Student Performance Prediction using Online Behavior Discussion Forum with Data Mining Techniques

Febrianti Widyahastuti[1] and Viany Utami Tjhin[2]
[1]*School of Information Technology, Deakin University, 221 Burwood Hwy, Burwood, Victoria 3125, Australia*
[2]*Department of Information Systems, Bina Nusantara University, Jakarta, Indonesia*

Keywords: E-learning, prediction, students' performance, Education Data Mining, Classification

Abstract: This paper aims to analytically compare and evaluate the students' performance results by applying several classification techniques in WEKA based on frequency students' using online discussion forum. We then compare and evaluate the performance results from five classification techniques, there are linear regression, multilayer perceptron, random forest, IBK and Kstar, which give the best results in terms of accuracy, performance and error. The basis of the data was derived from extraction and analysis of distance learning students' (202 undergraduate students majoring in English Professional in one of private university in Indonesia) e-learning logged- post in discussion forum and attendance. Based on the result, it has been concluded that using Linear Regression technique provides best predication results of final examination. Finally, we apply Linear Regression as benchmarking of data mining techniques to predict students' performance and interpret the results that the features in online discussion forum can predict students' performance.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

E-learning system provide a huge amount of students' activities in the course. The more students' active in E-learning could possibly improve their academic performance. E-learning is the major challenge for researcher that interest in educational sector by explosive growth of educational data and to use this data to improve students' performance. Online discussion forum is part of E-learning data, we used the features like login, posting, participation and attendance. The assessment of students' grade is based on quizzes, assignments, examinations and activities during their course.

It is important to use special tools to analyse and reveal hidden patterns from E-learning datasets. One of the best prediction tools to analysing data in E-learning is Data mining. Data mining focuses on educational data called as Educational Data Mining (Romero & Ventura, 2010). The main objective of Educational data mining is to analyse the different types of data by applying data mining methods to solve the educational problems (Romero, 2010). This will help educational institutions to plan, evaluate, improve and decide their programs.

Classification is the most popular data mining techniques used for predicting student performance (Romero 2007, Baker 2009). There are several algorithms in the classification techniques that have been applied to predict students' performance. We propose the use of various classification techniques that easily interpretable models like linear regression, multilayer perceptron, random forest, IBK and K-star. Those techniques will be applied using WEKA tool to investigate which technique is the best for predicting students' performance in online discussion forum. Some features in online discussion forum such as gender, login, students' posting, students' participation, attendance and grade are potential as predictors to predict students' performance in final examination.

Accordingly, this paper focus on some features in online discussion forum to predict students' performance using various classification techniques to address two research questions:

- Are those particular features in online discussion forum that significantly impact students' performance?
- Which classification techniques is the best to predict students' performance?

Widyahastuti, F. and Tjhin, V.
Student Performance Prediction using Online Behavior Discussion Forum with Data Mining Techniques.
DOI: 10.5220/0009017000002297
In *Proceedings of the Borneo International Conference on Education and Social Sciences (BICESS 2018)*, pages 90-95
ISBN: 978-989-758-470-1
Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

2 RELATED WORK

Research on students' performance prediction have been studied from various attributes in students' environment such as students' behaviour, demographics, students' information, psychological and socio-economic. Each attributes consist of several elements used in the measurement. Firstly, The students' behaviours (Romero, Ventura & Garcia, 2008) are measured based on the frequency of students' access to each feature extracted from E-learning logged data. The students' demographics (Shahiri, 2015) include place of residence, hobbies, family size, employment and education of parents and others. Next, The students' information (Gašević et al 2016, Kabakchieva 2013), similar to admission or enrolment data, includes name, age, gender, all grades (Ahmad et al 2015, Lin 2012), attendance and skill (Mayilvaganan 2014). Psychological attributes (Sembiring et al 2011) concern with abilities, attitudes, behaviour and motivation; meanwhile, Socio-economic (Pradeep 2015, David 2015) describe about the socio-economic background of the student and family as well as their academic background.

There are various data mining techniques (Romero & Ventura 2007, Baker 2009) such as classification, regression, density estimation, clustering and relationship mining have been implemented in educational data research. Furthermore, Methods like Decision Trees (Pradeep & Thomas 2015), Bayesian Network (Sundar 2013), Naïve Bayes (Dominick 2014) and Neural Network (Jai & David 2014) have also been used in predicting students' performance and mining educational data.

3 METHODOLOGY

Figure 1 represents the methodology which is used to predict independence attribute using training data from correlation result which are analysing next by a classification algorithm. We propose the use of various classification techniques that easily interpretable models like linear regression, multilayer perceptron, random forest, IBK and K-star. Using WEKA as open software machine learning can provides several features of selection models. Finally these algorithms will be executed, validated, evaluated and compared the results in order to determine which one give the best result with high accuracy.

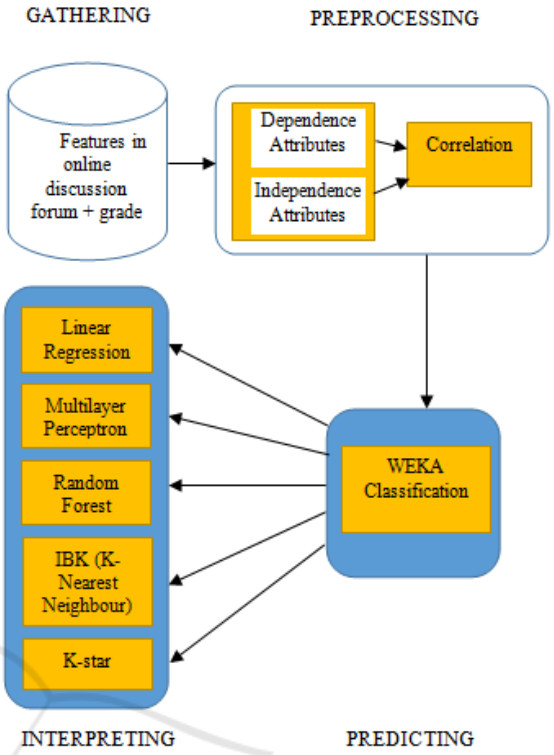

Figure 1: Methodology of Classification Techniques.

3.1 Gathering

Data gathering is the data may be obtained from many different data sources. This experiments are based online discussion forum dataset of 202 undergraduate students' majoring in Information System Management. Features in online discussion forum contain details from students' behaviour in E-learning. Table 1 shows features that consists of gender, login, students' posting, participate, attendance and final grade.

Table 1: Some Features in Online Discussion Forum.

Attributes	Description
Gender	Students' gender (Male / Female)
Login	The number of logins for each student to the online discussion forum in a semester
Posting	The number of postings for each student to the online discussion forum in a semester
FOD	The number of activation for each student to the online discussion forum in a semester

ATT	The attendance rate or percentage of each student to the online teaching session
Grade	8: 100-90 7: 89-80 6: 79-70 5: 69-60 4: 59-50 3: 49-40 2: < 40 1: 0

3.2 Pre-processing

Pre-processing of data is considered as a very important task in this work as we need quality and reliability of available information which directly affects the results attained. Before applying the data mining algorithms, it is important to carry out some pre- processing tasks such as data cleaning, integration, transformation and discretization.

We divided the attributes into dependent attributes and independent attributes. Independent attributes are the features in online discussion forum to be predictor. Grade is the selected dependent attribute to be predicted. We converted the grade into categorical values (8 to 1). Those attributes in online discussion forum and grade were studied and analysed to find out the main attributes or predictor that may affect the students' performance.

The Correlation analysis are to identify the dependent attribute and independent attribute if there is a significant relationship exist with academic results. The analysis done with the condition: If the sig. (2-tailed) value is less than or equal to 0.05, the correlation value is significant, or it can be proposed that there is a correlation. If the sig. (2-tailed) value is more than 0.05, the correlation value is no correlation between the two data.

Correlation coefficient describes the degree of connectedness between the actual value and the value predicted. The Range of correlation coefficient is between -1 and 1. If the coefficient value is 0, it means that there is no correlation. If it close to 1, it means that there is a positive relationship (if the actual value, the value of the predicted value is also a value) and vice-versa.

The correlations between grade as dependent variable and independent variables in online discussion forum (gender, login, posting, active and attendance) are displayed as in Table 2.

Table 2: Correlation analysis between features in online discussion forum and grade.

Attributes	Description
Gender	Students' gender (Male / Female)
Login	The number of logins for each student to the online discussion forum in a semester

	Variable	Correlation Coefficient	Sig. (2-tailed)	Result
Grade	Gender	0.047	0.511	No significant
	Student posting	0.151	0.032	Significant
	Login	0.400	0.000	Significant
	Forum active	0.307	0.000	Significant
	Attendance	0.778	0.000	Significant

3.3 Predicting

From the result of best predictor in correlation can process of classification models using WEKA. WEKA stands of Waikato Environment for Knowledge Analysis (Witten et al 1999, Hall et al 2009). WEKA is an open source of machine learning algorithm used for analysing and predicting students' performance of classification algorithms with binary variable which are, then, applied directly to a dataset.

We processed the pre-processing task in WEKA. As shown in Figure 2, Pre-processing task includes finding incorrect or missing data, removal of noise or outliers and collecting necessary information to model or account for noise. We train the data into 10-fold cross validation in WEKA.

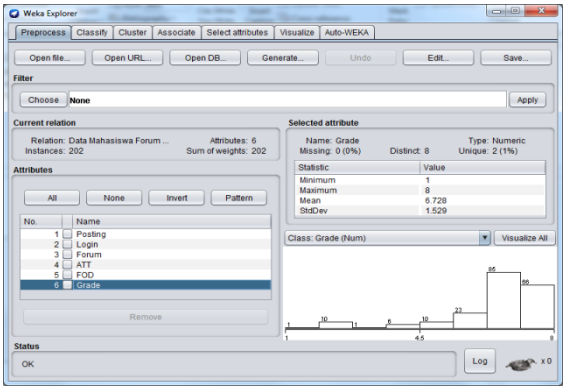

Figure 2: Pre-processing Data in WEKA.

3.4 Interpreting

We propose the use of various classification techniques that easily interpretable models like Linear Regression, Multilayer Perceptron, Random Forest, IBK and K-star. Using WEKA as open software machine learning can provides several features of selection models. Finally these algorithms will be executed, validated, evaluated and compared the results in order to determine which one give the best result with high accuracy.

3.4.1 Linear Regression

Linear regression is the best predication model to test the cause of one dependent variable (final grade) effect on one or more independent variables (features in online discussion forum). The initial judgement of a possible relationship between two continuous variables should always be made on the basis of a scatter plot (scatter graph).(Schneider et al 2010). Moreover, linear regression approach is quite easy and faster processing for large size datasets. The time to build this algorithm is 0.05 seconds. Below the result shows the formula of linear regression:

Grade = 0.4779 * attendance + 0.4614 * posts + 26.2719 (1)

3.4.2 Multilayer Perceptron

Multilayer perceptron is a supervised learning algorithm that uses the concept of neural network that interact using weighted connections. Each node will have a weight which, then, multiply the input node that generate the output predication. The Weight measure the degree of correlation between activity levels of neuron of which they connect. (Pal & Mitra, 1992). Generally, result from multilayer perceptron more accurate than linear regression but require a longer processing time for large datasets because the algorithm will always update the weight for each instance of the data. Thus, considering such factor, the disadvantage of Multilayer perceptron is sensitive to feature scaling (Pedregosa, 2011).

There are three hidden node labelled sigmoid node 1, 2 and 3. Attribute Posting, ATT and FOD seem to have nearly the same weight and sign in all the neurons. Below show the result of multilayer perceptron with the time taken to build 0.11 seconds:

Sigmoid Node 1
Inputs Weights
Threshold -0.3389339469178622
Attrib Posting 0.6356339310638692
Attrib Login -1.971194964716918
Attrib Forum 0.1528793652016145
Attrib ATT -2.9824012894200167
Attrib FOD -1.2565096616525258
Sigmoid Node 2
Inputs Weights
Threshold -0.3319752049637097
Attrib Posting 0.8489632795859472
Attrib Login 0.8981808286647163
Attrib Forum 1.1775792813836161
Attrib ATT -0.2727426863562934
Attrib FOD -1.4842188659857705
Sigmoid Node 3
Inputs Weights
Threshold -1.4238193757464874
Attrib Posting 2.516298013366708
Attrib Login 0.7532046884360826
Attrib Forum -0.15476793041226244
Attrib ATT -0.010654173314826458
Attrib FOD 2.257937779725289

3.4.3 Random Forest

The random forest was founded by Breiman in 2001 (Breiman 2001), as implemented in WEKA, is an ensemble of unpruned classification trees that use majority voting to perform prediction. The Random forest combines the predictions from classification trees using an algorithm similar to C4.5 (J48 in Weka). (Khoshgoftaar 2007).

3.4.4 IBK (K-Nearest Neighbour)

IBK is a k-nearest-neighbour classifier. It is also known as ‘lazy learning’ technique for the classifier construction process needs only little effort and, mostly, the work is performed along with the process of classification.(Khoshgoftaar 2007). Various combinations of search algorithms can be utilized to ease the task of finding the nearest neighbours. Normally, linear search is the most commonly used but there are other options which are also potential including KD-trees, ball trees and cover trees”. (Vijayarani & Muthulakshmi 2013)

Predictions made by considering more than one neighbour can be weighted based on the distance from the test instance and, then, the distance is converted into the weight by implemented two different formulas. (Vijayarani & Muthulakshmi 2013).

3.4.5 K-Star

K* algorithm is an instance-based learner using entropy to quantify the distance. It is considerably

beneficial because the elements therein contained providing a consistent approach in handling real valued attributes, symbolic attributes and missing values (Vijayarani & Muthulakshmi 2013).

4 RESULTS AND DISCUSSION

Using several classification algorithms such as Linear Regression, Multilayer Perceptron, Random Forest, IBK and KStar, the dataset is tested and analysed in WEKA. Those algorithms are verified by 10 fold cross validation check. The results show that linear regression are performing better than multilayer perceptron in terms of accuracy, performance and error. Even though the Multilayer perceptron have slight high 0.01% in correctly classified instance than linear regression. Accuracy of each classifier is shown in Table 3 and Figure 3. Performance and error are shown in Table 4 and Figure 4 as the result of mean absolute error and root mean squared error.

Table 3: Accuracy of correctly classification result.

Classification Algorithm	Correctly Classified Instance
Linear Regression	85.37%
Multilayer Perceptron	85.38%
Random Forest	79.66%
IBK	74.13%
KStar	77.68%

Table 4: Mean Absolute Error.

Classification Algorithm	Mean Absolute Error	Root Mean Squared Error
Linear Regression	0.6214	0.7947
Multilayer Perceptron	0.6323	0.8106
Random Forest	0.709	0.9296
IBK	0.7822	1.1158
KStar	0.758	0.9783

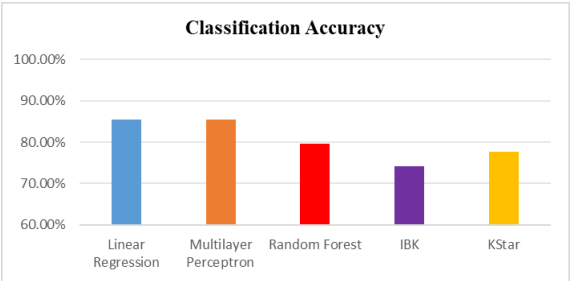

Figure 3: Graphical of Accuracy.

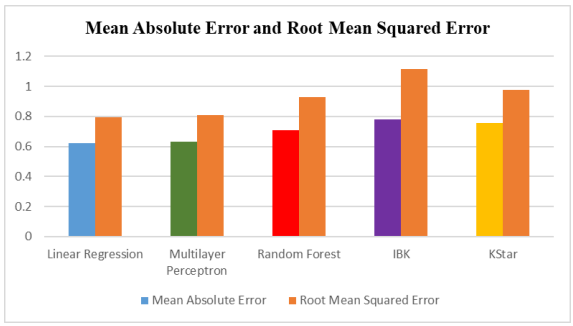

Figure 4: Graphical of Mean Absolute Error.

5 CONCLUSIONS

In this paper, final examination have been predicated using five classification algorithms namely Linear Regression, Multilayer Perceptron, Random Forest, IBK and Kstar and, then, are compared and evaluated which give best result based on the value of correctly classified instance, mean absolute error and root mean squared error.

The analysis of the experiment and the comparison of five classification algorithms has demonstrate evidence that Linear Regression is the most appropriate to predict students' performance result. The overall accuracy (85.37%) and overall error was extremely satisfactory (0.6214 and 0.7947). Moreover, the Linear Regression algorithm is the straightforward with formula and the best predication models in terms of accuracy, performance and error rate to compare their feasibility.

REFERENCES

Ahmad F, Ismail NH, Aziz AA. The prediction of students' academic performance using classification data mining techniques. Applied Mathematical Sciences. 2015;9(129):6415-26.

Baker RS, Yacef K. The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining. 2009;1(1):3-17.

Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

David L, Karanik M, Giovannini M, Pinto N. Academic Performance Profiles: A Descriptive Mode based on Data Mining. European Scientific Journal. 2015;11(9).

Dominick S, Razak TA. Analyzing the Student Performance using Classification Techniques to find the Better Suited Classifier. International Journal of Computer Applications. 2014;104(4).

Gašević D, Dawson S, Rogers T, Gasevic D. Learning analytics should not promote one size fits all: The

effects of instructional conditions in predicting academic success. The Internet and Higher Education. 2016;28:68-84.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10-8.

Jai R, K.David. Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study. IJRASET International Journal for Research in Applied Science & Engineering Technology. 2014;Vol. 2(Issue XI).

Kabakchieva D. Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies. 2013;13(1):61-72.

Khoshgoftaar TM, Golawala M, Van Hulse J, editors. An empirical study of learning from imbalanced data using random forest. Tools with Artificial Intelligence, 2007 ICTAI 2007 19th IEEE International Conference on; 2007: IEEE.

Lin S-H. Data mining for student retention management. Journal of Computing Sciences in Colleges. 2012;27(4):92-9.

Mayilvaganan M, Kalpanadevi D, editors. Comparison of classification techniques for predicting the performance of students academic environment. Communication and Network Technologies (ICCNT), 2014 International Conference on; 2014: IEEE.

Pal SK, Mitra S. Multilayer perceptron, fuzzy sets, and classification. IEEE Transactions on neural networks. 1992;3(5):683-97.

Pedregosa F, Ga, #235, Varoquaux l, Gramfort A, Michel V, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.

Pradeep A, Thomas J. Predicting College Students Dropout using EDM Techniques. International Journal of Computer Applications. 2015;123(5).

Romero C, & Ventura,S. Educational data mining: A review of the state of the art. IEEE Transactions on systems man and Cybernetics Part CApplications and review. 2010;40(6):601-18.

Romero C, Ventura S, García E. Data mining in course management systems: Moodle case study and tutorial. Computers & Education. 2008;51(1):368-84.

Romero C, Ventura S. Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 2010;40(6):601-18.

Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. Expert systems with applications. 2007;33(1):135-46.

Schneider A, Hommel G, Blettner M. Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. Deutsches Ärzteblatt International. 2010;107(44):776-82.

Sembiring S, Zarlis M, Hartama D, Ramliana S, Wani E, editors. Prediction of student academic performance by an application of data mining techniques. International Conference on Management and Artificial Intelligence IPEDR; 2011.

Shahiri AM, Husain W. A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science. 2015;72:414-22.

Sundar PP. A Comparative Study For Predicting Student's Academic Performance Using Bayesian Network Classifiers. IOSR Journal of Engineering (IOSRJEN) e-ISSN. 2013:2250-3021.

Vijayarani S, Muthulakshmi M. Comparative analysis of bayes and lazy classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering. 2013;2(8):3118-24.

Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Cunningham SJ. Weka: Practical machine learning tools and techniques with Java implementations. 1999.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS