Advances in Social Science, Education and Humanities Research, volume 410
1st International Multidisciplinary Conference on Education, Technology, and Engineering (IMCETE 2019)

A Review of Students' Graduation Classification: A Comparison of Naive Bayes Classifier and K-Nearest Neighbour

1st Via Tuhamah Fauziastuti
Department of computer science
Banten Jaya University
Serang Banten, Indonesia

2nd Lilis Aslihah Rakhman
Department of computer science
Banten Jaya University
Serang Banten, Indonesia
lilis_arachman@yahoo.co.id

Abstract—Students are the most crucial aspects in determining the successful implementation of every program offered within educational institutions. Monitor the progress and students' achievement, enhance the ability of students, consider the number of students who have graduated, the ratio of the total number of students, the competence of graduates, are several major factors that require attention and serious consideration from the higher educational institutions. This study is mainly based on the data mining technique by implementing two common algorithms namely Naive Bayes Classifier and K-Nearest Neighbour due to classifying students' graduation (on time and overtime). The main objectives of this paper are to compare the achievement of both algorithms (NBC and KNN) towards students' graduation classification. This paper also focuses to identify the most important variables to predict students' performance. Leading to two categories of dependent variables namely graduate on time or graduate overtime. The considerations of these variables are based on the importance of each towards classifying the target. A Cross-validation technique is applied to evaluate both algorithms. This study is beneficial for the center of graduate studies, educators, policymakers, and others in order to identify the main factors that impact the students' graduation status in higher educational institutions.

Keywords: NBC, KNN, student graduation, classification

I. INTRODUCTION

One of the main goals of higher education institutions is to provide quality education to every student [1]. Universities are encouraged to provide quality education to every student in order to produce knowledgeable, skilled, creative and competitive human resources [2]. According to Abdul Rohman (2015), one of the quality measures for higher education is students and graduates. Students are often referred to as community groups with more intellectual characteristics than their non-student peers. with this intelligence that students are able to deal with and find the source of the problems systematically and can be applied later in life and able to compete in the job market [3].

Students are an important aspect in evaluating the success of a program of study at an educational institution. Therefore, it is important that Institutions of Higher Education focus on monitoring student progress and achievement, improving student capacity, taking into account undergraduate and total student ratios, and graduate competencies in the job market.

Data mining (DM) is one of the processes for obtaining knowledge of patterns from data sets that can be used to classify student data so that its output can be used as a decision-making tool [4]. Data mining can be used for a variety of applications in the education sector for the purpose of enhancing student performance and accreditation of educational institutions [5]. It is called educational data mining. Educational data mining is a process used to extract useful information and patterns from a huge educational database [6]. Therefore, a systematic review is proposed to support the objectives of this study, which are:

- To compare the achievement of both algorithms (NBC and KNN) towards students' graduation classification
- To identify the most important variables to predict students' performance

II. METHODOLOGY

The aim of the systematic literature review is to provide the answers to proposed research questions [7]. Furthermore, the research questions also useful to identify the scope of the study and to guide the result.

A. Research Questions

The research questions (RQ) were designed by following the Kitchenhams steps, which consists of Population, Intervention, Outcome, and Context (PIOC). Table 1 shows the (PICO) structure of the research questions.

TABLE 1. STRUCTURE OF THE RESEARCH QUESTIONS

Criteria	Details
Population	University (student graduation)
Intervention	Methods or techniques for prediction
Outcome	Prediction accuracy, important variables
Context	Studies in academic institutions

Therefore, the research questions in this study are :

- Q1: What are the important variables used to predict students' graduation?
- Q2: Which one has the highest accuracy between NBC and KNN?

B. Search Strategy

In a systematic review, the search process consists of selecting digital libraries and defining the search string. An

extensive search for research papers was conducted to try answering the proposed research questions. The resulting search strings are as follows: (student graduation) AND (systems OR application OR method OR process OR system OR technique OR methodology OR procedure) AND (educational data mining) AND (prediction OR classification OR assessment). Searched databases: IEEE Xplore, Science Direct, ACM digital Library, Springer Link. Search items: Journal articles and conference papers. Publication period: Since 2012. This search items was limited until mid-year of 2019.

III. Research Result and Discussion

This section will discuss the important variables used to predict students' graduation and accuracy score between NBC and KNN. Figure 1 shows a list of common variables and methods used in predicting student's graduation. The first subsection will focused on the important variables to predict student's graduation and second will be focused on accuracy score between NBC and KNN.

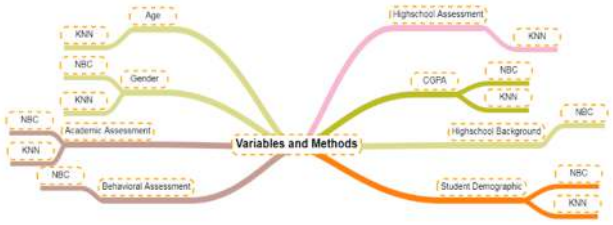

Figure 1. list of common variables and methods used in predicting student's graduation

A. *The important variables to predict student's graduation*

The variables that have been frequently applied is Cumulative Grade Point Average (CGPA) and academic assessment. Seven papers have used CGPA as their main variable to predict or classify student's graduation. Most of the researchers using CGPA as the main attributes because it has a significant input variable to predict student's graduation. Academic assessment such as assignment mark, attendance, quizzes. The third variable that has been frequently used also Gender. The reason why gender is frequently used by reasearcher to predict student's graduation is because they have different style in learning process between male and female.

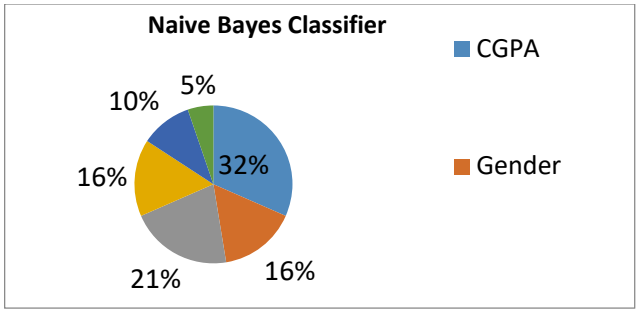

Figure 2. Variables frequently used by NBC

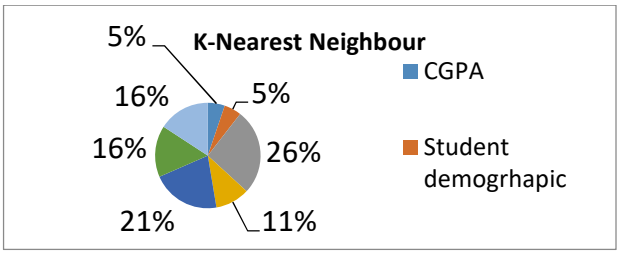

Figure 3. Variables frequently used by KNN

B. *The accuracy score between NBC and KNN*

The aim of the comparison between NBC and KNN is to know which technique is better to predict student's graduation. most researchers use a cross-validation technique to evaluate the algorithm and to get the accuracy score in predicting student's graduation. Table 2 shows the accuracy score of NBC and KNN. By looking at the table 2, NBC has a higher score of 91.10% rather than KNN 85.15%.

TABLE 2. THE ACCURACY OF NBC AND KNN IN PREDICTING STUDENT'S GRADUATION

Algorithm	Accuracy
Naive Bayes Classifier (NBC)	68.5%
	75.16%
	91.10%
	69.07%
	83.65%
	70%
K-Nearest Neighbour (KNN)	83%
	68.32%
	85.15%
	70.49%
	70.5%

IV. Conclusion

This paper reviewed the current research on the important factors for predicting or classifying students' graduation time. Most of the researchers have used CGPA and academic assessment as datasets. The accurate prediction will help the university to manage graduation rates, due to good graduation rates will improve the university's rank.

Acknowledgment

Special thanks to the computer science department of Banten Jaya University, which has supported this research.

References

[1] Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.

[2] Rohman, A. (2015). Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa. *Neo Teknika*, *1*(1).

[3] Azwar. (2004). Penyusunan Skala Psikologi. Yogyakarta: Pustaka Pelajar

[4] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[5] Anuradha, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian Journal of Science and Technology*, *8*(15).

[6] Angeline, D. M. D. (2013). Association rule generation for student performance analysis using apriori algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), 1(1), 12-16.

[7] Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

[8] Shahiri, A. M., & Husain, W. (2015). A review on predicting student's

performance using data mining techniques. Procedia Computer Science, 72, 414-422.

[9] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261.

[10] Tahyudin, I., Utami, E., & Amborowati, A. (2013). Comparing clasification algorithm of data mining to predict the graduation students on time. ISICO 2013, 2013.

[11] Amalia, N., Shaufiah, S., & Sa'adah, S. (2015). Penerapan Teknik Data Mining untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma Naive Bayes Classifier. *eProceedings of Engineering*, *2*(3).

[12] Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *jurnal EECCIS*, *7*(1), 59-64.

[13] Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. *International journal of computer science and management research*, *1*(4), 686-690.

[14] Sriram, K., Chakravarthy, T., & Anastraj, K. A COMPARATIVE ANALYSIS OF STUDENT PERFORMANCE PREDICTION USING MACHINE LEARNING TECHNIQUES WITH DEEDS LAB.

[15] Murtopo, A. A. (2016). Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes. *CSRID (Computer Science Research and Its Development Journal)*, *7*(3), 145-154.

[16] Mayilvaganan, M., & Kalpanadevi, D. (2014, December). Comparison of classification techniques for predicting the performance of students academic environment. In *2014 International Conference on Communication and Network Technologies* (pp. 113-118). IEEE.

[17] Nurafifah, M. S., Abdul-Rahman, S., Mutalib, S., Hamid, N. H. A., & Ab Malik, A. M. (2019). Review on predicting students' graduation time using machine learning algorithms. *International Journal of Modern Education and Computer Science*, *11*(7), 1.

[18] Tampakas, V., Livieris, I. E., Pintelas, E., Karacapilidis, N., & Pintelas, P. (2018, June). Prediction of students' graduation time using a two-level classification algorithm. In *International Conference on Technology and Innovation in Learning, Teaching and Education* (pp. 553-565). Springer, Cham.