International Journal of Technology and Engineering Studies
volume 5 issue 5 pp. 151-156 doi: https://dx.doi.org/10.20469/ijtes.5.10002-5

Development of Educational Data Mining Model for Predicting Student Punctuality and Graduation Predicate

Rianto*
Universitas Teknologi Yogyakarta,
Yogyakarta, Indonesia

Muhammad Fachrie
Universitas Teknologi Yogyakarta,
Yogyakarta, Indonesia

Abstract: This paper discusses the Educational Data Mining (EDM) to predict the punctuality and graduation predicate. Both are considered as important aspects that represent the student's academic performance. The model was developed by using academic records of 100 students from the vocational school of Informatics Management at Universitas Teknologi Yogyakarta. The dataset consisting of three features and two different labels was obtained by creating an Application Programming Interfaces (APIs) that connected to an academic database. Two classification algorithms were used to obtain knowledge from the dataset, i.e., Support Vector Machine (SVM) and Naive Bayes (NB). From the observations, SVM achieved the level of accuracy for punctuality of graduation on 0.68 while NB on 0.60. On graduation predicate, both algorithms achieved the same accuracy level on 0.92.

Keywords: *EDM, graduation predicate, NB, punctuality, SVM*

Received: 19 May 2019; ***Accepted:*** 2 October 2019; ***Published:*** 26 October 2019

I. INTRODUCTION

University is the highest level of education that must be completed by someone to get a bachelor's degree. Graduation is an ending process for students in completing of their education. In higher education learning outcomes are called Grade Point Average (GPA) which are divided into two: the semester and cumulative GPA. A GPA is a student's representation in education process. GPA will affect the duration study and graduation predicate. The duration study of higher education in Indonesia is 6 semesters for diplomas and 8 semesters for graduates. Graduation tolerance is given up to 10 semesters for diplomas and 14 semesters for graduates. In the learning process, students have an academic supervisor to supervise in their learning, so they can finish the study on time, 6 semesters for diplomas, and 8 semesters for graduates. The academic supervisor cannot provide full supervision because of other duties as a lecturer. This condition makes the academic supervision process not as expected. Academic supervisor must be able to advise and recommend so the students finish their study on time with impressive achievement. Therefore, we need an alternative digital academic supervisor who can advise and recommend for the students.

The information technology development presents several technologies for humans to complete their work. One technology is data mining, which is exploring data to find the hidden information. This process produces conclusions, patterns, rules, and others [1, 2]. The results of data exploration were used as basic knowledge to make predictions and recommendations [3].

In learning process, the students have academic data about the courses and grades. The academic data will be explored to find the patterns on predictions and recommendations. Data were classified to develop a classifier model by a classification algorithm. They are Naive Bayes and SVM [4, 5, 6] generally. The classifier models are trained and tested to assess their accuracy. The accuracy determines the predictions and recommendations correctly [7].

*Correspondence concerning this article should be addressed to Rianto, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia. E-mail: ans3981@naver.com

This research aims to develop a model for predictions and recommendations. It supports a student in their study. The model can be applied to the chatbot applications. The chatbot is a computer program which developed using artificial intelligence [8]. It will help students to get online academic suggestions and recommendations.

II. METHODS

The research uses academic data obtained from the Academic Information System of Universitas Teknologi Yogyakarta (SIA-UTY). This information system has been used from 2014 to the present. By API, researchers obtain data, consist of student, year of entry and graduation, and GPA. The research uses the data of Diploma program of Informatics Management from 2014 to the present.

The data cleaning and labeling process produces the data to be analyzed [9]. Data cleaning removes the unnecessary data. Data modification converts the decimal to an integer in conversion number, by GPA * 100, because GPA is a decimal number with 2 digits after the decimal point. The process produces 100 records with punctual and cumlaude labels using by binary numbers 0 and 1. The labeling process was done using an algorithm automatically. An algorithm can be defined in pseudocode [10] as shown:

```
Sub punctuality ()
    For i =1 to 100
                Length=year of graduation – year of intake
        If Length > 3
                Label=0
        Else
                Label=1
        End if
End sub
```

Fig. 1. Pseudocode 1

```
Sub cumlaude ()
    For i=1 to 100
                FinalGPA= (GPA1+ GPA2+ GPA3)/3
        If FinalGPA > 350
                Label=1
        Else
                Label=0
        End if
End sub
```

Fig. 2. Pseudocode 2

The research uses the dataset, for example are shown in Table 1 and Table 2.

TABLE 1
THE EXAMPLES OF CUMLAUDE AND NOT CUMLAUDE DATA

GPA 1	GPA 2	GPA 3	GPA 4	Label
333	361	369	373	1
304	314	360	378	1
376	352	382	365	1
290	300	278	304	0
295	290	321	347	0
285	290	286	295	0

TABLE 2
THE EXAMPLES OF PUNCTUAL AND NOT PUNCTUAL DATA

GPA 1	GPA 2	GPA 3	GPA 4	Label
333	361	369	373	1
304	314	360	378	1
376	352	382	365	0
342	352	356	369	1
366	361	369	373	1
366	379	365	400	0

Data cleaning and labeling were completed using Microsoft Excel by 2 research assistants and it is done in eight hours. The results are 2 excel files with punctual and cumlaude labels. Each file is classified using the Naive Bayes and SVM algorithms. The classification process is done using Python programming by Google Collaboratory in 2 hours. The results are confusion matrix, accuracy, precision, and recall of the model.

III. RELATED WORKS

The existence of data mining and artificial intelligence causes the subjectivity of decision making is reduce. The objectivity is increase because of the decision-making model based on the data. Data mining and artificial intelligence can be applied to some sectors such as health, commerce, education, and others. Data mining in education is Educational Data Mining (EDM). EDM aims to analyze educational institution data to find behavior patterns and predict student performance.

Research in EDM has been done by previous researchers. It was done to observe student performance and identify the factors that influence in learning process. Data analysis was done using the Waikato Environment for Knowledge Analysis (WEKA). The classification used Naive Bayes classifier, J48 Decision Tree, and Multi-Layer Perceptron (MLP). The results showed that social factors are dominant. In the other factor, the research

found the most influential factors in learning process such as alcohol consumption, family relationship , and parent's education [11, 12].

The research on Educational Data Mining identifies student performance so the failure risk can be minimized. The dataset obtained from the UCI Machine Learning Repository. The research compared several classification algorithms such as Bayes-based, ANN-based, Regression-based, SVM-based, Instance-based, Tree-based, and Rule-based classification. The attributes that used in the research are defined by G1, G2, and G3. The attributes of G1 and G2 are internal grades while G3 is the final exam grade. The results showed that the best classification algorithm was Random Forest. This research also found that family relationships influence the success study. The results of the research in [11] showed a correlation in the study success factors i.e., family relationships.

The other research that has been done in Educational Data Mining to improve the student performance. This research concerned in student's psychology factors. By using SVM, this research implemented RBF kernel to classify the data with category level: high, average, and low. This research also compared the classification algorithms which is used by previous researchers. The result showed that the SVM classifier could increase the level of accuracy from 89% to 90% [13].

Several research in Educational Data Mining focus on the improvement of student performance using grades. However, there are several factors such as student's background, social activities, and previous study can influence student academic performance. The research has been done to identify the significant and impact of student's background related to academic performance. The data classification used Naive Bayesian, Multilayer Perceptron, Decision Tree J48, and Random Forest. The result shows that student's background and social activities have an impact on academic performance significantly [14].

IV. RESULTS AND DISCUSSION

A. *Results*

The dataset in this research consist of punctuality and graduation predicate. The classification has been done by using 100 records on each dataset. Data were splitted into 75% (data training) and 25% (data testing). The classification algorithms that used were SVM and NB. The result of classification was a confusion matrix to calculate accuracy, precision, and sensitivity. The confusion matrix of the SVM classification in punctuality of graduation is shown in Fig. 3.

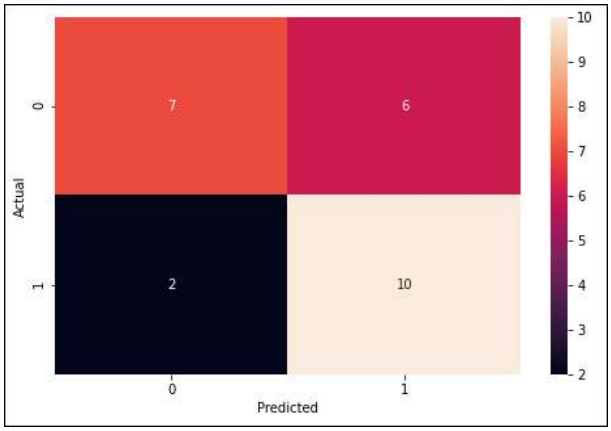

Fig. 3. Confusion matrix of punctuality of graduation using SVM

The classification report of punctuality of graduation using SVM is shown in Fig. 4.

	precision	recall	f1-score	support
0	0.78	0.54	0.64	13
1	0.62	0.83	0.71	12
accuracy			0.68	25
macro avg	0.70	0.69	0.68	25
weighted avg	0.70	0.68	0.67	25

Fig. 4. Classification report of punctuality of graduation using SVM

The confusion matrix of punctuality of graduation using NB is shown in Fig. 5.

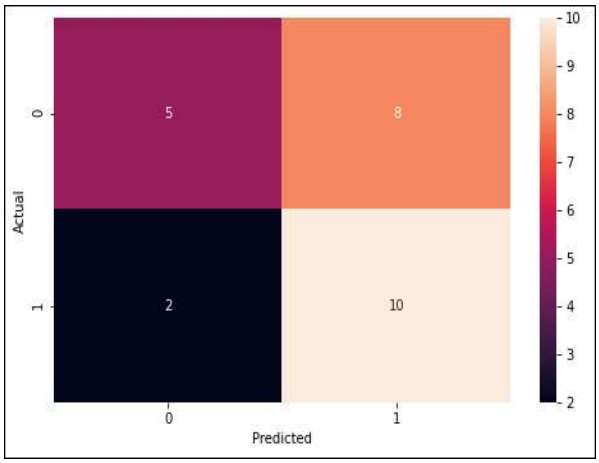

Fig. 5. Confusion matrix of punctuality of graduation using Naive Bayes

The classification report of punctuality of graduation using NB is shown in Fig. 6.

	precision	recall	f1-score	support
0	0.71	0.38	0.50	13
1	0.56	0.83	0.67	12
accuracy			0.60	25
macro avg	0.63	0.61	0.58	25
weighted avg	0.64	0.60	0.58	25

Fig. 6. Classification report of punctuality of graduation using Naive Bayes

The graduation classification classifies the punctuality. The research classifies the graduation predicate too, namely cumlaude or not. The dataset that used in the graduation predicate is 100 records. The classification has been done by splitted data into 75% (data training) and 25% (data testing) using SVM and Naive Bayes Classifier. The result of classification in the graduation predicate using SVM is shown in Fig. 7.

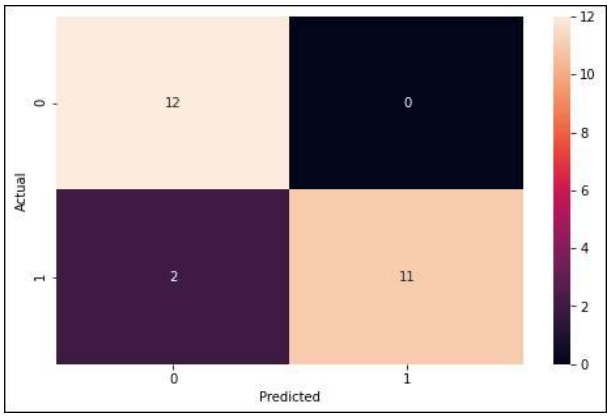

Fig. 7. Confusion matrix of graduation predicate using SVM

The classification report of graduation predicate using SVM is shown in Fig. 8.

	precision	recall	f1-score	support
0	0.86	1.00	0.92	12
1	1.00	0.85	0.92	13
accuracy			0.92	25
macro avg	0.93	0.92	0.92	25
weighted avg	0.93	0.92	0.92	25

Fig. 8. Classification report of graduation predicate using SVM

The confusion matrix of graduation predicate using NB is shown in Fig. 9.

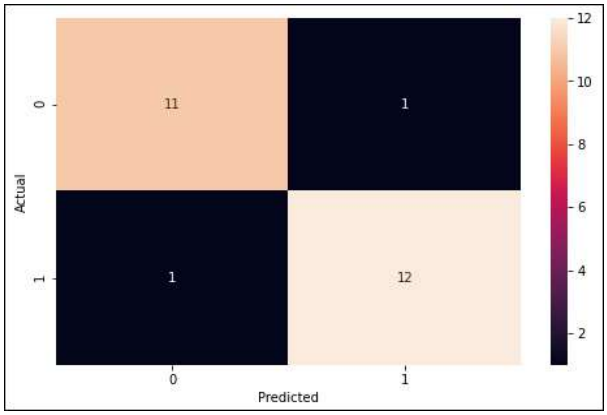

Fig. 9. Confusion matrix of graduation predicate using NB

The classification report of graduation predicate using NB is shown in Fig. 10.

	precision	recall	f1-score	support
0	0.92	0.92	0.92	12
1	0.92	0.92	0.92	13
accuracy			0.92	25
macro avg	0.92	0.92	0.92	25
weighted avg	0.92	0.92	0.92	25

Fig. 10. Classification report of graduation predicate using NB

B. *Discussion*

In a classification method, this research was included in the binary classification. The binary classification classified instances into one of two classes. The result of the classification process is a classifier model that summarized by performance indicators such as accuracy, precision, and recall that calculated by using the value in the confusion matrix. The confusion matrix was used to evaluate how often a certain behavior is detected by true and false [15]. The elements of the confusion matrix were True Positive (TP), True Negative (TN), False Positive (FP), and False-Negative (FN). True positive and true negative showed that actual and prediction were true. False-positive was a condition when the actual was negative, but it was predicted positive, while false-negative was a condition when the actual positive but it was predicted negative [15]. The elements in a confusion matrix are shown in Fig. 11.

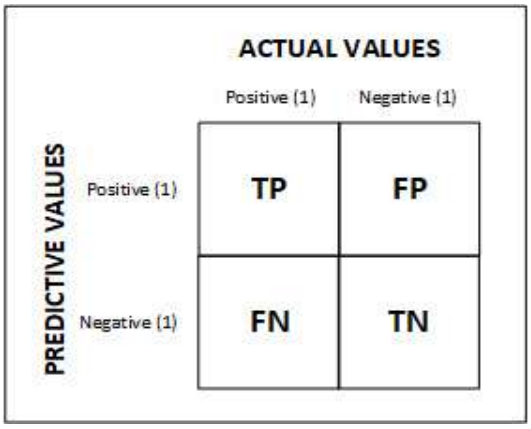

Fig. 11. The elements of confusion matrix

The accuracy of the classifier model was calculated (1) by using a formula [16]:

$$Accuracy = (TP+TN)/(TP+FP+TN+FN) \quad (1)$$

The accuracy was calculated based on data training which was setted in 25% of total data. The total data that used on each classification were 100 records, so the number of data training was 100 * 25% = 25. Based on a result of punctuality of graduation classification was known that accuracy level of SVM was 0.68 and 0.60 on NB. According to the accuracy formula, the detail of calculation based on Fig. 3 and Fig. 5 were shown:

$$SVMAccuracy = (7+10)/(7+10+6+2) = 0.68 \quad (2)$$

$$NBAccuracy = (5+10)/(5+10+8+2) = 0.60 \quad (3)$$

The level of accuracy between SVM and NB was known that SVM was better than NB. Based on the confusion matrix as shown Fig. 3 and Fig. 5 was known that the punctuality of graduation classifier model on both classification algorithms had false-negative value 6 on SVM and 8 on NB. The false-positive value on both algorithms was 2. The error rate on each classification algorithms was 0.32 on SVM and 0.40 on NB. The low level of accuracy in the punctuality of graduation classifier model was caused by the limitation of attributes in the dataset. It showed that to predict punctual or not, it's not enough using student's GPA.

The graduation predicate classifier model had a higher accuracy level was compared with the punctuality of graduation classifier model. In both classification algorithms, the accuracy level was 0.92. In predicting cumlaude, NB was better and accurate than SVM. In predicting not cumlaude, SVM was better and accurate than NB. Finally, SVM Algorithm using RBF kernel was better than NB. It can be proven by comparing the confusion matrix that shown in Fig. 3 and Fig. 5, for punctuality of graduation and Fig. 7 and Fig. 9 for graduation predicate.

This research has limitations in punctuality of graduation, so the future work in predicting punctuality of graduation several attributes should be added. The student's background that related to their previous study, the teacher and parent's impact [17, 18], and family relationship influenced the academic performance.

V. CONCLUSION

This research compared two classification algorithms i.e., SVM and NB to classify the punctual and graduation predicate. On punctuality of graduation classification, SVM had a level of accuracy was higher than NB on 0.68. However, in the graduation predicate, both algorithms had the same level of accuracy on 0.92. Based on analysis using the confusion matrix, SVM was more accurate. The future works are adding the features related to students such as demographic data, family background, student activity, and others.

VI. ACKNOWLEDGMENT

The research was supported by the Ministry of Education and Culture of the Republic of Indonesia. The researcher thanks to Universitas Teknologi Yogyakarta for providing the data and the great advising so this paper is published.

REFERENCES

[1] G. Schuh, G. Reinhart, J.-P. Prote, F. Sauermann, J. Horsthofer, F. Oppolzer, and D. Knoll, "Data mining definitions and applications for the management of production complexity," *Procedia CIRP*, vol. 81, pp. 874–879, 2019.

[2] N. Ugtakhbayar, B. Usukhbayar, S. H. Sodbileg, and J. Nyamjav, "Detecting TCP based attacks using data mining algorithms," *International Journal of Technology and Engineering Studies*, vol. 2, no. 1, pp. 1–4, 2016. doi: https://doi.org/10.20469/ijtes.2.40001-1

[3] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471–1483, 2020. doi: https://doi.org/10.1016/j.procs.2020.03.358

[4] M. W. Rodrigues, S. Isotani, and L. E. Zárate, "Educational data mining: A review of evaluation process in the e-learning," *Telematics and Informatics*, vol. 35, no. 6, pp. 1701–1717, 2018. doi: https://doi.org/10.1016/j.tele.2018.04.015

[5] S. Maitra, S. Madan, R. Kandwal, and P. Mahajan, "Mining authentic student feedback for fac-

ulty using naïve bayes classifier," *Procedia Computer Science*, vol. 132, pp. 1171–1183, 2018. doi: https://doi.org/10.1016/j.procs.2018.05.032

[6] T. A. Cardona and E. A. Cudney, "Predicting student retention using support vector machines," *Procedia Manufacturing*, vol. 39, pp. 1827–1833, 2019. doi: https://doi.org/10.1016/j.promfg.2020.01.256

[7] R. Medar, V. S. Rajpurohit, and B. Rashmi, "Impact of training and testing data splits on accuracy of time series forecasting in machine learning," in *International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India*, 2017.

[8] A. Androutsopoulou, N. Karacapilidis, E. Loukis, and Y. Charalabidis, "Transforming the communication between citizens and government through ai-guided chatbots," *Government Information Quarterly*, vol. 36, no. 2, pp. 358–367, 2019.

[9] B. Bilalli, A. Abelló, T. Aluja-Banet, and R. Wrembel, "Presistant: Learning based assistant for data pre-processing," *Data & Knowledge Engineering*, vol. 123, pp. 10–17, 2019. doi: https://doi.org/10.1016/j.datak.2019.101727

[10] T. Dirgahayu, S. N. Huda, Z. Zukhri, and C. I. Ratnasari, "Automatic translation from pseudocode to source code: A conceptual-metamodel approach," in *IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Phuket, Thailand*, 2017, pp. 122–128.

[11] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," in *4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Utar Pradesh, India*, 2017, pp. 568–572.

[12] C. L. S. Tablatin, F. F. Patacsil, and P. V. Cenas, "Design and development of an information technology fundamentals multimedia courseware for dynamic learning environment," *Journal of Advances in Technology and Engineering Research*, vol. 2, no. 6, pp. 202–210, 2016. doi: https://doi.org/10.20474/jater-2.6.5

[13] I. Burman and S. Som, "Predicting students academic performance using support vector machine," in *Amity International Conference on Artificial Intelligence (AICAI)*, 2019, pp. 756–759.

[14] C.-C. Kiu, "Data mining analysis on students academic performance through exploration of students background and social activities," in *Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA),* Subang jaya, Malaysia, 2018.

[15] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behavioural Processes*, vol. 148, pp. 56–62, 2018. doi: https://doi.org/10.1016/j.beproc.2018.01.004

[16] P. Galdi and R. Tagliaferri, "Data mining: Accuracy and error measures for classification and prediction," *Encyclopedia of Bioinformatics and Computational Biology*, pp. 431–436, 2018.

[17] J. Caddell and D. Newell, "Evaluating teacher impact on student performance: A case study at the United States Military Academy," in *International Systems Conference (SysCon),* Orlando, FL, 2019.

[18] D. R. Padhi and A. Joshi, "A correlational study between the parent and the teacher's self-reported assessments on the child's performance," in *Tenth International Conference on Technology for Education (T4E),* Goa, India, 2019, pp. 280–281.