 ISBN: 978-623-92207-1-6

Predicting Academic Performance In Blended Learning By Using Data Mining Classification Techniques

1st Dafid
Universitas Sriwijaya
Palembang, Indonesia
03013622025009@student.unsri.ac.id

STMIK GI MDP
Palembang, Indonesia
dafid@mdp.ac.id

2nd Ermatita
Universitas Sriwijaya
Palembang, Indonesia
ermatita@unsri.ac.id

***Abstract*—Internet and Web technologies provide students with the ease of communication to lecturers, accessing learning resources, and submitting the assignment. One of that technologies is Blended Learning. Blended learning has been growing in demand and popularity and widely used in the modern higher education system with the implementation of a learning management system (LMS). As a result, Blended Learning generates large amounts of student's information that can be used to invent valuable patterns and get hidden and useful information. This study analyzed data extracted from a Moodle-based blended learning course in STMIK XYZ that is called SIMPONI and SPON, to build a student model that predicts academic performance. Data Mining Classification Techniques were used to predict it based on Semester GPA. Various factors like demographics, economics and information about previous education are considered in analyzing academic performance. Rapid Miner software package was used for prediction and the result is a model by which could determine academic performance: passed with a CGPA above or equal to 2.5 and failed with a CGPA below 2.5 This model may give information for STMIK XYZ to early intervention and to outcome the poor GPA. Various classification algorithms are used to select the best algorithm in finding the better prediction model.**

Keywords—academic performance, blended learning, classification

I. INTRODUCTION

STMK XYZ is one of higher education that concern to generate qualified graduates with the best academic performances by improving the quality of learning. The need to achieve that goal has led it to embrace innovative practices like blended learning. Blended learning is designed to engage students in numerous interactions through a Moodle Learning Management System (LMS) that is called SIMPONI (Sistem Pembelajaran Online dan Interaktif) and SPON (Sistem Pembelajaran Online) in STMIK XYZ. SIMPONI and SPON allowed students to download learning materials, upload their assignments, communicate and collaborate in teams, discuss or ask a question with the lecturers and take an online test. Blended learning is a combination of traditional face-to-face learning and online learning delivery methods which aim to create a learning atmosphere that supports self-directed learning[1]. Blended learning has become the primary means for delivering learning materials for both online and traditional modes of education. Generally, there is a problem that always a matter for education institution especially for STMIK XYZ. The problem is about academic performance. Most of the studies stated that graduation is a measure of student success. STMIK XYZ used Cumulative Grade Point Average (CGPA) to evaluate academic's student performance. CGPA refers to the overall Semester GPA, which includes dividing the number of points earned in all courses by the total semester credit unit in all courses. The points is gained by quizzes, assignment, mid-term exam and final exam. The evaluation is important to recognize students which need special attention to maintain students performance and to reduce risk of academic failure. As a result the information allow STMIK XYZ to take a right decision. Academic performance is an outcome influenced by many factors. Yassein et al. stated the student's academic performance is affected by factors like socio economic, personal and other environmental variable[2]. The factors can be grouped into academic factors and nonacademic factors. Academic factors such as semester GPA and nonacademic factors such as attendance, time spent learning, login frequency, work during studies, demographic, high school type, scholarship, etc. As a result blended learning generate large amounts of data during the semester that is stored in the log files. Predicting academic performance has become challenging due to huge number of data. These data can be used to invent valuable pattern and get hidden and useful information. Shahiri et al. stated that presently there are several techniques used to evaluate the academic performance of students[3]. Data mining is one of the most familiar techniques use to examine the academic performance of students. It is an interdisciplinary area that brings together techniques from statistics, artificial intelligence, database systems, machine learning, pattern recognition, data visualization, knowledge acquisition, and information theory to find useful patterns and, thus, helps understand students' behavior and how they learn[4]. Several studies have reported by many researcher and provided a great results in prediction of student's academic performance. In most of these studies, classification is the most popular technique to predict student's academic performance. The goal of classification is to separate attributes (factors) into several predefined classes as accurately as possible. The algorithms that is used under classification technique are Decision Tree, Artificial Neural Network, Naïve Bayes, K-Nearest Neighbour and Support Vector Machine[1].

 ISBN: 978-623-92207-1-6

Devasia et al's. experiment is conducted by using Naive Bayesian mining technique for the extraction of useful information. The variable used are student admission details, course details, subject details, student marks details, attendance details and student's academic history as input. The results show that Naive Bayesian algorithm is more accurate than other methods like Regression, Decision Tree, Neural networks etc., for comparison and prediction[5]. Khobragade[6] proposed an approach where they have predicted the students' academic failure using decision tree, Naive Bayes and using classifiers that are based on induction rule and decision tree. Data used for classifications involved social, academic and background information of the student. These data have been collected through surveys. A total of 11 features were used for prediction after applying the feature selection algorithm. Classifiers have then been evaluated based on accuracy. Naive Bayes provided the best accuracy of above 87 percent. However, data were gathered through surveys which are time consuming and also involved methods that make overall prediction and do not consider early prediction. Zacharis[7] used data from a Moodle-hosted blended learning course to explore online activities that could be used to predict academic performance. The log file, containing the interactions of 134 students with the course material, mates and stuff, was processed to reveal predictors of success, such as the time spent in various activities or total LMS hits. Fourteen features were found to have significant correlation with student grades and used as the independent variables in a stepwise multivariate regression. The most predictive variables in student outcomes were the number of messages sent or viewed, the number of quiz efforts made, the number of files viewed and the number of contributions in different team tasks. To evaluate the predictive power of the regression model consisting of these four variables, a binary logistic analysis was performed that achieved a prediction accuracy of 81%. Kabakchieva[8] used various data mining classification algorithms, including a decision tree classifier, a neural network, a rule learner and a Nearest Neighbour classifier to develop a data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics. The main goal of her research was to demonstrate the great potential of data mining applications for university management. Hamsa[9] proposed academic performance prediction model using two selected classification method; Decision Tree and Fuzzy Genetic Algorithm. Variable used are internal mark, sessional mark and admission score. Result form decision tree makes lecturers a decision to take more care for the student at risk class.Result from fuzzy genetic algorithm give more passed student because of considering those who are in between risk and safe, to safe state that gives students a mental satisfaction.

This research used data mining classification technique in order to predict academic performance that focus on Decision Tree and Naïve Bayes algorithm. The goal was to propose a predictive model that will classify students who gain CGPA above or equal to 2.5 into passed categories and CGPA below 2.5 into failed categories. This model may give information for STMIK XYZ to early intervention and reduce the percentage of dropout students. The rest of this paper was organized in the following manner: Section II gives the methodology of this research. Section III discusses the research findings and introduces the study implications. Finally, Section IV outline conclusion and possible future directions.

II. RESEARCH METHODOLOGY

2.1 CRISP-DM (Cross Industry Standard Process for Data Mining)

The need to use a standard which would be a guideline to data mining projects is essential to succesfull data mining. This research use CRISP-DM methodology that has six main steps. The steps are business understanding, data understanding, data preparation, modeling, evaluation and deployment as in figure 1.

Fig. 1. The Steps of CRISP-DM Methodology[10]

Business understanding focus on understanding the project objectives and requirement. This step converts the business problem to a data mining problem. The second step is data understanding concerned with establishing the main characteristics of data. This step includes task for initial data collection. Next step is data preparation that involves all the activities for constructing the final data set on which modeling tools can be applied directly. The different task in this phase are select data,clean data, construct data, integrate data and format data. Data Preparation is the most important step due to in data mining its related to data quality. In the modeling step involved selecting modeling techniques and assessing the model created. Evaluation's phase validates the model from the data analysis point of view. The last step is deployment in which the results of data mining are presented[11].

2.2 Data

This research analyzed data of 144 courses during first semester until eighth semester. The data collected from 380 students in Information System, Informatic Enginering and Informatic Management study program at STMIK XYZ. There are three different kind of data used in this research to create a prediction model. The data are :

1. Demographic data and information about previous education
   These data are availaible on enrolment form that consist of question about gender, major of the high school they attended, if they work during studies or not, what category of their income, if they get scholarship or not, what their accomodation during studies and what their parents education

 ISBN: 978-623-92207-1-6

2. SIMPONI and SPON logs about the activity
   These online learning system give students online lessons with videos, online materials and modules, exercises, quizzes, assignment, mid and final exam, forum, chat and group discussion. The SIMPONI and SPON logs give the information about students activity regarding time spent learning like total login time and total login frequency. Total login time was computed by calculating the total of time spending between login and logout. Total login frequency was calculated by adding up the number of login time into SIMPONI and SPON.
3. Points gained during the semester
   The points the student gained each course they registered is divided into assignment, quiz, mid exam and final exam. Next, all of the points was calculated in the form of Semester GPA.

   Table I presents the list of all attributes used as input in this research as a result of data preparation step.

TABLE I. STUDENTS RELATED ATTRIBUTES

Attribute Name	Description	Domain
gnd	gender	male or female
mjr	major school	natural science, social studies, technical high school, technical high school non technic, others
wrk	work during studies	yes and no
inc	average monthly income	considerably below average, little below average, average, little above average and considerably above average
bea	scholarship	yes, no
acc	accommodation during studies	with parents/relatives, payment of rent
edu	parent's education	secondary education, bachelor's , master's or doctor's degree
tlt	total login time	long. medium, short
tlf	total login frequency	often, rarely, very rarely
gpa	semester GPA points	0.00 – 4.00

2.3 Model

The goal of this research is to create a prediction model about students academic performance based on three kind of with the above-mentioned data. This model could help the institution to classify students who are passed or failed in education. In the STMIK XYZ, CGPA is used as a standard measure. Students who gained CGPA above or equal to 2.5 are classified into passed categories and CGPA below 2.5 are classified into failed categories. In most of these studies, classification is the most popular technique to predict student's academic performance. Under the classification techniques, the proposed model was create used decision tree algorithm and Naïve Bayes algorithm.

2.4 Tool

For the model, this research used the Rapid Miner Studio version 9.8 software package. This software is very powerful to build predictive analytic models because it's a suitable platform for data preparation, machine learning and model deployment as well [12].

III. RESULT

After importing data to Rapid Miner Studio software, the software goes to read the earlier prepared data set and goes through five phases: data analysis, data pre-processing, design, training and testing as in figure 2. In the data analysis phase, we are determining the target attribute that we want to predict. The target predict is Status which has the value of passed if CGPA above or equal to 2.5 and failed if CGPA below 2.5. STMIK XYZ used CGPA to evaluate academic's student performance. In the data analysis step, Rapid Miner Studio software divides data into two sets: training (90%) and testing (10%).

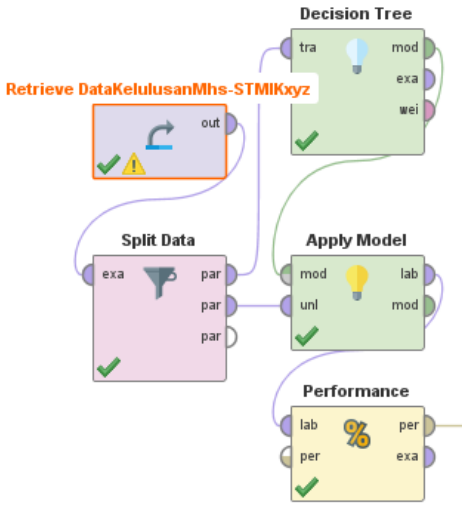

Fig. 2. The steps of process with Rapid Miner Studio software using Decision Tree Algorithm

3.1 The Prediction Model

Training process train the model used decision tree algorithm which requires criterion options like accuracy. During the testing process, we used two operations: The Apply Model on the test dataset and the Performance operation for measuring the model performance. In order to analyze the performance of these students a decision tree based model is created (see figure 3 and 4) which in the end helped to predict which students may passed or failed.

Tree

```
tlf = jarang
|   IPS 5 > 2.815: Lulus {Lulus=7, Tidak=0}
|   IPS 5 ≤ 2.815
|   |   IPS 4 > 1.780
|   |   |   mjr = IPA
|   |   |   |   tlt = sebentar: Tidak {Lulus=0, Tidak=5}
|   |   |   |   tlt = sedang
|   |   |   |   |   IPS 5 > 2.125: Lulus {Lulus=4, Tidak=0}
|   |   |   |   |   IPS 5 ≤ 2.125
|   |   |   |   |   |   IPS 2 > 2.485: Tidak {Lulus=0, Tidak=3}
|   |   |   |   |   |   IPS 2 ≤ 2.485: Lulus {Lulus=1, Tidak=1}
|   |   |   mjr = SMK Teknik: Lulus {Lulus=2, Tidak=0}
|   |   IPS 4 ≤ 1.780: Tidak {Lulus=0, Tidak=7}
tlf = sangat jarang: Tidak {Lulus=0, Tidak=33}
tlf = sering: Lulus {Lulus=312, Tidak=4}
```

Fig. 3. Decision Tree Model (Description)

ISBN: 978-623-92207-1-6

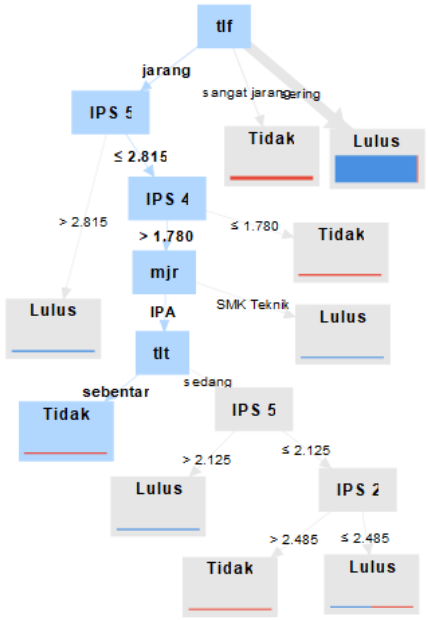

Fig. 4. Decision Tree Model (Graph)

3.2 Testing of Model

To test the accuracy of the model, this research used value got from confusion matrix and Area Under Curve (AUC). After executing the model, the result are presented in table confusion matrix as in table II which display the value of accuracy.

TABLE II. CONFUSION MATRIX OF THE DECISION TREE MODEL

	true Lulus	true Tidak	class precision
pred. Lulus	31	0	100.00%
pred. Tidak	2	5	71.43%
class recall	93.94%	100.00%	

The accuracy of model is 94.74% according to the confusion matrix and value of AUC is 0.939 (positive class: Tidak). The value of accuray is computed by the formula as in (1) :

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$
$$= (31+5)/(31+5+0+6)$$
$$= 94.74\%$$

Where:

TP (True Positive) is sum of data that predicted is passed and the fact is passed.
TN (True Negative) is sum of data that predicted is failed and the fact is failed.
FP (False Positive) is sum of data that predicted is passed and the fact is failed.
FN (False Negative) is sum of data that predicted is failed and the fact is passed.

The value of AUC is 0.939 (positive class: Tidak) means excellent classification with AUC (optimistic): 1.000 (positive class: Tidak) and AUC (pessimistic): 0.939 (positive class: Tidak) [13].

To increase the efficiency of this research, another model has created which is based on the Naive Bayes algorithm by using the previous used data set as well. The synthesis of this model is similar to the synthesis of the decision tree. It just had to replace the Decision Tree operation to the Naive Bayes operation as in figure 5.

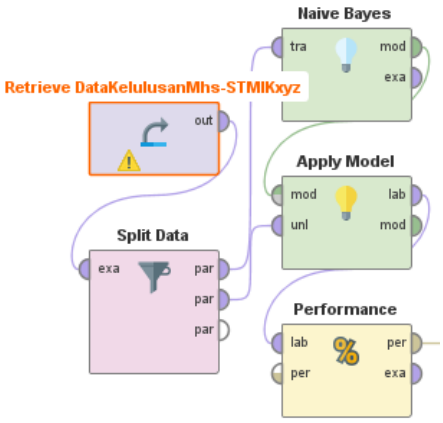

Fig. 5. The steps of process with Rapid Miner Studio software using Naive Bayes Algorithm

The result of Naive Bayes model is a simple distribution functions. These functions represent in each interval how much chance the students have to be passed or failed. The result show the ratio of passed (86%) is much higher than failed (14%) as in figure 6.

SimpleDistribution

```
Distribution model for label attribute STATUS

Class Lulus (0.860)
14 distributions

Class Tidak (0.140)
14 distributions
```

Fig. 6. Distribution function of the Naive Bayes model regarding the STATUS (passed or failed) of the students

After executing the Naïve Bayes model, the result are presented in table confusion matrix as in table II which display the value of accuracy.

TABLE III. CONFUSION MATRIX OF THE NAÏVE BAYES MODEL

	true Lulus	true Tidak	class precision
pred. Lulus	32	0	100.00%
pred. Tidak	1	5	83.33%
class recall	96.97%	100.00%	

The accuracy of model is 97.37% according to the confusion matrix and value of AUC is 1.000 (positive class: Tidak).

The value of AUC is 1.000 (positive class: Tidak) means excellent classification with AUC (optimistic): 1.000 (positive class: Tidak) and AUC (pessimistic): 1.000 (positive class: Tidak)

 ISBN: 978-623-92207-1-6

IV. CONCLUSION

This research presented a prediction model that found those key elements that help to identify those students who are likely to be failed in their studies because of various factors such as academic and non-academic (like socio economic). The final academic status is always a result of the previous semesters, which gives an opportunity for the prediction of a subsequent semester performance based on the previous ones. This research rely on the reliability of data mining classification techniques like Decision Tree and Naïve Bayes.

As a result, the model accurately predict the performance of students at different levels of their studies using various academic performance factors with the score above 90%. Naïve Bayes generates the result 1.03% more accurate than Decision Tree.

By adapting this model, education institution could recognize students which need special attention to maintain students performance and to reduce risk of academic failure and give information to take a right decision.

REFERENCES

[1] D. R. Garrison and H. Kanuka, ‘‘Blended learning: Uncovering its transformative potential in higher education,’’ Internet Higher Educ., vol. 7,no. 2, pp. 95–105, 2004.

[2] Hu, S.&Kuh, G.(2002).Being disengaged in educationally purposeful activities: the Influences of student and institutional characteristics. Research in Higher Education, 43(5), 555-575.

[3] Shahiri, A. M., Husain, W., Rashid, N. A. A review on predicting students performance using data mining techniques. The 3rd Information System International Conference, Elsevier, 72, p414 – 422, 2015.

[4] Sumathi, S., Sivanandam, S. N. (2006). „Introduction to data mining and its applications“, Springer. Vol. 29..

[5] Devasia, T., Vinushree, T. P., Hegde, V. (2016, March). „Prediction of students performance using Educational Data Mining“. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). pp. 91-95.

[6] Khobragade, L.P. (2015), “ Students’ academic failure prediction using data mining ”, Vol. 3 No. 5, pp. 2321-7782.

[7] Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. Internet and Higher Education, 27, 44

[8] Kabakchieva, D. (2012). „Student performance prediction by using data mining classification algorithms.“ International Journal of Computer Science and Management Research. Vol. 1 No 4, pp. 686-690.

[9] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, “Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm,” *Procedia Technol.*, vol. 25, pp. 326–332, 2016, doi: 10.1016/j.protcy.2016.08.114.

[10] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0. Letöltés dátuma: 2017. 03 02, forrás: IBN Corporation (2000). ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[11] Oracle (2019), "Data Mining Concepts", available at: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON002 (07 November 2019)

[12] RapidMiner: A complete platform for predictive analytics. Letöltés dátuma: 2017. 03 10, forrás: RapidMiner 7.3 Data Science Platform (2016). https://1xltkxylmzx3z8gd647akcdvov-wpengine.netdna-ssl.com/wp-content/uploads/2016/12/rm-platform-fact-sheet-12-2.pdf

[13] Gorunescu, F. (2011). Data Mining Concepts, Models and Techniques. (Springer, Ed.) (12th ed., Vol. 12). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-19721-5