ATLANTIS PRESS

*Advances in Intelligent Systems Research (AISR), volume 144*

2017 International Conference on Education and Technology (2017 ICEduTech)

# Predicting Student Academic Performance by Mining Their Internet Usage Behaviour

Utomo Pujianto[1], Mahmuda Muthmainnah[2]

[1,2]Department of Electrical Engineering, Universitas Negeri Malang, Indonesia

utomo.pujianto.ft@um.ac.id, muthmm17@gmail.com

***Abstract*— Students who are active this time belong to the Generation Z. Namely, the generation that most of his time is spent interacting with computing devices, especially the virtual world. It can bring positive and negative effects on their performance in the learning process. This study proposes the use of students' internet behavior data to predict their academic performance. The method used in this study is Naive Bayes. There are six attributes of student internet behavior used in this study, including the number of social media accounts owned, the number of hours spent accessing social media, the number of hours spent on non-social media entertainment on the internet, the number of hours on the internet used for learning, and number of internet sessions in a week. The accuracy and sensitivity metric values of the experiments exceeded 85%, so it can be concluded that the mining of students' internet behavior data has the potential to be used to assist educational institutions in monitoring the performance of their students.**

***Keywords—student; academic performance; internet; behavior; naive bayes***

## I. INTRODUCTION

The Internet has become a part of everyday life for millennials. The time they spend more on surfing in the virtual world than doing any other physical activity. Such behavior is also done by the students. Especially students who belong to the generation Z..

Generation Z, also known as “Gen Next” or “Gen I,” includes people born between the early 1990s and the early 2000s (Posnick-Goodwin, 2010). Some consider members of Generation Z to be smarter, more selfdirected, and more able to quickly process information than previous generations; but there is one thing they may not be—team players. And that just might be the best reason to pay attention to new research about cooperative learning.

Chen et al. [1] explain that differences in academic grades and learning satisfaction between heavy and nonheavy Internet users were statistically significant. Nonheavy users had better grades and greater learning satisfaction than heavy users. The data suggested students who spend a significant amount of time online experience academic and learning difficulties. Young [3] suggested that students who spend excessive time on the Internet may have difficulty completing homework assignments, studying, and getting sufficient sleep to meet their academic responsibilities.

## II. RELATED WORK

There are previous researches that developed algorithms to predict students' achievement, namely, Backpropagation Neural Network Method to Measure Student Achievement Correlation Level (Case Study at Dian Nuswantoro University Semarang) and in the research resulted prediction of 61% and considered less than maximum [5]. Naive Bayes became the solution to predict student achievement. Naive Bayes is a simple probabilistic classifier that computes a set of probabilities by summing the frequency and value combinations of the given dataset. The algorithm uses the Bayes theorem and assumes all the independent or non-dependent attributes given by the value in the class variable [6]. The Bayes naive method can show the level of truth of output through probability theory, but it can not process inference from many interrelated rules. The advantages of this research using the Naive Bayes Method are having a fairly high accuracy [7].

Based on the character of the algorithm, naive bayes have an opportunity to improve good predictive results. For that, the study was entitled "Prediction of Student Achievement Result Based on Internet Usage Using Naive Bayes Method". The purpose of this study is to determine the effect of internet network usage on student achievement results so that students can make improvements to achieve maximum performance by utilizing internet network well, if students use internet network in campus environment, hence university or majors can block site which can disrupt the learning process and students can follow the learning activities effectively by not opening social media or entertainment media that can disrupt lectures. So the results of student achievement can increase.

## III. METHODOLOGY

### *A. Research Instruments*

This study focuses on predictions of student achievement outcomes based on the use of the Internet network using the Naive Bayes method using WEKA 3.8 program assistance as an instrument to implement, test and measure the accuracy of the Naive Bayes algorithm.

Weka is a software that has many machine learning algorithms for data mining purposes. Weka also has many tools for data processing, ranging from pre-processing, classification, regression, clustering, association rules, and

visualization [17]. Weka is a Java-based open source software and we can use it directly or through Java programs [18].

### *B. Research Data*

To collect the data obtained by conducting a survey using google form distributed to students who are still active college. Test data used as many as 285 records consisting of 6 attributes and classified into 2 classes namely GPA> 3 and GPA <3. Attributes used are several parameters that influence in predicting student achievement results in internet network usage activities. Some attributes used in this study are as follows:

1. Number of Social Media. This attribute contains the amount of social media that is owned and used by the student.
2. Number of Internet Usage. This attribute contains the internet network usage time of the week.
3. The amount of Internet Usage Time. This attribute contains the amount of internet network usage time in one day (24 hours).
4. Number of Internet Usage for Entertainment. This attribute contains the amount of time internet network usage for entertainment media in one day (24 hours).
5. Number of Internet Usage for Social Media. This attribute contains the amount of time internet network usage for social media in one day (24 hours).
6. Number of Internet Usage to Learn. This attribute contains the amount of time the internet network usage to learn in one day (24 hours).
7. GPA. This attribute is a class attribute or an output, there are two groups namely GPA> 3 and GPA <3.

### *C. Preprocess Stages*

This preprocess is performed to prepare the data before it is processed further [17]. The goal is to get clean and ready data for research and the data can be processed on the method used. Some of the steps taken are as follows:

1. Data cleaning (data cleaning). Data cleaning is a process of eliminating inconsistent noise and data or irrelevant data [8].
2. Data integration (data integration). Data integration is a combination of data [8].
3. Data Reduction. This process aims to combine the information contained in a large dataset into a small dataset.
4. Transformation of data (data transformation). Data is altered or merged into the appropriate format for processing in data mining [8].

### *D. Naive Bayes*

Naive Bayes is a simple probabilistic classifier that computes a set of probabilities by summing the frequency and value combinations of the given dataset. The algorithm uses Bayes's theorem and assumes all the independent or non-dependent attributes given by the value of the class variable [8]. Another definition says Naive Bayes is a classification with the probability and statistical methods brought by British scientist Thomas Bayes, predicting future opportunities based on past experience [7].

Naive Bayes is based on the simplifying assumption that attribute values are conditionally independent if given an output value. In other words, given the value of output, the probability of observing collectively is the product of the individual probability [9]. The advantage of using Naive Bayes is that this method requires only a small amount of training data to determine the estimated parameters required in the classification process. Naive Bayes often work much better in most real-world situations that are complex than expected [10].

Naïve Bayes is a simple probabilistic classification method and is designed to be used with the assumption that one class with another is not independent. In the Naïve Bayes classification, the learning process is more emphasized on estimating probabilities. The advantage of this approach is that classification gets a smaller error value when the data set is large [11] [14].

The Naïve Bayes classification is proven to have high accuracy and speed when applied to a large number of databases [12].

### *E. Evaluation Method*

Measurement of algorithm performance can be done by using confussion matrix. Confusion matrix is one of the techniques used to perform performance calculations on the data mining classification algorithm using a matrix. By knowing the amount of data in the correct classification, it can be seen the accuracy of the prediction results. The error rate of the predicted result can be determined from the amount of data clarified by false [19]. This test is suitable for testing datasets that have two classes but for some cases such as classification into several classes, the confussion matrix can be modified. Confusion matrix for datasets that have two classes is shown in Table I.

TABLE I. CONFUSION MATRIX

| Factual Class | Classified as | |
|---|---|---|
| | + | - |
| + | True Positive | False Positive |
| - | False Negative | True Negative |

In the above confusion matrix table, True Positive is the number of positive records classified as positive, false positive is the number of negative records classified as positive, false negative is the number of positive records classified as negative, true negative is the number of negative records that are classified as negative, then enter test data [19]. Evaluation using confusion matrix yield three test result that is:

Accuracy. Is a percentage of accuracy of data that is classified correctly after testing using an algorithm. Accuracy can be calculated using Equation 1.

$$\text{Accuracy} = (tp + tn) / (tp + tn + fp + fn) \times 100\% \quad (1)$$

Sensitivity. Used to compare the proportion of TP to positive-valued data. Sensivity is calculated using Equation 2.

$$\text{Sensitivity} = tp / (tp + fn) \times 100\% \quad (2)$$

PPV (Positive Predictive Value). Is the proportion of cases with a positive diagnosis. PPV is calculated using equation 3.

$$PPV = tp / (tp + fp) \times 100\% \quad (3)$$

## IV. Result and Discussion

In this study the data are classified using the attributes of the GPA into the class attribute. And grouped into 2 classes of IPK class above 3 (GPA> 3) and GPA below 3 (GPA <3). Based on data that has been classified with the help of WEKA program produce classification by using confussion matrix. The result of classification using confusion matrix can be seen in Table II.

TABLE II. Confusion Matrix with result

| Factual Class | Classified as | |
|---|---|---|
| | + | - |
| + | 244 | 3 |
| - | 35 | 3 |

Confusion matrix generally generates two values, ie, accuracy and error rate [13]. From the Confusion matrix result indicates that the classification result in GPA class> 3 is predicted correctly as 244 data, and the classification result in GPA class <3 is predicted with wrong is 3 data. Then for the classified prediction result of GPA <3 correctly predicted 3 data and the result of classification in GPA class <3 is predicted with wrong is as much as 35 data.

## V. Conclusion

Based on the results of research conducted to predict student achievement based on the use of internet network using naive bayes method showed a high accuracy of 86.6667%, has an error rate of 13.3333%, sensitivity of 98.7854% and PPV of 87, 4551.

## References

[1] Y.-F. Chen and S. S. Peng, "University Students' Internet Use and Its Relationships with Academic Performance, Interpersonal Relationships, Psychosocial Adjustment, and Self-Evaluation," CyberPsychology & Behavior, vol. 11, no. 4, pp. 467–469, Aug. 2008.
[2] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Computer Science, vol. 72, no. Supplement C, pp. 414–422, Jan. 2015.
[3] D. J. Kuss, M. D. Griffiths, and J. F. Binder, "Internet addiction in students: Prevalence and risk factors," Computers in Human Behavior, vol. 29, no. 3, pp. 959–966, May 2013.
[4] R. Rusno, "PENGARUH PENGGUNAAN INTERNET SEBAGAI SALAH SATU SUMBER BELAJAR TERHADAP PRESTASI MAHASISWA PENDIDIKAN EKONOMI UNIVERSITAS KANJURUHAN MALANG," Jurnal Ekonomi Modernisasi, vol. 6, no. 2, pp. 161–172, Jun. 2010.
[5] Aminudin. 2009. Terampil Menggunakan Internet. Bandung: Satu Nusa.
[6] Andayani, Sri Wahyuni. 2015. "Optimalisasi Penggunaan Media Internet Sebagai Sumber Belajar Mahasiswa Program Studi Pendidikan Kesejahteraan Keluarga, Jurusan Pendidikan Teknologi Dan Kejuruan, Fakultas Keguruan Dan Ilmu Pendidikan, Universitas Sarjanawiyata Tamansiswa Yogyakarta."
[7] Novianto, Iik. "Perilaku Penggunaan Internet Di Kalangan Mahasiswa." Fisip Unair: 1–40.
[8] Mustafidah, H. dan Suwarsito. 2012. Student Learning Achievement Prediction Based on Motivation, Interest, and Discipline Using Fuzzy Inference System. Proceeding International Conference on Green World and BusinessmTechnology 2012 (IC-GWBT2012) Technopreunership Based on Green Business and Technology, Ahmad Dahlan University Yogyakarta, ISBN: 978-979-3812-25-0, 23 – 24 March 2012.
[9] Jelita. 2013. Penggunaan Fasilitas WiFi dan Pengaruhnya terhadap Indeks Prestasi Mahasiswa Prodi Pendidikan Matematika (Studi pada Mahasiswa Prodi Pendidikan Matematika STAIN Zawiyah Cot Kala Langsa). Vol. 1, No.01, Januari 2013.
[10] Kartini, Aprilliani Dwi. 2016. Sistem Pendukung Keputusan Produk Olahan Kayu Jati (Bahan Baku) Menggunakan Metode Naive Bayes Pada Perum Perhutani.
[11] Patil, T. R., Sherekar, M. S., 2013, Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science and Applications, Vol. 6, No. 2, Hal 256-261.
[12] Ridwan, M., Suyono, H., Sarosa, M., 2013, Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier, Jurnal EECCIS, Vol 1, No. 7, Hal. 59-64.
[13] Bustami., 2013, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI : Jurnal Penelitian Teknik Informatika, Vol. 3, No.2, Hal. 127-146.
[14] Pattekari, S. A., Parveen, A., 2012, Prediction System for Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294.
[15] Berry, I. H. and Browne, M. 2006. Lecture Notes in DATA MINING. USA: World Scientific.
[16] Han, J and Kamber, M. 2006. Data Mining Concepts and Techniques, second edition. California: Morgan Kaufman.
[17] Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: ANDI Yogyakarta.
[18] R. E. Putri et al., Perbandingan Metode Klasifikasi Naïve Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012. Jurnal Gaussian, Vol. 3, No.4 (2014) ISSN:2339-2541.
[19] Sulistyaningsih, Y., Djunaidy, A., & Kusumawardani, R. P. (n.d.). Pengklasifikasian Pengaduan Masyarakat pada Laman Kantor Pertanahan Kota Surabaya I dengan Metode, 1–6.
[20] Pratama, W. A. (2013). Analisa Perbandingan Algoritma Decision Tree , Naive Bayes , dan k-NN dalam Penentuan Target Tindakan Terorisme di Indonesia.
[21] http://www.cs.waikato.ac.nz/ml/weka/index.html
[22] Rahmansyah, Arif. 2014. Getting Started: Weka. https:// ariefrahmansyah.wordpress.com/2014/11/18/getting-started-weka/
[23] Han, Jiawei. Kamber, Micheline. Data Mining: Consepts and Techniques. 2001 San Fransisco, USA. Morgan Kaufmann Publisehers.