USING DATA MINING MODELS TO PREDICT STUDENTS' ACADEMIC PERFORMANCE BEFORE THE ONLINE COURSE START

Tonghui Xu, University of Massachusetts Lowell

ABSTRACT

The early detection of students' academic performance or final grades helps instructors prepare their online courses. In the Open University Learning Analytics Dataset, I found many online students clicked the course materials before the first day of class. This study aims to investigate how data mining models can use this student interaction data to predict their academic performance. In this study, this interaction information is called "week 0" data. The results suggest that "week 0" interaction data can be used to identify the academic success of online students and predict first assignment performance.

Keywords: *educational data mining, online learning, early detection*

INTRODUCTION

During COVID 19, all universities had to replace traditional face-to-face courses with distance courses. Distance education includes e-learning, blended learning, and online learning. In the 2009 PISA assessment report, about 15% of students in OECD countries reported that they did not have Internet at home, but in 2018, that number dropped to less than 5% (Schleicher, 2019). Because of the popularity of the Internet, online distance learning has become a trend in higher education, because it is a convenient and flexible learning environment for students. However, in online learning education, instructors have less interaction with students. In a face-to-face course, an instructor can observe the students' in-class activities or combine their assignments to identify students with high or low risk in their studies. However, in the early stage of an online class, the instructor does not have sufficient data to identify the academic success of students. So, online interaction with web materials becomes an important piece of information to help the instructor to evaluate the students' final grades.

The most common university online data system includes massive open online courses (MOOCs), virtual learning environments (VLE), and learning management systems (LMSs) (Hussain et al., 2018). In this study, I selected the Open University Learning Analytics Dataset (OULAD) as the study sample. The OULAD is a virtual learning environment system and belongs to the Open University (Kuzilek et al., 2017). This system offers resources, activities, and interactions, as well as different stages of assessment (Britain and Liber, 1999). The online interactions include online discussion, question posts, the total sum of material clicks, and frequency of viewed materials. In addition, designing an early detection system is an important way to identify students' academic success and predict students' assignment performance. Hussain et al. (2018) use the sum of material clicks of students and their demographic data to predict their first assessment score. They employed different decision tree models to prove that student engagement (material clicks) will impact their performance. Hlosta et al. (2017) used the students' click information to identify at-risk online students. Those studies proved that data mining methods can use students' online

interaction information to identify at-risk students and predict assignment' performance.

In this study, I found that many Open University online students have already clicked the course materials before the first day of courses. No research articles focus on this part of the data. Therefore, I named "week 0" data for that information. This study aims to integrate different classification data mining algorithms into those interaction data and student personal information data to identify academically successful students and predict first assignment performance. Suppose the data mining models can employ the "week 0" data to predict students' academic success. In that case, they will prove that "week 0" can be used by educational data mining research. Accordingly, I designed two following research questions:

RQ1. Can I use the classification data mining algorithms to identify academically successful students by their "week 0" interactions with and without student information?

RQ2. Can I use the classification data mining algorithms to predict first assignment performance by their "week 0" interactions with and without student information?

RQ3. If the data is usable, based on the outcomes of the following seven online courses, what are their similarities and differences? Which model is the better one?

I used the WEKA software to conduct my research. WEKA is data mining software designed by the University of Waikato. This software can be freely downloaded in http://old-www.cms.waikato.ac.nz/~ml/weka/. Hall et al. (2009) introduced how to use WEKA to conduct data mining models in big data analysis. The literature review is discussed in section two. Methods and data are presented in section three. Section four displayed the results and discussion. Finally, section five discussed the conclusion.

LITERATURE REVIEW

Many articles discussed using WEKA to conduct the data mining analysis in educational research. Bresfelean (2007) used "WEKA J 48 decision tree classifiers" to predict students' choice in continuing their education with post-university studies. Aher and Lobo (2011) employed different data mining models to analyze the educational database in the WEKA environment. Many educational research papers use WEKA to predict students' academic performance (Hussain et al. 2018; Roy and Garg, 2017) and identify the high-risk dropout students (Jayaprakash et al. 2014). Those articles show that WEKA is an effective data mining software for analyzing educational data.

A number of studies focused on the week-wise design or early detection to predict the students' academic performance. Marbouti et al. (2015) utilized data mining models to predict that first-year students have a risk of failure in the 2nd, 4th, and 9th weeks of an online course. Summers et al. (2020) used the first three weeks of student records to measure the online students' engagement. Student engagement is an important educational factor to evaluate the students' performance. Aljohani et al. (2019) employed machine learning models to predict at-risk students based on the first ten weeks' clickstream data in a virtual learning environment. They indicated that the model could predict pass/fail with around 90% accuracy. Alyahyan and Düştegör (2020) summarized important research articles that focus on predicting students' academic success in higher education.

Several studies have been conducted to investigate student performance, learning behaviors, and status by data mining models in the online learning environment. Aldowah et al. (2019) summarized influential educational data mining and learning analytics articles and classified each paper. Al-Radaideh et al. (2006) found that the classification models can be used to predict students' final grades in a specific course. Several research papers used classification models to predict students' dropout rate from distance learning and e-learning courses (Kotsiantis et al., 2007; Lykourentzou et al., 2009). Sabourin et al. (2011) monitored students' online activities to understand students' self-regulated learning behaviors. Such articles supported unity of the data mining method in online educational research.

In addition, some of the papers contribute the researchers' educational data mining studies on the OULA dataset. These include the following (a) identifying the student at risk of low performance or engagement (Hussain et al. 2018; Kuzilek et al. 2018); Peach et al. 2019; and (b) identifying student academic performance (Azizah et al. 2018; Rizvi

et al. 2019). For example, according to a recent research paper, Waheed et al. (2020) used an artificial neural network, SVM, and linear regression models on a set of unique handcrafted features extracted from the virtual learning environments clickstream data to investigate at-risk students providing measures for early intervention. They found those data mining approaches can effectively predict the at-risk students.

Overall, most of the previous research did not emphasize that the "week 0" interaction data can be used to predict and assignment scores or identify the students' final results. Therefore, this study will focus on the "week 0" online interaction data to design an early detection to assist the online course instructors in understanding and helping online students to improve their performances and achieve academic success in courses.

METHOD

Data Source

In this study, the Open University Learning Analytics (OULA) dataset is designed by a standard query language (SQL) database. The data can be freely download at https://analyse.kmi.open.ac.uk/open_dataset. The structure of OULA data, including six data files and seven online courses from 2013 to 2014 (Kuzilek et al., 2017). The studentInfo, studentRegistration, and course files include student personal information, final results, and course information. The student assessment scores, and assessment information are stored in assessments and studentAssessment data files. The studentVLE and VLE datasets include student online interaction types and the sum of clicks. Three social science and four STEM online courses are recorded in this system. The total population consist of approximately 32000 students from seven individual online courses.

Missing data

The Open University data is a public resource and incomplete records exist in all seven courses. Therefore, I used the listwise deletion method to disregard all the records which have missing value in the final results and first assignment scores, and then employed the mean and frequency values to impute other missing values.

Sample

The date variable indicates the date of students' interaction with course materials. I only need the date value less than 1. It means this student has already accessed the online course materials before the first day of the course. The total number of this study is approximately 18500 students. Table 1 displays the basic statistical description of student information of each course. Table 2 shows the frequencies of "week 0" interaction data variables. The statistical descriptions of category variable include frequency and percentage. For continuous variable, the descriptions consist of minimum, maximum, mean, and standard deviation.

Dependent variables

Final result. This original variable consists of four categories, such as "Fail," "Pass," "Distinction," and "Withdrawn." I converted the "Distinction" and "Pass" to "Pass" and transferred "Fail" and "Withdrawn" to "Fail." Only passed students assumed as academically successful. I assumed that "Fail" students would fail or withdraw from their courses. Student withdrawal courses can be assumed a waste of educational resources.

Score. The original variable is a continuous variable. I converted it to a binary variable. The new binary variable consists of two outcomes: (a) high and (b) low. The criterion of conversion is if the assignment score is equal to or lower than 60, then this observation will be labeled as "low", otherwise the label is "high."

Independent variables

Activity type. The original activity type variable consists of different serval types of online learning activities. I covert each category to new dummy variables. This student interaction variables include forum, subpage, homepage, resources, ouelluminate, and oucontent. The forum variable indicates the student interacts with the online discussion forum. The subpage reveals the student's navigation path through the VLE structure, and the resources variable includes course materials, lecture slides, and books (Wolff and Zdrahal, 2012). The homepage variable reflects the homepage of every course in the VLE system. The oucontents variable indicate the content information of online course, and the ouelluminate variable represents the virtual space where the university hold classes or meetings.

Count. I created a dummy variable called

Table 1 The statistical descriptions of each course

				Frequency			
Variable	AAA	BBB	CCC	DDD	EEE	FFF	GGG
Type	Social	Social	STEM	STEM	STEM	STEM	Social
Gender							
Female	266	3661	442	1752	220	787	1233
Male	356	435	1205	2513	1644	3559	236
Age band							
0-35	279	2641	1124	3146	1402	3188	888
35-55	343	1455	523	1119	462	1168	581
IMD band							
0-30%	119	1431	452	1201	481	1309	498
30%-60%	185	1330	512	1356	575	1400	471
60%-100%	330	1335	683	1708	808	1647	500
Education							
HE degree	151	564	387	747	392	671	71
Less than A level	136	1681	431	1400	589	1653	877
A level or high	347	1851	829	2118	883	2053	521
Region							
London	53	307	190	369	135	447	165
South and Yorkshire	202	972	402	1138	425	1095	433
North Region	108	673	292	813	348	820	304
West Region	188	1046	380	1107	447	1123	462
Ireland, Scotland, Wales	83	1098	383	838	509	871	105
Score							
High	527	3954	1031	2388	1480	3897	1363
Low	107	142	616	1877	384	459	106
Final result							
Pass	484	2813	856	2305	1398	3207	1209
Fail	150	1283	791	1960	466	1149	260
Total observation	634	4096	1647	4265	1864	4356	1469
				Mean			
date registration	-78	-67	-67	-74	-64	-68	-54
studied credits	82	82	74	82	64	86	34

Table 2 The average frequencies of student interaction data of each course

	Mean Frequency						
	AAA	BBB	CCC	DDD	EEE	FFF	GGG
Sum_click	168	57	74	97	97	148	52
homepage	34	14	16	29	19	33	15
ouelluminate	73	9	11	13	32	38	12
oucontent	73	9	11	13	32	38	12
resource	5	5	8	7	2	6	9
subpage	12	6	6	20	3	23	4
forumng	39	20	3	19	29	31	10
count	39	18	41	43	30	49	17

"count." This variable is accumulated by the frequencies of activity type was accessed by students. For instance, if the number of counts is equal to 10, it means this student access 10 times of online materials.

Sum_click. This variable reflects the total number of online students who click course materials before the start of the course. This is an important attribute to count the students' frequencies of interaction data.

Covariate variables

Age band. This variable indicates the different categories of online course students' age intervals. The range of student age between 0 and 55.

Date registration. The date variable specific date of students' registration. For instance, if the date value is "-10", it means this student registered for this course 10 days before the first day of the online course.

Gender. Gender is an important demographic variable that may impact the students' academic success in educational research.

IMD band. This is a composite variable that includes living place income, education, crime rates, etc. The higher score of IMD band indicates the student living areas has high living conditions, otherwise, the lower score means the student living areas has lower living conditions.

High education. The high education is a categorical variable that includes three categories: High school, less than A level, and A level or higher.

Studied credits. This variable accumulates the total students' credits of each online student. The range of student credits between 30 and 655.

EDUCATIONAL DATA MINING

Knowledge discovery in databases (KDD) focuses on the overall process of knowledge discovery from data, including how the data are stored and accessed, how algorithms can be scaled to massive data sets and still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported (Fayyad et al.,1996). Figure 2 shown the overview of KDD steps of this study. First, select the sample from the Open University online data, and process the data to clear the sample. Second, transfer the variable type or create dummy variables. Third, use data mining models to analyze the data and evaluate the models. Finally, based on the result to construct a knowledge pattern.

Educational data mining (EDM) applies data mining methods to educational data (Baker & Yacef, 2009). Data mining is increasingly gaining significance in studying online learning behavior and student performances in educational research. In this paper, I choose the classification decision tree, random forest, and K-nearest neighbors to construct the data mining models. The classification model is supervised machine learning. Supervised learning is required to split the entire data into two disjoint datasets: a training set and a testing set. The training set is used to construct the model, and testing data will evaluate the performance of the model.

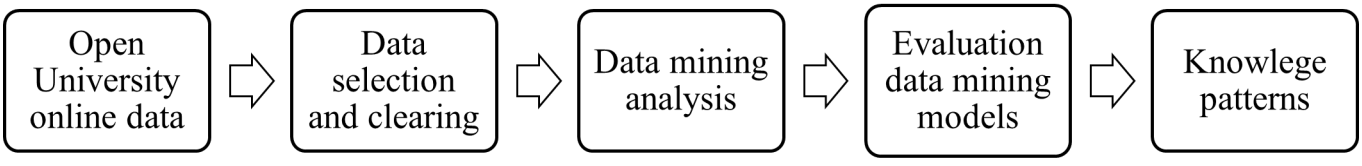
Decision tree (DT)

The decision tree is a tree-like structure supervised algorithm and consists of a root and several branches that represent attributes and consequences of the test. In this study, I choose J-48 decision tree to conduct my research. An internal node may contain two or more leaf nodes. Internal nodes represent the characteristics of the database. The general rule of DT is if condition A is true, then print the outcomes. If not, check other conditions until there is no path. The advantages of the decision tree include that (a) DT uses an "If-then" statement, which is easy to understand, interpret and apply, and (b) DT can visualize the outcomes of data predictions.

Random forest (RF)

Random forest is one of the decision tree

Figure 1. The KDD Processes



algorithms methods. This approach will randomly construct different decision trees in the forest and input data into each tree. Each tree will compute the outcome, and the outcome with the highest frequency is selected as the prediction of the model. The strengths of the random forest include (a) RF has high performance in prediction and low error rate, and (b) RF can be understood and used easily.

K-nearest neighbors (K-NN)

K-nearest neighbors (K-NN) will search for the nearest neighbors in the entire training dataset, and it does not have a training phase. So, the K-NN is also called the lazy learner algorithm. The K parameter is to set up an integer value of numbers of nearest neighbors. This approach is based on the K value to search for the nearest neighbors to predict the outcomes. The strengths of K-NN contain (a) short running time, (b) the K-NN can be interpreted and used easily, and (c) good performance in the prediction.

DATA ANALYSIS

The classification data mining model is a supervised method applied in educational data mining research. Accuracy, recall, and precision are the most important scales to evaluate the performance of the data mining model. In this study, I employed the 10 cross-validation (CV) method to divide my dataset into training and testing sets. The basic rule of the 10 CV approach is to split the data into 10 pieces. In the first iteration, CV chooses the first piece as the testing set to evaluate the model and others as a training set to construct a model. In the second iteration, CV selects the second piece as the testing set, and the first and the remaining pieces are the training set. After 10 times of iterations, the 10 CV approach stops. The cross-validation method is used to avoid the overfitting issue in data mining and improve the performance of the data mining model. Over-fitting is a common problem in data mining, and it will cause the model cannot objectively and reliably predict future observations.

In data analysis part, I used a decision tree, K-nearest neighbors, and random forest to analyze the students' "week 0" interaction data with and without student personal information data to identify the successful students. I constructed three different classification data mining models for each online course. After that, I computed the accuracy, precision, and recall to evaluate my classification models in each course data. In the next step, I repeated the same analysis process but only used student personal information to identify the successful students.

In part two, I applied the same classification data mining models and sample to predict the students' first assessment scores of each course and calculate the accuracy, precision, and recall to evaluate the classification models in each online course. Finally, I chose to use interaction data and student personal information data separately to repeat the same analysis process.

CONFUSION MATRIX

The confusion matrix is used to present the data mining prediction results. Table 3 showed examples of confusion matrix tables. For example, one of my dependent variable is binary and consists of "Pass" and "Fail." The Ture positive (TP) and Ture negative (TN) means the predicted result march the actual result. In this study, I am interested in whether the student will pass online courses, so I set "Pass" as the positive. It meant I would predict a student would pass this course. False-positive (FP) shows the predicted value is "Fail," but the actual value is "Pass." Ture negative (FN) displays the predicted value as "Pass," but the actual value is "Fail."

Evaluation model

Based on the confusion matrix results, the performance of machine learning can be computed by single detection theory and diagnostic accuracy theory (Swets, 1988; Zweig & Campbell,1993), and accuracy, precision, and recall will evaluate

Table 3 Confusion Matrix Table

		Predication	
Actual		Positive	Negative
	Positive	Ture-positive (TP)	False-positive (FP)
	Negative	Ture-negative (FN)	Ture-negative (TN)

the performance of machine learning model. In general, the high score of each measurement indicates high performance. Therefore, the accuracy of a classification data mining model over 70 % will be considered a good model and then check other measurements' results.

Accuracy is the intuitive performance measure, and it is the total correctly predicted records divided into the total records.

Precision is the percentage of correctly predicted positive records to the total predicted positive records. For instance, for all students who are labeled as "Pass", how many students actually pass the course.

Recall is the percentage of correctly predicted positive records to all records in the actual positive class. For instance, among the students who actually pass the course, how many they are marked as "Pass" by prediction.

RESULTS

Figures 2 and 3 showed the accuracy of prediction in students' final grades and their performances in their first assignments. Both figures consisted of 3 rows and columns, and 9 histograms. The first histogram showed the accuracy of a decision tree model, the middle one indicated the accuracy of a K-nearest neighbor model, and the right histogram indicated the accuracy of a random forest model. The first row showed the accuracy of three models with "week 0" interaction data and student personal information; the second row represented accuracy of three models with "week 0" data only; and the last row showed the accuracy of three models with student personal information data. I only discussed the course with accuracy value over 70%. In the first experiment, for instance, the first histogram of figure 2 and first column of table 4 (Appendix A) showed the accuracy values of decision trees to predict students' final grades in different courses. The accuracy of the "AAA" course was 74%; the accuracy of the "EEE" course was 72%; and the accuracies of the "FFF" and "GGG" courses were 74% and 82% respectively. Table 5 (Appendix B) and 6 (Appendix C) indicated the precision and recall of three models. For example, the precision of the "AAA" course was 77%; the accuracy of the "EEE" course was 76%; and the accuracies of the "FFF" and "GGG" courses were 74% and 83% respectively. Finally, the recall of the "AAA" course is 94%; the accuracy of the "EEE" course is

Figure 2. Accuracy of Classification Models to Predict Students' Final Grades

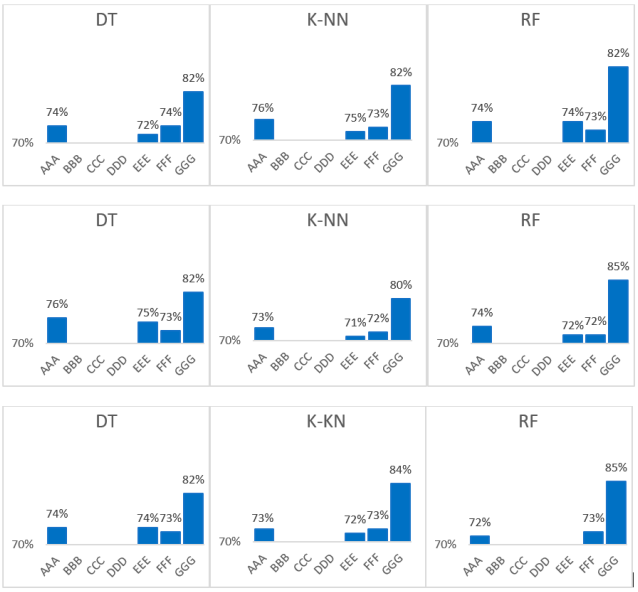
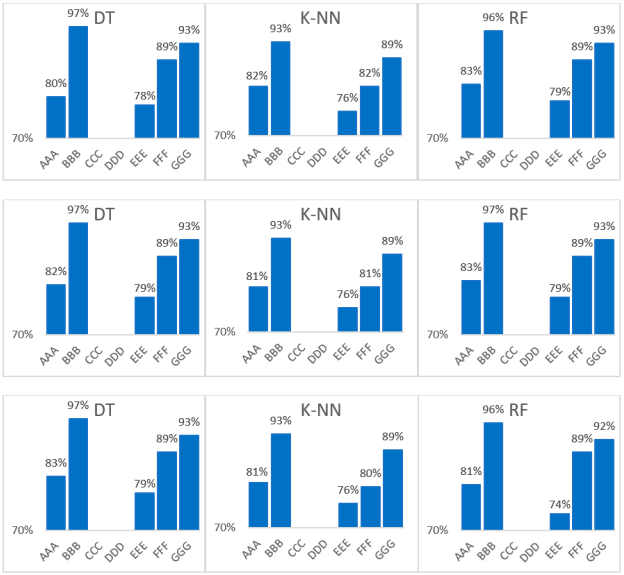


Figure 3. Accuracy of Classification Models to Predict Students' Performance on First Assignment



93%; and the accuracies of the "FFF" and "GGG" courses are 98% and 99% respectively.

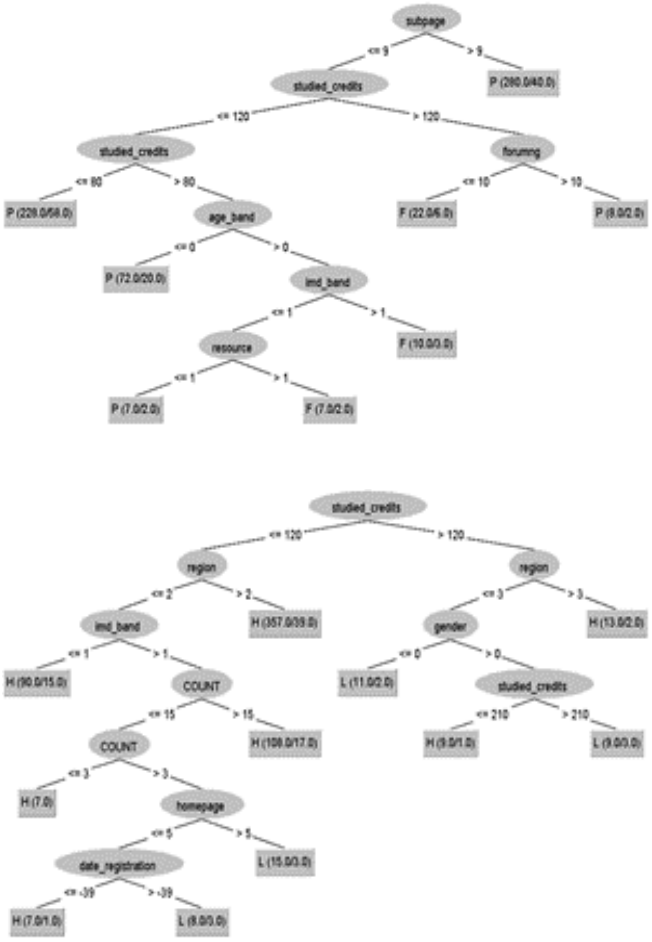
In the second experiment, five courses' accuracy values were over 70%. For example, the first histogram of figure 3 and first column of table 7 (Appendix D) indicated the accuracy values of decision trees to predict students' performance on first assignment in different courses. The accuracy of the "AAA" course was 80%; the accuracy of the "BBB" course was 97%; the accuracy of the "EEE" course was 78%; and the accuracies of the "FFF"

and "GGG" courses were 89% and 93% respectively. Table 8 (Appendix E) and 9 (Appendix F) showed the precision and recall of three models. For example, the precision of the "AAA" course was 82%; the accuracy of the "BBB" course was 97%; the accuracy of the "EEE" course was 80%; and the accuracies of the "FFF" and "GGG" courses were 90% and 93% respectively. Finally, the recall of the "AAA" course was 85%; the accuracy of the "BBB" course was 100%; the accuracy of the "EEE" course was 97%; and the accuracies of the "FFF" and "GGG" courses were 100% and 97% respectively.

Figure 4 indicated the tree-like outcomes of decision tree prediction in the course "AAA". The right side was the decision tree of predicted students' final grades, and the left side was the decision tree of predicted student assignment performance. The circle indicated the names of independent variables, and the rectangles show the prediction outcomes. In the left tree, the outcome includes: (a) "Pass" and (b) "Fail," and the outcomes of right tree consist of "High" and "Low." For instance, in the left decision tree, "subpage" was the root independent variable. In the first split, if a student's clicks of "subpage" materials exceeded 9, then moved to the right branch and printed "P" in the outcome. The outcome showed that in this condition, 280 students passed the course, and only 40 students did not pass the course. It meant that if a student has this characteristic, the decision tree model will predict that this student passed the course. In the left side, if a student's clicks of "subpage" materials were less or equal to 9, then move to the next level to check student credits.

In the first experiment, I used DT, K-NN, and RF to predict students' final grades. Based on the performances of model evaluation, I found that the "week 0" data can be used to predict their final grades by classification data mining models. Figure 2 showed the accuracy of predicted students' final grades. Four online courses' accuracy values were over 70% by "week 0" data with and without student information. These results proved that "week 0" can be used to identify the academically successful student before the course starts. Compared with the performance of the three classification data models with different datasets, the three models of the three courses have similar accuracy, precision, and recall. However, I found in some cases, the values of recall of decision trees were equal to 100%.

Figure 4. *Selected Outcomes of Decision Trees*



This meant the decision tree model would predict all students to be successful. In addition, in some cases, values of recall of random forest were close to 100%, and it meant random forest would more likely predict more students to be successful. Both models may predict some at risk students to be successful students. The decision tree model has more bias in its prediction.

In the second experiment, I found that five courses' accuracy values were over than 70% by "week 0" with and without student information. It meant that the "week 0" data can be used to predict their first assignment performances by classification data mining models. Compared with the performance of the three classification data models with different datasets, "BBB" course and "GGG" course have higher evaluation performance and "AAA" and "EEE" courses have similar evaluation performance. Similarly of first experiment, I found in

some cases, the values of recall of decision trees were equal to 100%. This meant the decision tree model will predict all students to be successful in the first assignment. In addition, in some cases, some values of recall of random forest approach 100%, which meant random forest will more likely predict more students to be successful in the first assignment. Both models may predict some at risk students to be successful students. The decision tree model has more bias in its prediction.

In both experiments, I also chose to use student information data only to predict students' final grades and their first assignment performances. These evaluation performances proved that student personal information is also an important data source to predict students' academic performances and final grades. Based on the two experiments, the evaluation performances of the predicted first students' assignment performances are better than the predicted students' final grades. It meant the "week 0" and student personal information has better performance to predict student assignment performance in early stage of online courses. Compared to the overall evaluation performance, although the K-NN model did not have higher performance in prediction, this model was the most stable one. The decision tree model had good accuracy and recall values, but this model has more bias. The random forest had higher prediction performances, and the bias was lower than the decision tree. However, the decision tree can be used to construct a tree-like figure to show the data mining process and important independent variables. Therefore, using multiple classification data mining models is necessary.

Finally, this study also proved that the quality of data is the most important factor in data mining models. For example, the course "DDD" had similar interaction data with the course "EEE," but all three classification models could not predict well in the course "DDD." In summary, "week 0" data can be used to predict student final grades and assignment performances, but it depends on the quality of the data and prediction model. The educators should not ignore this part of the data or aggregate it to first week interaction data.

CONCLUSIONS

Identifying students' academic success and predicting the assignment performance are important research topics in educational research. Based on the results, the "week 0" data can support useful information to the online course instructors, but it depends on the quality of the data and data mining models. The utility of the "week 0" data is limited but not meaningless. Both experiments indicated that the random forest (DT) and K-nearest neighbor (K-NN) are the two appropriate algorithms for predicting students 'academic success. Decision tree model has more bias in prediction. However, the decision tree can be used in data mining visualization. In addition, these results also prove that the early detection system and related studies should include the "week 0" data to predict student final grades or assignment performance. Arnold and Pistilli (2012) indicated that learning analytics dashboards would induce positive drives in students learning, consequently impacting performance. Hussain et al. (2018) designed an instructor dashboard to provide the prediction results and explanations to an instructor.

In my future studies, I consider using the feature selection models to study the "week 0" data to identify important independent variables that will impact the students' academic success and assignment performances. It can help the instructor know the important online activities and student characteristics are relevant variable of student final grades. For students, feature selection can offer them to knowledge about which course materials are important. They will know the core activities of this course. Finally, I hope my study can engage more course instructors to post the courses materials before the first week of courses to improve students' engagement to study and suggest researcher does not disregard the interaction data before the beginning of the course.

REFERENCES

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st-century higher education:Aa review and synthesis. Telematics and Informatics, 37, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Aljohani, N. R., Daud, A., Abbasi, R. A., Alowibdi, J. S., Basheri, M., & Aslam, M. A. (2019). An integrated framework for course adapted Student Learning Analytics Dashboard. Computers in Human Behavior, 92, 679–690. https://doi.org/10.1016/j.chb.2018.03.035

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1). https://doi.org/10.1186/s41239-020-0177-7

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK 12. https://doi.org/10.1145/2330601.2330666

Azizah, E. N., Pujianto, U., & Nugraha, E., (2018). Comparative performance between C4. 5 and Naive Bayes classifiers in predicting student academic performance in a virtual learning environment. 2018 4th international conference on education and technology (ICET), IEEE (2018), pp. 18-22

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Bresfelean, V. P. (2007). Analysis and predictions on students' behavior using decision trees in Weka Environment. 2007 29th International Conference on Information Technology Interfaces. https://doi.org/10.1109/iti.2007.4283743

Britain, S., & Liber, O., (1999). A framework for pedagogical evaluation of virtual learning environments. JISC Technology Applications Programme (Report 41).

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad et al. (Eds). Advances in Knowledge Discovery and Data Mining, MIT Press. (pp. 1–34).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The Weka Data Mining Software. ACM SIGKDD Explorations Newsletter, 11(1), 10–18. https://doi.org/10.1145/1656274.1656278

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. Proceedings of the Seventh International Learning Analytics & Knowledge Conference. https://doi.org/10.1145/3027385.3027449

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. Computational Intelligence and Neuroscience, 2018, 1–21. https://doi.org/10.1155/2018/6347186

Hussain, S., Abdulaziz Dahan, N., Ba-Alwi, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using Weka. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P., (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160(1), 3–24

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. Scientific data, 4(1), 1-8.

Kuzilek, J., Vaclavek, J., Fuglik, V., & Zdrahal, Z. (2018). Student drop-out modelling using virtual learning environment behaviour data. Lifelong Technology-Enhanced Learning, 166–171. https://doi.org/10.1007/978-3-319-98572-5_13

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers & Education, 53(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

Marbouti, F., Diefes-Dux, H., & Strobel, J. (2015). Building course-specific regression-based models to identify at-risk students. 2015 ASEE Annual Conference and Exposition Proceedings. https://doi.org/10.18260/p.23643

Peach, R. L., Yaliraki, S. N., Lefevre, D., & Barahona, M. (2019). Data-driven unsupervised clustering of online learner behaviour. NPJ Science of Learning, 4(1). https://doi.org/10.1038/s41539-019-0054-0

Roy, S., & Garg, A. (2017). Analyzing performance of students by using data mining techniques a literature survey. 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). https://doi.org/10.1109/upcon.2017.8251035

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning: A decision tree based approach. Computers & Education, 137, 32–47. https://doi.org/10.1016/j.compedu.2019.04.001

Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. Lecture Notes in Computer Science, 534–536. https://doi.org/10.1007/978-3-642-21869-9_93

Schleicher, A. (2019). Pisa 2018: Insights and interpretations. OECD.

Summers, R. J., Higson, H. E., & Moores, E. (2020). Measures of engagement in the first three weeks of higher education predict subsequent activity and attainment in first year undergraduate students: A UK case study. Assessment & Evaluation in Higher Education, 46(5), 821–836. https://doi.org/10.1080/02602938.2020.1822282

Swets, J. A. (1988). Measuring the accuracy of Diagnostic Systems. Science, 240(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE Big Data using Deep Learning Models. Computers in Human Behavior, 104, 106189. https://doi.org/10.1016/j.chb.2019.106189

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry, 39(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561

APPENDIX A

Table 4 The Accuracy of Predict Student Final Grade

	Personal & "Week 0"			"Week 0"			Personal		
	DT	KNN	RF	DT	KNN	RF	DT	KNN	RF
AAA	74%	75%	75%	76%	73%	74%	74%	73%	72%
BBB	68%	65%	69%	68%	63%	68%	68%	66%	68%
CCC	56%	53%	58%	55%	54%	53%	57%	53%	56%
DDD	59%	56%	61%	58%	55%	56%	59%	54%	58%
EEE	72%	72%	75%	75%	71%	72%	74%	72%	68%
FFF	74%	73%	73%	73%	72%	72%	73%	73%	73%
GGG	82%	83%	88%	82%	80%	85%	82%	84%	85%

APPENDIX B

Table 5 The Precision of Predict Student Final Grade

	Personal & "Week 0"			"Week 0"			Personal		
	DT	KNN	RF	DT	KNN	RF	DT	KNN	RF
AAA	77%	78%	77%	76%	78%	78%	77%	77%	77%
BBB	71%	72%	70%	68%	69%	69%	69%	72%	70%
CCC	58%	56%	58%	59%	56%	55%	57%	55%	58%
DDD	66%	61%	63%	59%	59%	59%	60%	58%	60%
EEE	76%	76%	76%	75%	75%	76%	75%	75%	75%
FFF	74%	75%	75%	74%	74%	74%	74%	75%	74%
GGG	83%	90%	88%	82%	87%	87%	83%	89%	88%

APPENDIX C

Table 6 The Recall of Predict Student Final Grade

	Personal & "Week 0"			"Week 0"			Personal		
	DT	KNN	RF	DT	KNN	RF	DT	KNN	RF
AAA	94%	95%	95%	99%	91%	92%	95%	93%	90%
BBB	90%	80%	95%	99%	85%	98%	98%	82%	93%
CCC	60%	48%	60%	46%	49%	57%	70%	50%	58%
DDD	61%	54%	68%	76%	53%	63%	72%	55%	67%
EEE	93%	91%	96%	98%	92%	92%	99%	93%	87%
FFF	98%	97%	97%	99%	97%	94%	99%	97%	97%
GGG	99%	90%	99%	100%	89%	96%	99%	92%	95%

APPENDIX D

Table 7 The Accuracy of Predict Student First Assignment Performance

	Personal & "Week 0"			"Week 0"			Personal		
	DT	KNN	RF	DT	KNN	RF	DT	KNN	RF
AAA	80%	82%	83%	82%	81%	81%	83%	81%	81%
BBB	97%	93%	96%	97%	93%	96%	97%	93%	96%
CCC	61%	58%	62%	60%	58%	61%	61%	58%	58%
DDD	58%	54%	58%	67%	64%	61%	56%	53%	51%
EEE	78%	76%	79%	79%	76%	74%	79%	76%	74%
FFF	89%	82%	89%	89%	81%	89%	89%	80%	89%
GGG	93%	89%	93%	93%	89%	92%	93%	89%	92%

APPENDIX E

Table 8 The Precision of Predict Student First Assignment Performance

	Personal & "Week 0"			"Week 0"			Personal		
	DT	KNN	RF	DT	KNN	RF	DT	KNN	RF
AAA	82%	84%	84%	84%	83%	83%	84%	83%	85%
BBB	97%	94%	97%	97%	97%	97%	97%	94%	97%
CCC	65%	66%	66%	64%	66%	66%	63%	65%	65%
DDD	61%	60%	61%	70%	68%	65%	59%	58%	56%
EEE	80%	79%	80%	79%	79%	79%	79%	79%	79%
FFF	90%	90%	90%	90%	90%	90%	90%	89%	90%
GGG	93%	94%	93%	93%	93%	93%	93%	93%	93%

APPENDIX F

Table 9 The Recall of Predict Student First Assignment Performance

	Personal & "Week 0"			"Week 0"			Personal		
	DT	KNN	RF	DT	KNN	RF	DT	KNN	RF
AAA	85%	97%	98%	99%	96%	97%	99%	97%	94%
BBB	100%	96%	97%	100%	96%	99%	100%	97%	99%
CCC	80%	69%	82%	81%	68%	76%	89%	72%	74%
DDD	71%	54%	69%	72%	65%	67%	75%	56%	58%
EEE	97%	95%	98%	99%	95%	91%	100%	95%	91%
FFF	100%	90%	99%	100%	90%	99%	100%	89%	99%
GGG	97%	95%	97%	99%	95%	98%	100%	95%	99%