e-ISSN 2963-2870
Vol 5 2024

UNLOCKING THE FUTURE: HOW EDUCATIONAL DATA MINING IS SHAPING MARKET SEGMENTATION

Diky Paramitha[1], Novita Nugraheni[2]

[1, 2]*Universitas Terbuka (INDONESIA)*

dikyparamitha@ecampus.ut.ac.id

Abstract

In an organization that wants to remain sustainable in its business, one of the important factors is to have good data management. Many of the organizations that have big data are "rich in information but poor in knowledge". Data processing that is complex and difficult to predict is a challenge in the field of data mining. A lot of data is available in the education system but there is a lack of effective analysis tools to find hidden relationships and trends in a data. The use of various techniques can be done to produce valuable information that can support a decision in an organization. The data calculation uses the Growth Ratio and Naïve Bayes Classifiers (NBC) data mining techniques. The method proposed in this study utilizes data obtained from the data of the national high school of the Ministry of Education and Culture. Based on the results of the analysis through the calculation of the growth ratio of grouping per region nationally with the projected percentage of absorption of high school graduates nationally of 1.2%, so that the projected absorption of high school graduates in 2024 is 7.2% and in 2025 of 8.4% and 9.6% in 2026, reflecting the increase in the projection of new students of the Universitas Terbuka that occurred in that period.

Keywords: Data Mining, Growth Ratio

1 INTRODUCTION

In the era of increasingly advanced technology, especially the development of Artificial Intelligence, it makes it easier to complete work, especially those related to data. The level of accuracy of a data is needed in daily life. Every information that exists is important to determine every decision in a certain situation. This causes the provision of information to be a means to be analyzed and summarized into knowledge from data that is useful when making a decision. Data mining is the extraction to obtain important information that is implicit and previously unknown from a data (Witten, 2011). Data mining is an activity that includes the collection and use of historical data to find regularities, patterns and relationships in large data sets (Santosa et al., 2007). Data mining is an activity that includes the collection and use of historical data to find regularities, patterns and relationships in large data sets. Data mining is also an extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) pattern of

knowledge from huge amount of data (Han et al., 2006). By using data mining, each data collection or warehouse can provide important knowledge that becomes very valuable information for an organization, such as in an educational organization.

The Open University is a Legal Entity State University established by the government with the status of an autonomous public legal entity. UT has achieved the highest class in the form of a state educational institution through Government Regulation No. 39 of 2022 concerning State Universities, Legal Entities of Open Universities, Open Universities are now increasingly strengthening their milestones to become an institution that provides distance education with increasingly high quality. Especially since UT has changed its status to a Legal Entity State University (PTN-BH). Currently, many universities are implementing education by implementing a distance learning system because after the post-COVID-19 pandemic and the community's attitude towards education that implements a distance learning system has become better. As has been conveyed by Prof. Ojat Darojat, M.Bus., Ph.D that the implementation of distance education is no longer a monopoly market of the Open University, which has become a perfect competition market for other universities. Currently, the distance education system is an education trend because with increasingly advanced technology, it can make it easier for all educational institutions to implement a distance learning system. Accuracy in market segmentation in getting prospective new students is very necessary where market segmentation is an effort needed to improve the accuracy of marketing in an organization or company so that it can make it easier for businesses to run marketing programs or approach customers (Kotler, 2012). To face technological developments and the growing development of every higher education institution in implementing a distance learning system, the Open University must be quick in responding to these changes, one of which is by utilizing its big data which is an important asset for the university. The big data must be managed properly in order to provide important information and knowledge, especially in the main sectors, namely the recruitment of prospective new students.

The recruitment process is one of the important functions of the human resources department and is the first step towards the creation of competitive strength, while in campus recruitment it is the main recruitment mode for new talented graduates (Sivaram et al., 2010). In fact, a lot of data that is Big Data in a system in an organization has not been utilized optimally and converted into more useful knowledge, especially for predicting the future. Most of them are

just statistical data to see the development of a data. For this reason, further research related to data mining on big data is needed. Data mining is a tool to strengthen the perception of data.

This study aims to analyze and process national data that has graduated from high school which is a national big data that may be rarely touched and is expected to describe an important and useful information for the recruitment process of new students at the Open University. All data in the national data of high schools is used as data attributes. In this study, analysis and processing using data mining techniques with Growth Ratio. The expected benefit of this research is to provide knowledge to the Open University in mapping prospective new students from various regions in Indonesia.

2 METHODOLOGY

The research uses a quantitative approach that aims to achieve an understanding of how it is properly constructed and built and how it works (Berndtsson et al., 2008). The type of data used in this study uses primary and secondary data. Primary data in this study were obtained directly through the process of interviews, observations, triangulation, surveys and questionnaires. As for secondary data, it was obtained through coordination with the data team of the Ministry of Education, Culture, Research and Technology. The data was processed using the growth ratio formula to see the growth of the data. The subject of this study is data operators. Meanwhile, the object of the research is national graduate data for Senior High School. In data analysis, the steps used in this study use analysis with data collection at the beginning of the study, reduction of collected data, data processing of data presentation of reduction results, and finally drawing conclusions or called verification.

In this study, the data used used is data from the Ministry of Education and Culture's National High School graduates, and is time series data from 2021 to 2023 and projection data for the next 3 years, namely 2024 to 2026. In looking at the growth of data in each series, the growth ratio calculation is used to see the percentage growth of data variables in a period of time. The growth ratio equation is as follows:

Growth Ratio = ((Present – Past)/Past) x 100%

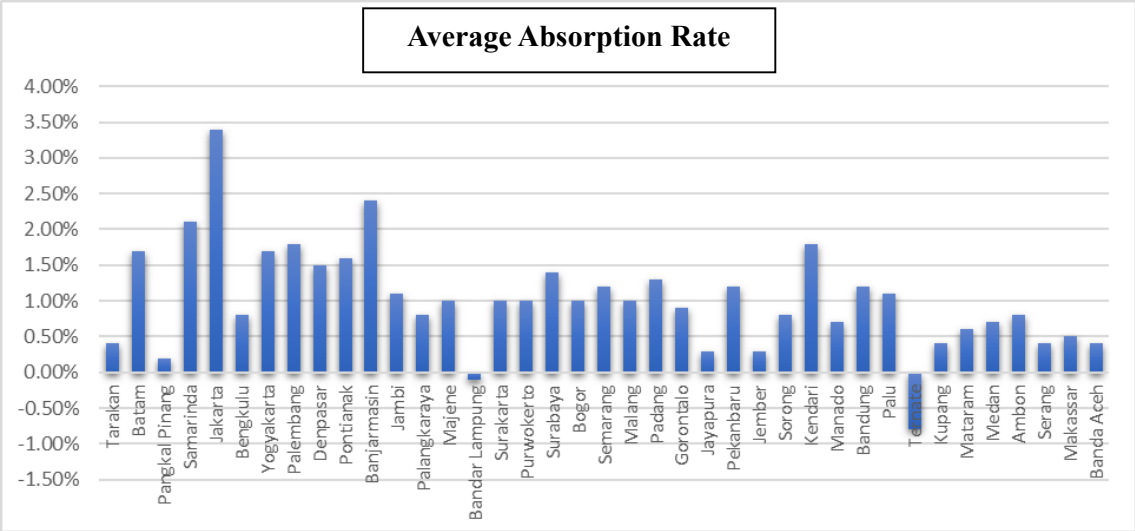
Present is the value of a current data or present value, while Past is the value of a data in the past. The result of the growth ratio is a reflection of the growth of a data.

The data collection technique was carried out by participatory observation, interviews, reviewing documents, as well as the results of triangulation or the results of the combination of data collection techniques used by the author. The data collection technique in this study is through primary data collection. The process carried out by the author is the processing of national big data, as well as reviewing documents related to high school big data. There are various techniques to perform data analysis procedures with three main methods, namely interpretive, recursive abstraction, and mechanical techniques (Oun et al., 2014).

3 FINDINGS AND DISCUSSION

The time series data processed is data on National High School graduates and data on new students of open universities for the last 3 years with the period 2021 to 2024. The data is divided into components to be used as projections for the future. The time series is based on data that has the same components in time, namely semesters and years, so it is expected to produce forecasts for the next year. The following are the results of data cleansing trends for the absorption of high school graduates per Regional UT.

Graph 1. Average Graduate Absorption Rate



Source : Data from the Ministry of Education and Culture, processed data

Based on the results of data processing shown in the average development graph of the absorption rate, the highest absorption rate of high school graduates is in the UT Jakarta area with an average absorption rate of 3.40% during the 2021-2024 period, and the second and third places are in the UT Banjarmasin and UT Samarinda areas with an average absorption

rate of 2.4% and 2.1%. But not all regional UTs experienced an increase in that period, there were several regions that experienced a decrease in the absorption rate of high school graduation during that period such as what happened at UT Ternate which experienced a decrease with an average absorption rate of high school graduation of -0.8% and UT Bandar Lampung by -0.1%. If viewed from the entire regional UT, the average high school graduation absorption rate nationally is 1.2% with details of the period 2021-2024 as shown in table 1.

Table 1. Comparison of High School Graduates and UT New Students

Senior High School Graduate Year	Amount	Number of New Students	% Absorption of High School Graduates	Average Absorption Rate
2021	3.366.588	123.078	3.7%	1.2%
2022	3.463.453	160.955	4.6%	
2023	3.449.918	207.187	6.0%	

Source: Data processing

Based on the data shown in the table above, UT's absorption rate nationally during the 2021-2023 period has increased with details in 2021 the number of graduates was 3.36 million students with the number of new students in that year of 123,078 or 3.7% of the absorption of high school graduates nationally. In 2022, the number of high school graduates increased to 3.46 million students, when viewed from the absorption trend in UT new students from the previous year there was an increase of 0.9% so that the % absorption became 4.6% in 2022. In 2023, high school graduates experienced a downward trend to 3.44 million students, when viewed from the absorption trend in UT new students, there was an increase in absorption with the number of new students at 207,187 or an increase of 1.4% from 2022 to 6.0%.

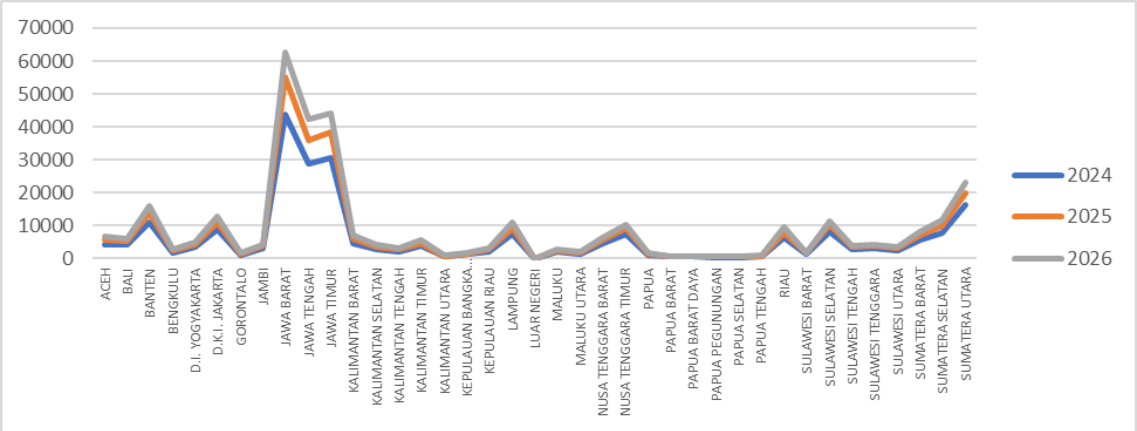
Table 2. Projection of UT High School Graduates and New Students

Year of graduation	High School Graduates	Projected Absorption of High School Graduates	UT New Student Projections
2024	3,331,960	7.2%	239,901
2025	3,579,941	8.4%	300,715
2025	3,626,150	9.6%	348,110

Source: Data processing

Based on the calculation of the projected trend of absorption of high school graduates for the next 3 years, the projected absorption value of high school graduates which will later be absorbed by UT is 7.2% in 2024 or as many as 239,901 students, the projected absorption value will increase to 8.4% in 2025 or as many as 300,715 students and 9.6% in 2025 or as many as 348,110 students. The average annual increase trend during the 2024-2025 projection period is 1.2%.

Graph 2. Projected Absorption of High School Graduates in 2024-2026



Source: Data processing

Based on the calculation of the projected trend of absorption of National High School graduates per region in Indonesia, it can be seen that National High School graduates and the highest absorption during the projection period of 2024-2026 are found in 3 regions in Java Province, namely West Java, Central Java and East Java. With a projected absorption in 2024 of 7.2% in

West Java Province, a total of 43,962 prospective new students were obtained, in 2025 it was 8.4% or a total of 55,368 prospective new students and in 2026 it was 9.6% or a total of 62,804 prospective new students in the projected absorption in that year. Meanwhile, the lowest absorption in 2024 is in the North Kalimantan region, which is a total of 653 prospective new students with a projection rate of 7.2%, in 2025 there are 863 prospective new students with a projection rate of 8.4%, and in 2026 there are 998 prospective new students with a projection rate of 9.6%.

4 CONCLUSION

Based on the results of grouping per region nationally with the projected percentage of high school graduate absorption in 2024 of 7.2% and in 2025 of 8.4% and 9.6% in 2026, reflecting the increase in new student projections that occurred in that period nationally. If you look at the regional results, the largest dominance of the projected absorption rate of new students is in the West Java, Central Java, and East Java regions, while the lowest absorption projection is in the Papua and North Kalimantan.

With the information generated from this study, recommendations can be made that in increasing the absorption rate of students in the region there are several ways that can be considered.

a. Increase socialization and promotion by conducting intensive information campaigns regarding study programs, facilities, and learning benefits in the region. Use social media, local media, and community events to reach prospective students. Then collaborate with high schools by holding seminars, workshops, or presentations that explain the benefits of continuing higher education at the Open University.

b. Provision of Scholarships and Financial Aid Local scholarships by providing special scholarships or financial aid for prospective students from underserved areas. This could include tuition discounts, living allowances, or transportation cost assistance. Then the Fee Waiver Program by offering tuition fee waiver programs for outstanding students or students who are experiencing financial difficulties.

c. Improving the quality and relevance of study programs in the form of Curation of Study Programs by offering study programs that are relevant to the needs and interests of the local job market. Do research on fields that have good job prospects in the region. Then

use the latest curriculum periodically to stay relevant to the latest industrial and technological developments.

d. Development of infrastructure and facilities by improving educational facilities such as laboratories, libraries, and study rooms. Adequate facilities can attract more prospective students. Then improve connectivity such as fast internet access and transportation services to facilitate student mobility.

e. Collaboration with industry and the government in the form of building partnerships with local companies and industries to create internship programs, industry visits, and job opportunities for students.

For regions that are already high, they can manage and maintain superior quality of education while continuing to attract students and maintain the good reputation of the educational institution

REFERENCES

Alvi, M. (2016). *A manual for selecting sampling techniques in research*. https://mpra.ub.uni-muenchen.de/id/eprint/70218

Berndtsson, M., Hansson, J., Olsson, B., & Lundell, B. (2008). Developing your Objectives and Choosing Methods. *Thesis Projects*, 54–70. https://doi.org/10.1007/978-1-84800-009-4_8

Bustami, B. (2013) Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI : *Jurnal Penelitian Teknik Informatika,* Vol. 3, No.2, Hal. 127-146

Dawson, C. W. (2009). Projects in Computing and Information Systems A Student's Guide (2nd ed.). Great Britain: Pearson Education

Eka Setyaputri, K., Wildan Fadholi, M., Cahyana Mukti, F., & Muhammadiyah Paguyangan Brebes, S. (2021). Rancang Bangun Sistem Pelaporan A1 Berbasis Web Di Badan Pengawas Pemilu Kabupaten Brebes. *Jurtisi.Stmikmpb.Ac.IdKE PutriJURNAL TEKNIK INFORMATIKA DAN SISTEM INFORMASI, 2021•jurtisi.Stmikmpb.Ac.Id*, *1*(1), 1–7.

Han, J., Kamber, M., Berzal, F., & Marín, N. (2002). Data Mining: Concepts and Techniques. *SIGMOD Record*, *31*(2), 66–68. https://doi.org/10.1145/565117.565130

Han, J., Kamber, M., kaufmann, D. (2006). Concepts and techniques. *Liacs.Leidenuniv.Nl*.

Jantawan, B., in, C. (2007). A comparison of filter and wrapper approaches with data mining

techniques for categorical variables selection. *Citeseer*, *3297*.

K. E. Setyaputri, Megawati, M. W. Fadholi, and F. C. Mukti, "Rancang Bangun Sistem Pelaporan a1 Berbasis Web Di Badan Pengawas Pemilu Kabupaten Brebes," *Jurnal Teknik Informatika dan Sistem Informasi (JURTISI)*, vol. 1, no. 1, pp. 1–7, 2021

Kotler, P. (2012). *Kotler on marketing*.

Negara, E. S., Kerami, D., Wiryana, I. M., & Maulana Kusuma, T. B. (2017). Researchgate data analysis to measure the strength of Indonesian research. *Eprints.Binadarma.Ac.IdSN Edi, K Djati, MK TubagusFar East Journal of Electronics and Communications, 2017•eprints.Binadarma.Ac.Id*, *17*(5), 1177–1183. https://doi.org/10.17654/EC017051177

Oun, M., (2014). Qualitative research method summary. *Jmest.Org*, *1*, 3159–3199.

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.

R. Amanda, E. S. Negara, A. Info, D. Mining, E. Method, and Y. D. Classification, "Analysis and Implementation Machine Learning for YouTube Data Classification by Comparing the Performance of Classification Algorithms," Jurnal Online Informatika vol. 5, no. 1, pp. 61-72, 2020, doi: 10.15575/join.v5i1.505

Santosa, B., Conway, T., Biomedicine, T.. (2007). A hybrid knowledge based-clustering multi-class SVM approach for genes expression analysis. *SpringerB Santosa, T Conway, T TrafalisData Mining in Biomedicine, 2007•Springer*.

Santosa, B., Conway, T., & Trafalis, T. (2007). A hybrid knowledge based-clustering multi-class svm approach for genes expression analysis. *Springer Optimization and Its Applications*, *7*, 231–274. https://doi.org/10.1007/978-0-387-69319-4_15

Senika, A. (2022). Implementasi Metode Naïve Bayes Dalam Penilaian Kinerja Sales Marketing Pada PT. Pachira Distrinusa. *Ejurnal.Stmik-Budidarma.Ac.IdA Senika, R Rasiban, D IskandarJurnal Media Informatika Budidarma, 2022•ejurnal.Stmik-Budidarma.Ac.Id*. https://doi.org/10.30865/mib.v6i1.3331

Singh, S. (2013). Performance analysis of engineering students for recruitment using classification data mining techniques. *Ijcset.Net*. Retrieved November 29, 2023, from http://ijcset.net/docs/Volumes/volume3issue2/ijcset2013030202.pdf

Sivaram, N. (2010). Applicability of clustering and classification algorithms for recruitment

data mining. *Citeseer*, *4*(5), 975–8887.

Suwayudhi, S., Eka, I. (2022). Teknik klasifikasi dalam memprediksi penerimaan siswa baru menggunakan metode Naïve Bayes. https://doi.org/10.55123/jomlai.v1i3.963

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*, *5*(4), 2493–2518. https://doi.org/10.1214/11-AOAS493.