International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Mar-2022)

https://dx.doi.org/10.12785/ijcds/110191

Using Machine Learning to Predict the Low Grade Risk for Students based on Log File in Moodle Learning Management System

Anh Thi-Diem Nguyen[1]

[1]*Faculty of Information Technology Faculty of Information Technology, School of Engineering and Technology, Van Lang University, Ho Chi Minh City, Vietnam*

Received 10 Jun. 2021, Revised 21 Nov. 2021, Accepted 3 Feb. 2022, Published 31 Mar. 2022

Abstract: The ever-growing demand for online teaching and studying requires the deployment of an effective online learning model. With diminishing face to face learning opportunities, schools are now heavily invested in improving student's sense of active learning and ultimately reduce the rate of student failures and dropouts. Currently, researchers have aimed to build a solution to analyze learners' behavior from the data collected through the online learning site - Moodle LMS and use the Linear Regression algorithm to predict the learning outcomes upon the completion of a student's course. The purpose of the study is to provide lecturers with a set criterion to standardize student learning outcomes during in the teaching process. On that basis, the lecturer can filter out the list of students who are at risk of failing the subject, and promptly warn students to change their learning attitude more actively, so that students can achieve satisfactory results. at the end of the course, thereby reducing the rate of students failing and dropping out of school.

Keywords: Machine Learning, Linear Regression, Learning Management System, Log File

1. INTRODUCTION

The world is facing the global COVID-19 epidemic, so the need for online learning is very necessary, limiting concentrated learning and supporting social distancing. Along with the current rapid development of information technology, the establishment of educational curriculums that utilizes the E-learning models at universities around the world is becoming more and more popular.

In addition to improving the quality of education, an equally important activity is to reduce the rate of students dropping out of school, early warning students at risk of failing subjects to help them adjust their attitudes. timely learning, increase the rate of effective student learning and improve the success level of the E-learning model. Through the current centralized education method, the lecturer can only grasp a student's learning situation through tests and it takes a lot of time to compile a list of students at risk of failing a subject. But for the online learning model, learners' behavior is recorded on the log file of the Moodle LMS system, which is an invaluable resource that researchers can correlate and include in their analysis. The specific focus of this article will be on the statistics of all student activities in training courses through the use of the Moodle LMs log file research data and finding the correlation between these activities and the final learning outcomes of each student. This will assist teachers in establishing rules that can provide an indication of each student's progress, thus give early warning to students who are at risk of failing a subject, help teachers actively monitor and encourage students to be more active in their studies. This study, contributes in reducing the rate of students failing subjects as well as indirectly reducing the number of students leaving school due to excessive debt of subjects.

2. RELATED WORK

Moodle LMS stores all user interactions in a log file, which is an opportunity for data mining researchers to create utilities that support the online education model. Many previous studies have shown to have exploited these log file data sources in many different aspects and approaches. The studies apply data visualization methods and use data analysis tools to produce predictive results such as [1], research MOCLog research has developed a Learning Management system that provides some reporting tools to track student learning. Using a learning management system log file can help determine who was active in the course, what they did, and when they did it. Beer and Clarke's [2], independent research and collection of log file data from CQ University's Moodle, hereby compares the level of student participation before and after the University adopts Moodle as a system. Their sole learning management has highlighted that student

E-mail address : anh.ntd@vlu.edu.vn

engagement in the course through the e-learning site has become the criterion for measuring the quality of teaching and learning at universities. Wambua's article ”Adopting course completion tracking and conditional activities to enhance engagement in elearning for university students”[3], the article used SPSS software to analyze the logfile data and presents the findings of the adoption of completion tracking and conditional activities to enhance engagement in Moodle LMS. These results have significant implications on instructors conducting online classes and the development of student engagement for online courses. [4], the research is implemented via Python programs for the analysis of the behavior of learners in the LMS environment. This investigation means to watch the personal conduct standards of the learners in an e-learning condition and decide whether the conduct of the learners can influence the scholarly presentation.

And some studies have shown a positive correlation between learner behaviors and a student's results such as [5] in the study ”Using Learning Analytics to Predict Students Performance in Moodle LMS” which shows the correlation between student grades and file openings. This correlation has a positive meaning that students with a higher frequency of opening files will generally higher scores compared to those who didn't open files regularly. [6] conducted a study of an actual information technology course being taught and provided the following interesting results: Female students are often more actively engaged during the course and successful than their male counterparts. There was a correlation between the number of diary entries and the final results. [7], the research used the Rstudio tool to automate analysis and show relationships between variables obtained from Moodle logs (total visits, course views, file submissions, assignments, quizzes) with student scores during the study period. [8] using Excel Macros to analyze Moodle logs and visualize log file data, gives trainers an overview of similar individual student involvement in the course. [9] the article states that the “views” and “updates” activities of students and lecturers are highly correlated and have to use Excel to visualize this correlation, to evaluate the teaching effectiveness of the course.

In addition to the studies of data visualization for reporting and using tools to analyze data to make predictions, some studies have applied algorithms to analyze logfile data to obtain other related results. Positive results for practical application such as [10] in the research paper ”Analysis of Student Behavior and Success Based on Logs in Moodle ” used a vector space model to analyze data and simulate student activity levels through diagrams. Research results help Instructors track course progress and allow them to quickly identify students who are not performing well and adjust their pedagogical strategies accordingly. [11] in the article “Alignment of teacher's plan and students' use of LMS resources. Analysis of Moodle logs” used a Linear Temporal Logic based model to analyze log files from the Moodle site. Based on their research results, teachers can design courses more suitable for their students. [12] this study applied Educational Data Mining on Moodle Learning Management System (LMS) at an African University. Data collected from Moodle LMS was preprocessed and analyzed using machine learning algorithms of clustering, classification and visualization from the WEKA system tools. The findings indicated that there is significant relationship between the use of LMS resources and students' academics. These are useful for strategic academic planning purpose with LMS data at a university. [13], this study discusses the behavioral analysis of training participants in the data visualization course with Tableau on the Moodle LMS on the website of the Warung Kompetensi of Statistics Indonesia (Warkop BPS). The analytical methods used is k-means clustering. The results showed that the behavior of the trainees from this course could be divided into 3 groups or clusters based on the activities of the training participants. This study concludes that k-means clustering is able to provide information on the grouping of course participants' behavior through the LMS log data so that in the future they can see what interventions can be implemented to increase the learning enthusiasm of the training participants.

[14] learning outcomes in a C programming language course in the early stage of a semester according to the log files of the MOODLE LMS system to build models and prediction results for early learning warning. A hybrid classification decision mechanism is proposed to combine the results of different predictions based on the accumulated training cases to further improve the accuracy of prediction.

[15] Research objective is to detect at-risk, fail and excellent students in the early stages of the course. And used different classication models for each of those three student groups. Decision tree, naive Bayes, logistic regression, multilayer perceptron (MLP) neural network, and support vector machine models are created and evaluated.

3. DEFINITIONS

A. Linear regression algorithm by least squares

Simple Linear Regression Analysis is to find the relationship between two continuous variables: the independent variable (the predictor) on the x-axis with the dependent variable (the outcome variable) on the vertical axis. y. Then draw a regression line and from this line equation we can predict the variable y in general form [16]

$$y = w_0 + w_1 x \qquad (1)$$

Where x is the predictor variable, y is the value predicted with the corresponding x value (response variable). To determine the values of w0 and w1, we use the least squares formula to get a most suitable straight line. The calculations w0 and w1 are as follows:

$$w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2} w_0 = \bar{y} - w_1 \bar{x} \quad (2)$$

For example, the study will model the relationship between the point X (the total number of times the student interacts with the course) and the student's final grade Y score obtained by the linear regression model. The objective is to find out the relationship between the X point and the student's Y score, the study uses the scatter plot as a tool in determining the degree of association between the two variables included in the model. From the linear regression model, the corresponding Y score of the student will be predicted based on the X point

B. Overview of the Moodle LMS platform

Moodle LMS is a technology platform that supports the E-Learning online learning model, which has brought many utilities to help students and lecturers have more diverse and easier interactive activities [17] Specifically, it allows instructors to create online classes, post assignments and assessments, calculate grades, etc. Students can access classes and online resources, complete assignments, and communicate with faculty. Through the online learning site, instructors will choose activities that are interactive tools to engage students in learning and assess student progress, including activities and resources as shown in Fig 1

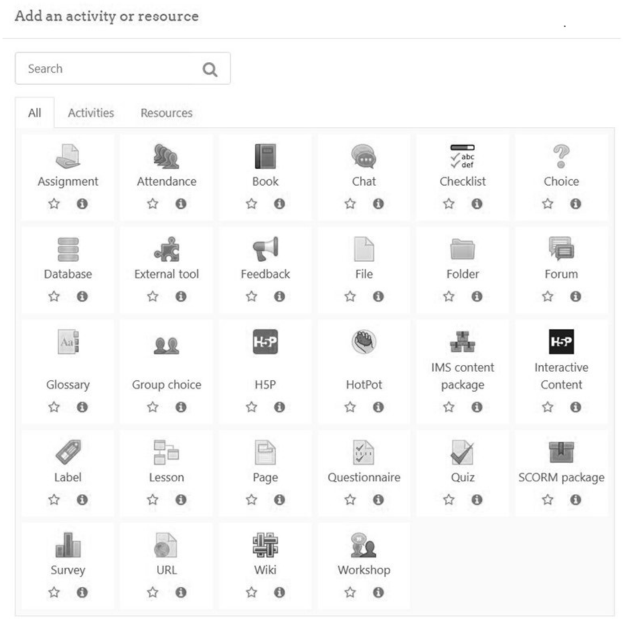


Figure 1. Activity and resource in Moodle LMS

C. Logfile – activity log

Besides convenient features for Instructors to easily create online courses, Moodle LMS also provides an extremely rich source of data stored in the system's database, which is an opportunity for Data mining researchers to create utilities to support an increasingly powerful and effective online education model. This data source is the logs file, which records all the relevant activities (Time, User name, Event contex, Component, Event name, Origin, IP address and etc..) of admins, lecturers, and students interacting with the course [18] during the training process of the course as shown in Fig 2

No	Component	Event name	Point
1	Assignment	Submitted.	1.0
2	Quiz	submitted	1.0
3	Page	viewed	0.2
4	URL	viewed	0.2
5	File	viewed	0.5
6	Folder	downloaded	2.0

TABLE I. Point estimation table for interactive activities on the e-learning site

D. The structure of the courses used in the study

The overall structure of courses built on the e-learning site that the study included for analysis is as follows:

At the beginning of each course, the instructor provides general information about the subject and learning materials for the whole course, through the following Moodle LMS activity selection:

Folder – Storage learning materials

File - Learning materials

Each session, the lecturer provides materials and means of student assessment:

File - Learning materials

Page – Video lectures

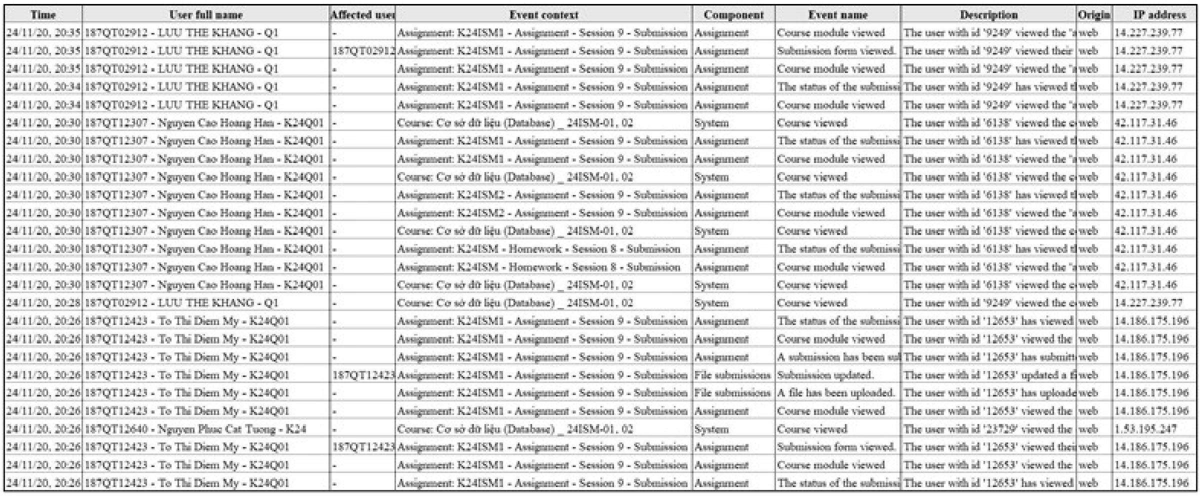
URL - Link to the lecture video

Assignment – Allows students to submit assignments they have done

Quiz - Assess the level of students' receptivity through the form of multiple-choice questions

The training program follows the credit regulations and the GPA and at the end of the course, if the score is less than 4.0, the subject the resultant score is a fail.

The influence of attributes: based on the subject assessment regulations in the detailed outline of the subjects Database – K25PM, Advanced Database System – K24PM, Requesting Technique – K25PM, Introduction to Software Engineering - K26PM, research to determine the data attributes (student's activities) that affect the final learning results, then calculate the point for each data's attribute in Table I



Time	User full name	Affected use	Event context	Component	Event name	Description	Origin	IP address
24/11/20, 20:35	187QT02912 - LUU THE KHANG - Q1	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '9249' viewed the '	web	14.227.239.77
24/11/20, 20:35	187QT02912 - LUU THE KHANG - Q1	187QT02912	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Submission form viewed.	The user with id '9249' viewed their	web	14.227.239.77
24/11/20, 20:35	187QT02912 - LUU THE KHANG - Q1	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '9249' viewed the '	web	14.227.239.77
24/11/20, 20:34	187QT02912 - LUU THE KHANG - Q1	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	The status of the submissi	The user with id '9249' has viewed t	web	14.227.239.77
24/11/20, 20:34	187QT02912 - LUU THE KHANG - Q1	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '9249' viewed the '	web	14.227.239.77
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Course: Cơ sở dữ liệu (Database) _ 24ISM-01, 02	System	Course viewed	The user with id '6138' viewed the c	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	The status of the submissi	The user with id '6138' has viewed t	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '6138' viewed the '	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Course: Cơ sở dữ liệu (Database) _ 24ISM-01, 02	System	Course viewed	The user with id '6138' viewed the c	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Assignment: K24ISM2 - Assignment - Session 9 - Submission	Assignment	The status of the submissi	The user with id '6138' has viewed t	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Assignment: K24ISM2 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '6138' viewed the '	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Course: Cơ sở dữ liệu (Database) _ 24ISM-01, 02	System	Course viewed	The user with id '6138' viewed the c	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Assignment: K24ISM - Homework - Session 8 - Submission	Assignment	The status of the submissi	The user with id '6138' has viewed t	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Assignment: K24ISM - Homework - Session 8 - Submission	Assignment	Course module viewed	The user with id '6138' viewed the '	web	42.117.31.46
24/11/20, 20:30	187QT12307 - Nguyen Cao Hoang Han - K24Q01	-	Course: Cơ sở dữ liệu (Database) _ 24ISM-01, 02	System	Course viewed	The user with id '6138' viewed the c	web	42.117.31.46
24/11/20, 20:28	187QT02912 - LUU THE KHANG - Q1	-	Course: Cơ sở dữ liệu (Database) _ 24ISM-01, 02	System	Course viewed	The user with id '9249' viewed the c	web	14.227.239.77
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	The status of the submissi	The user with id '12653' has viewed	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '12653' viewed the	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	A submission has been su	The user with id '12653' has submitt	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	187QT12423	Assignment: K24ISM1 - Assignment - Session 9 - Submission	File submissions	Submission updated.	The user with id '12653' updated a f	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	File submissions	A file has been uploaded.	The user with id '12653' has uploade	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '12653' viewed the	web	14.186.175.196
24/11/20, 20:26	187QT12640 - Nguyen Phuc Cat Tuong - K24	-	Course: Cơ sở dữ liệu (Database) _ 24ISM-01, 02	System	Course viewed	The user with id '23729' viewed the	web	1.53.195.247
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	187QT12423	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Submission form viewed.	The user with id '12653' viewed thei	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	Course module viewed	The user with id '12653' viewed the	web	14.186.175.196
24/11/20, 20:26	187QT12423 - To Thi Diem My - K24Q01	-	Assignment: K24ISM1 - Assignment - Session 9 - Submission	Assignment	The status of the submissi	The user with id '12653' has viewed	web	14.186.175.196

Figure 2. Log file of the course Database on the online learning page

4. METHODOLOGY

This study uses log file data of Van Lang University's online learning site, along with detailed outlines and actual summary transcripts of Database subjects - K25PM with 103 students, Advanced Database System – K24PM with 139 students, Requirements Engineering – K25PM with 87 students , Introduction to Software Engineering – K26PM with 107 students. According to the [10] approach, the research deploys sequentially in 3 phases:

Data collection and normalization

Application mining algorithm

Result evaluation.

A similar process to the few machine learning-oriented studies that have used the Python programming language to program and apply algorithms to make predictions like [9], [15] is utilised. In the process of extraction and analysis, the study uses the Python programming language to simulate data and applies the Linear Regression algorithm to predict the result. The reason for using this methods are as follows:

Python language is the most in-demand programming language used for AI and it offers a significant array of choices in the available libraries. In this research, I used two libraries: Scikit-learn libraries to manage linear regressions algorithm; Matplotib libraries to develop 2D charts and data visualisation.

Linear regression algorithm is often used to show a linear relationship between a dependent variable and one or more independent variables. Therefore, this algorithm is perfectly suitable to find the rule to predict the total interactive point of students on the online learning site (dependent variable) to achieve so that the final result of the course will above 4.0 (independent variable).

From the experimental results of the algorithm, the research will calculate the percentage of the total interactive point of students on the online learning site that the result of the course is at 4.0.

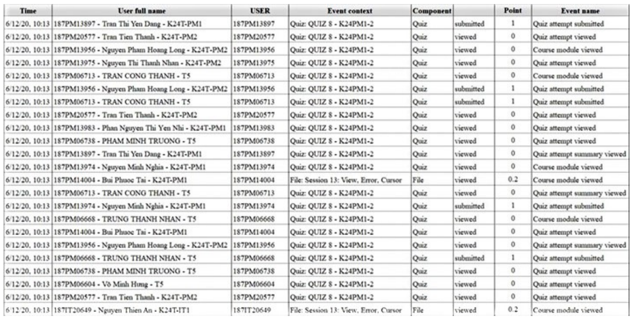
A. *Collect and normalize data*

Collect data:

Research uses log file data of 4 training courses taught in the first semester of 2020 - 2021 on the online learning website of Van Lang University at [19] and transcripts of those 4 Training Courses. Therefore, the data is highly reliable and accurate for analysis and prediction.

Normalize data:

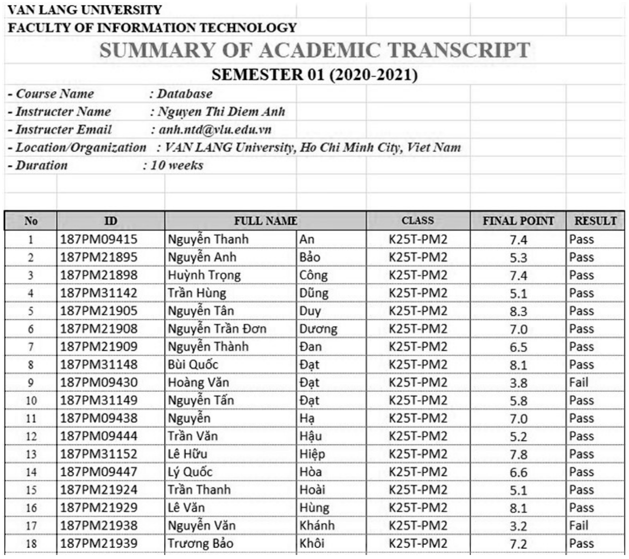
Applying method of extracting the data fields needed for the study from the log file of the 4 training courses. Eliminating confounding data streams such as log data lines made by faculty and administrators, keeping only data lines storing actions of students who have interacted in the course. Remove duplicate records. Assigning points to each student's action based on Table 1 Converting the normalized data file into a csv format named "Event.csv" which will contain the data attributes provided in Fig 3



Time	User full name	USER	Event context	Component		Point	Event name
6/12/20, 10:13	187PM13897 - Tran Thi Yen Dang - K24T-PM1	187PM13897	Quiz: QUIZ 8 - K24PM1-2	Quiz	submitted	1	Quiz attempt submitted
6/12/20, 10:13	187PM20577 - Tran Tien Thanh - K24T-PM2	187PM20577	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM13956 - Nguyen Pham Hoang Long - K24T-PM2	187PM13956	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Course module viewed
6/12/20, 10:13	187PM13975 - Nguyen Thi Thanh Nhan - K24T-PM2	187PM13975	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM06713 - TRAN CONG THANH - T5	187PM06713	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Course module viewed
6/12/20, 10:13	187PM13956 - Nguyen Pham Hoang Long - K24T-PM2	187PM13956	Quiz: QUIZ 8 - K24PM1-2	Quiz	submitted	1	Quiz attempt submitted
6/12/20, 10:13	187PM06713 - TRAN CONG THANH - T5	187PM06713	Quiz: QUIZ 8 - K24PM1-2	Quiz	submitted	1	Quiz attempt submitted
6/12/20, 10:13	187PM20577 - Tran Tien Thanh - K24T-PM2	187PM20577	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM13983 - Phan Nguyen Thi Yen Nhi - K24T-PM1	187PM13983	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM06738 - PHAM MINH TRUONG - T5	187PM06738	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM13897 - Tran Thi Yen Dang - K24T-PM1	187PM13897	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt summary viewed
6/12/20, 10:13	187PM13974 - Nguyen Minh Nghia - K24T-PM1	187PM13974	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Course module viewed
6/12/20, 10:13	187PM14004 - Bui Phuoc Tai - K24T-PM1	187PM14004	File: Session 13: View, Error, Cursor	File	viewed	0.2	Course module viewed
6/12/20, 10:13	187PM06713 - TRAN CONG THANH - T5	187PM06713	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt summary viewed
6/12/20, 10:13	187PM13974 - Nguyen Minh Nghia - K24T-PM1	187PM13974	Quiz: QUIZ 8 - K24PM1-2	Quiz	submitted	1	Quiz attempt submitted
6/12/20, 10:13	187PM06668 - TRUNG THANH NHAN - T5	187PM06668	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Course module viewed
6/12/20, 10:13	187PM14004 - Bui Phuoc Tai - K24T-PM1	187PM14004	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM13956 - Nguyen Pham Hoang Long - K24T-PM2	187PM13956	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt summary viewed
6/12/20, 10:13	187PM06668 - TRUNG THANH NHAN - T5	187PM06668	Quiz: QUIZ 8 - K24PM1-2	Quiz	submitted	1	Quiz attempt submitted
6/12/20, 10:13	187PM06738 - PHAM MINH TRUONG - T5	187PM06738	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM06604 - Võ Minh Hưng - T5	187PM06604	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187PM20577 - Tran Tien Thanh - K24T-PM2	187PM20577	Quiz: QUIZ 8 - K24PM1-2	Quiz	viewed	0	Quiz attempt viewed
6/12/20, 10:13	187IT20649 - Nguyen Thien An - K24T-IT1	187IT20649	File: Session 13: View, Error, Cursor	File	viewed	0.2	Course module viewed

Figure 3. Event.csv file data has been normalized

Integrating the new file (Event.csv) with the student's average score file (Fig.4 Score.csv) into a complete data table for analysis. (Fig.5 Data for visualization)



VAN LANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

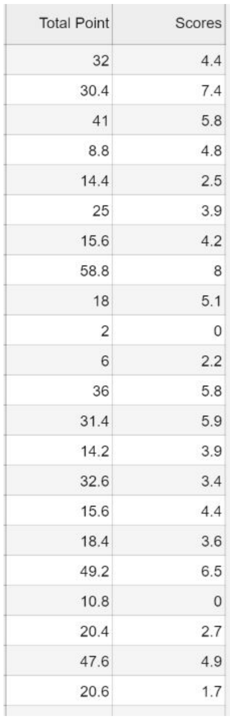
SUMMARY OF ACADEMIC TRANSCRIPT

SEMESTER 01 (2020-2021)

- Course Name : Database
- Instructer Name : Nguyen Thi Diem Anh
- Instructer Email : anh.ntd@vlu.edu.vn
- Location/Organization : VAN LANG University, Ho Chi Minh City, Viet Nam
- Duration : 10 weeks

No	ID	FULL NAME		CLASS	FINAL POINT	RESULT
1	187PM09415	Nguyễn Thanh	An	K25T-PM2	7.4	Pass
2	187PM21895	Nguyễn Anh	Bảo	K25T-PM2	5.3	Pass
3	187PM21898	Huỳnh Trọng	Công	K25T-PM2	7.4	Pass
4	187PM31142	Trần Hùng	Dũng	K25T-PM2	5.1	Pass
5	187PM21905	Nguyễn Tân	Duy	K25T-PM2	8.3	Pass
6	187PM21908	Nguyễn Trần Đơn	Dương	K25T-PM2	7.0	Pass
7	187PM21909	Nguyễn Thành	Đan	K25T-PM2	6.5	Pass
8	187PM31148	Bùi Quốc	Đạt	K25T-PM2	8.1	Pass
9	187PM09430	Hoàng Văn	Đạt	K25T-PM2	3.8	Fail
10	187PM31149	Nguyễn Tấn	Đạt	K25T-PM2	5.8	Pass
11	187PM09438	Nguyễn	Hạ	K25T-PM2	7.0	Pass
12	187PM09444	Trần Văn	Hậu	K25T-PM2	5.2	Pass
13	187PM31152	Lê Hữu	Hiệp	K25T-PM2	7.8	Pass
14	187PM09447	Lý Quốc	Hòa	K25T-PM2	6.6	Pass
15	187PM21924	Trần Thanh	Hoài	K25T-PM2	5.1	Pass
16	187PM21929	Lê Văn	Hùng	K25T-PM2	8.1	Pass
17	187PM21938	Nguyễn Văn	Khánh	K25T-PM2	3.2	Fail
18	187PM21939	Trương Bảo	Khôi	K25T-PM2	7.2	Pass

Figure 4. ”Score.csv” file the student's average score



Total Point	Scores
32	4.4
30.4	7.4
41	5.8
8.8	4.8
14.4	2.5
25	3.9
15.6	4.2
58.8	8
18	5.1
2	0
6	2.2
36	5.8
31.4	5.9
14.2	3.9
32.6	3.4
15.6	4.4
18.4	3.6
49.2	6.5
10.8	0
20.4	2.7
47.6	4.9
20.6	1.7

Figure 5. Data for visualization.csv

B. *Applying algorithms to data mining*

Research using Python programming language to run algorithms and simulate data using graphs to find the correlation between student activities on the E-learning page and the final score of the subjects in Fig. 6, Fig. 7, Fig. 8, Fig. 9.

Based on the results of the above 4 simulations, the research find that there is a correlation between the total point of students' interactions on the online learning site with the final results, so the research continues to apply the linear regression algorithm to give a linear equation (the closest line to the points on the graph). From the linear equation, the research can calculate the total points students need to achieve to have a final result of 4.0 (the minimum result to pass the subject). The results are as shown in Fig 10, Fig 11, Fig 12, Fig 13

5. EXPERIMENT

A. In this section, we showed the results that were implemented the linear regression algorithm on log files.

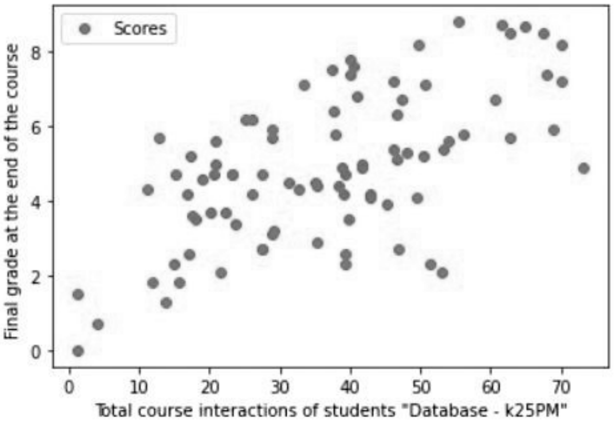


Figure 6

In Figure 6, the graph of the relationship between the total point of the activity and the average score of each student in the course Database - K25PM (87 data lines)

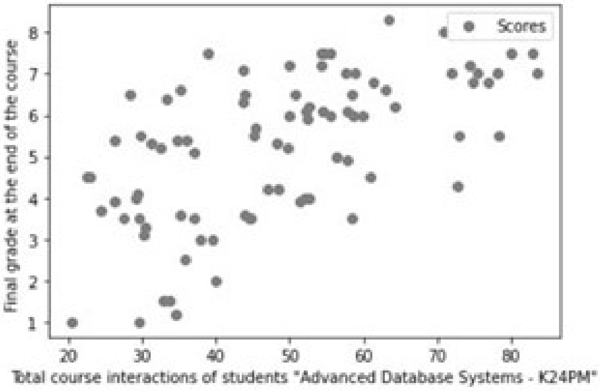


Figure 7

In Figure 7, the graph of the relationship between the total point of the activity and the average score of each student in the course Advanced Database System – K24PM (139 data lines)

In Figure 8, the graph of the relationship between the total point of the activity and the average score of each

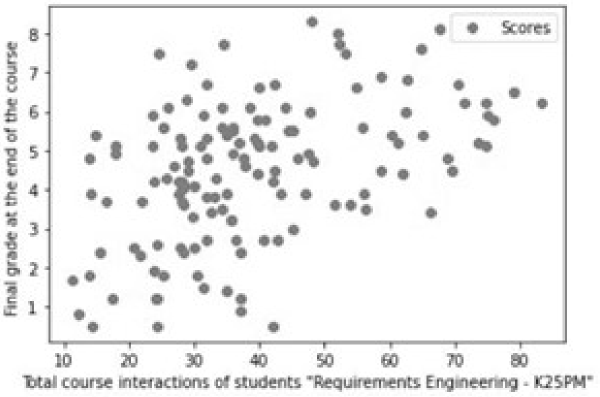


Figure 8

student in the course Requirements Engineering – K25PM (87 data lines)

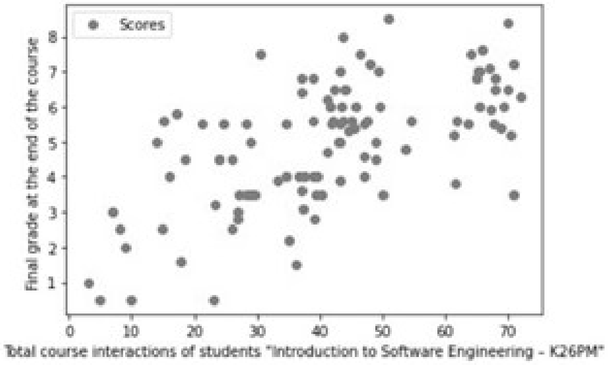


Figure 9

In Figure 9, the graph of the relationship between the total point of the activity and the average score of each student in the course Introduction to Software Engineering – K26PM (107 data lines)

We experimented the linear regression model (Equation 1, 2) to prediction results at the end of each course such as:

For the Database course – K25PM (Fig. 10)

The linear regression to find is: y = 2.23 + 0.07 x

With the requirement that student achieve a final result of 4.0 (variable y), the total student interaction point on the online learning site must be 25 (variable x)

Correspondingly, if the total student interaction point is less than 35 percent, the final final score will be less than 4.0 (fail).

For the Advanced Database Systems course – K24PM (Fig. 11)

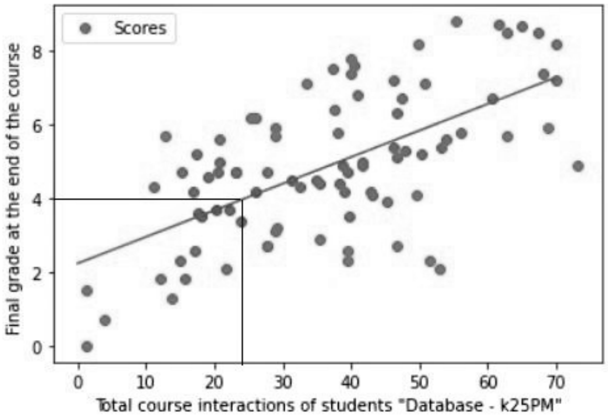


Figure 10. Linear regression line simulation graph with the Database course – K25PM data

The linear regression to find is: y = 2.069 + 0.064 x

With the requirement that student achieve a final result of 4.0 (variable y), the total student interaction point on the online learning site must be 30 (variable x)

If the student has a total interaction point for the whole course of 30, the predicted result will be 4.0

Correspondingly, if the total student interaction point is less than 35 percent, the final final score will be less than 4.0 (fail).

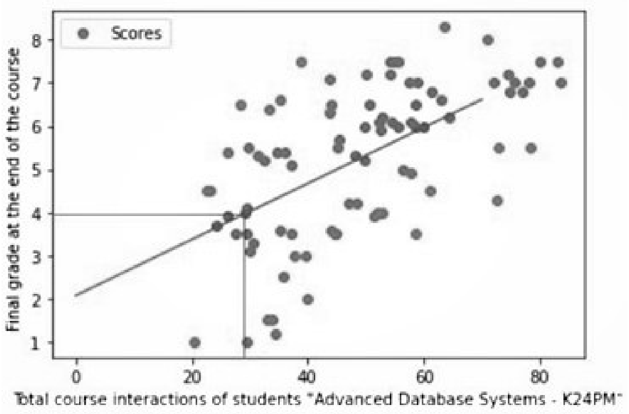


Figure 11. Linear regression line simulation graph with the Advanced Database Systems course – K24PM

For the Requirements Engineering course – K25PM (Fig .12)

The linear regression to find is: y = 2.535 + 0.048 x

With the requirement that student achieve a final result of 4.0 (variable y), the total student interaction point on the online learning site must be 30 (variable x)

Correspondingly, if the total student interaction point is less than 36 percent, the final score will be less than 4.0 (fail).

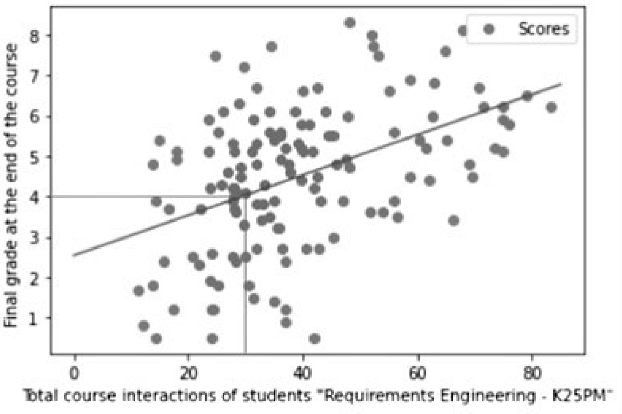


Figure 12. Linear regression line simulation graph with the Requirements Engineering course – K25PM

For the Introduction to Software Engineering course - K26PM (Fig. 13)

The linear regression to find is: y = 2.361 + 0.062 x

With the requirement that student achieve a result of 4.0 (variable y), the total student interaction point on the online learning site must be 26 (variable x)

Correspondingly, if the total student interaction point is less than 37 percent, the final score will be less than 4.0 (fail).

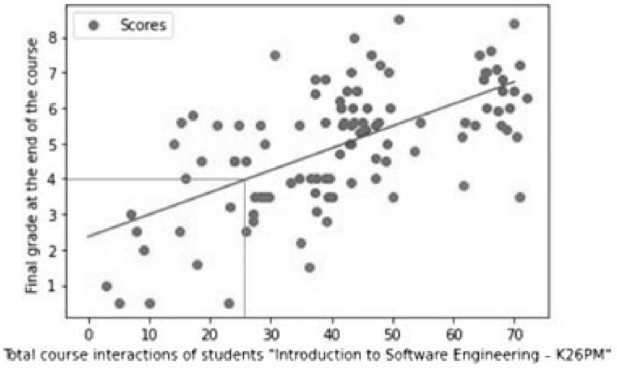


Figure 13. Linear regression line simulation graph with the Introduction to Software Engineering course - K26PM

B. Result Evaluation

Through a survey of 4 courses taught in practice in the first semester of the academic year 2020-2021 at the Faculty of Information Technology, Van Lang University, the study recorded the following specific results:

Database Course - K25PM: the total point of student interaction is less than 35 percent, the final score will be less than 4.0 (fail the subject)

Advanced Database System course - K24PM: the total point of student interaction is less than 35 percent, the final score will be less than 4.0 (fail the subject)

Requirements Engineering Course – K25PM: the total point of student interaction is less than 36 percent, the final score will be less than 4.0 (fail the subject)

Course Introduction to Software Engineering – K26PM: the total point of student interaction is less than 37 percent, the final score will be less than 4.0 (failure)

From the above analysis results, the study draws a general rule: With online training courses designed according to a common structure, during the learning process, students have little interaction on the online learning site (the total score of interactive activities on the e-learning site is less than 37 percent of the total maximum points of interactive activities calculated at the present time) there is a risk of failing the course. The lecturer can filter out this list of students to include in the list of students at risk of failing the course.

The results of this study are consistent with the results of similar published works such as [2], [5], [7], [9]. [4]. All studies have shown that there is a correlation between students' performance on e-learning websites and students' final results. And similar to the machine learning-oriented research of [15], using some algorithms to predict early student learning outcomes in the early stages of the course.

The differentiating focus of this research compared to previous studies is that it has given a specific criterion so that in the teaching process, teachers can easily filter out the list of students at risk of failing the subject.

6. MANAGERIAL IMPLICATIONS

Based on the criteria for predicting students at risk of failing a subject that the research brings, from the middle of the semester, the lecturers of the Faculty of Information Technology at Van Lang University can filter out a list of students at risk of failing the subject sent to the General Department. The Faculty will make statistics of students' academic performance in all subjects in the semester, promptly remind students and support them to improve their sense of active learning, contributing to reducing the annual dropout rate of students.

7. FUTURE RESEARCH

Data mining on online learning sites is very popular today and there have been many mining research works that bring many effective applications to society. The research work of this paper uses course data with the same general structure (weekly materials, video lectures, quizzes, exercises), so some data related to other student activities are ignored such as forum discussions, quiz results, assignment results, etc.

Although the scope of the research has been applied in Van Lang University, the results have contributed signifi-