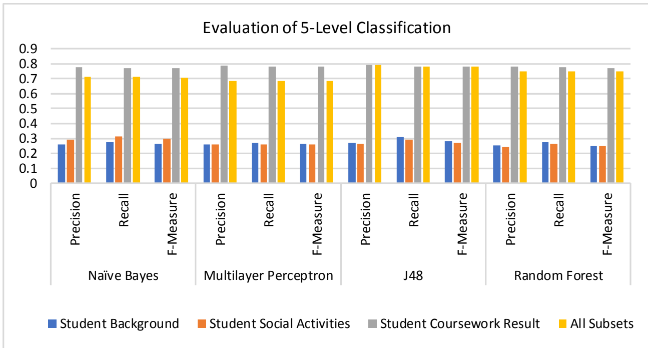


Fig. 4. Accuracy comparision of 5-level classification

V. CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated that student background and social activities are significant to be applied for early prediction on student performance and also can be used to identify at-risk student. Hence, early prediction with these models may be helpful for the teachers and students. Students may be able to perform better in the academic performance. Meanwhile, teachers can do early preparation to perform education interventions in teaching the subject.

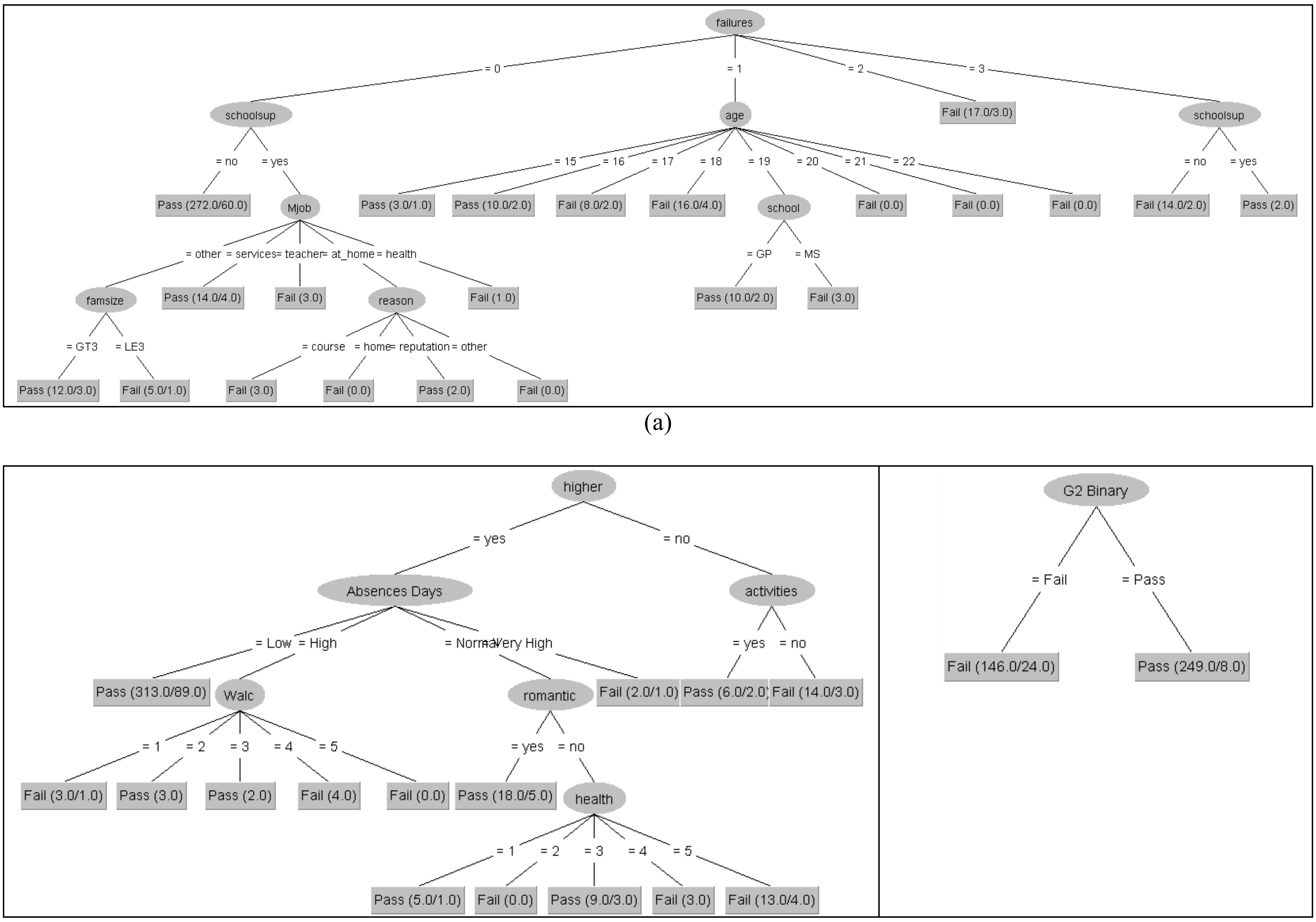
In future work, unsupervised education data mining techniques will be applied to discover correlation and impact of the attributes in clusters. In addition, we will discover the correlation and impact using attributes analysis and feature selection to provide more accurate prediction models for predicting early at-risk students' performance.

TABLE VI. ACCURACY OF 2-LEVEL CLASSIFICATION

		Student Background	Student Social Activities	Coursework Result	All Subsets
Naïve Bayes	Precision	0.648	**0.61**	**0.924**	0.889
	Recall	0.671	0.651	0.919	0.889
	F-Measure	0.653	0.612	0.92	0.889
Multilayer Perceptron	Precision	0.622	0.576	**0.924**	0.863
	Recall	0.625	0.575	0.919	0.863
	F-Measure	0.624	0.576	0.92	0.863
J48	Precision	**0.687**	0.595	**0.924**	**0.924**
	Recall	0.706	0.656	0.919	0.919
	F-Measure	0.681	0.587	0.92	0.92
Random Forest	Precision	0.614	0.572	**0.924**	0.894
	Recall	0.653	0.62	0.919	0.894
	F-Measure	0.616	0.583	0.92	0.894

TABLE VII. ACCURACY OF 5-LEVEL CLASSIFICATION

		Student Background	Student Social Activities	Coursework Result	All Subsets
Naïve Bayes	Precision	0.258	**0.291**	0.775	0.710
	Recall	0.278	0.316	0.772	0.711
	F-Measure	0.266	0.295	0.769	0.709
Multilayer Perceptron	Precision	0.261	0.259	0.788	0.687
	Recall	0.268	0.261	0.782	0.686
	F-Measure	0.264	0.26	0.78	0.687
J48	Precision	**0.273**	0.266	**0.791**	**0.791**
	Recall	0.306	0.294	0.785	0.785
	F-Measure	0.283	0.272	0.782	0.782
Random Forest	Precision	0.254	0.246	0.783	0.749
	Recall	0.278	0.263	0.777	0.752
	F-Measure	0.25	0.25	0.774	0.748



(b) (c)

Fig. 5. Decision tree structure for (a) student background (b) social activitiest and (c) coursework result or all components on 2-level classification

REFERENCES

[1] O. H. Lu, A. Y. Huang, J. C. Huang, A. J. Lin, H. Ogata, and S. J. Yang, “Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning”, Journal of Educational Technology & Society, 21(2), 2018, pp.220-232.

[2] P. Suhem, Z. Zain, and M. Fatima, “Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns”, 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), 2012, pp 1 – 4

[3] C. Romero, and S. Ventura, “Educational data mining: A Review of the state of the art”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010, 40(6), pp.601-618.

[4] C. Romero, M. I. López, J. M. Luna, and S. Ventura, “Predicting students' final performance from participation in online discussion forums”, Computers & Education, 2013, 68, pp.458-472.

[5] Y. H. Hu, C. L. Lo and S. P. Shih, “Developing early warning systems to predict students’ online learning performance”, Computers in Human Behavior, 2014, 36, 469-478.

[6] R. Asif, A. Merceron, and M. K. Pathan, “Predicting student academic performance at degree level: A Case study”, International Journal of Intelligent Systems and Applications, 2014, 7(1), 49-61.

[7] Y. Meier, J. Xu, O. Atan, and M. Schaar, “Predicting grades”. IEEE Transactions on Signal Processing, 2016, 64(4), 959-972.

[8] L. C. Yu, C. W. Lee, H. I. Pan, C. Y. Chou, P. Y. Chao, Z. H. Chen, S. F. Tseng, C. L. Chan, and K. R. Lai. "Improving early prediction of academic failure using sentiment analysis on self□evaluated comments." Journal of Computer Assisted Learning, 2018.

[9] P. Cortez and A. Silva, “Using Data Mining to Predict Secondary School Student Performance”, In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[10] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.