pijmath
Pattimura International Journal of Mathematics (PIJMath)
Volume 02 issue 01, pages 01-06, May 2023
E-ISSN: 2830-6791
RESEARCH ARTICLE ◙ *OPEN ACCESS*
DOI: https://doi.org/10.30598/pijmathvol2iss1pp01-06

Application of Classification Data Mining Technique for Pattern Analysis of Student Graduation Data with Emerging Pattern Method

Aditya Handayani[1], Neva Satyahadewi[2*], Hendra Perdana[3]

[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Tanjungpura University
Prof. Dr. H. Hadari Nawawi Street, Pontianak 78124, Indonesia.

*Corresponding author's e-mail: * neva.satya@math.untan.ac.id*

Article History

Received : 15th February 2023
Accepted: 02nd April 2023
Published: 01st May 2023

Keywords
Data mining;
Graduation;
Classification;
Emerging patterns;

ABSTRACT

Data mining has been applied in various fields of life because it is very helpful in extracting information from large data sets. Student graduation data is one example of data that can be extracted for information and become a recommendation. This study used a classification data mining technique to extract information from the student graduation data. The classification technique used was the Emerging Pattern method to search for patterns in the student graduation data. The data in this study were graduation data for students of the Statistics Study Program, Faculty of Mathematics and Natural Sciences, Tanjungpura University, from 2013-2018. The sample data used amounted to 186 records. Attributes used in this study include as many as four attributes, including gender, batch, GPA, and TUTEP scores. This research began by finding the class and frequency values obtained. It was continued by calculating each item set's support, growth rate, and confidence values. This study obtained the highest confidence value among all the attributes owned, namely 91% in the 2013 batch itemized list and the 2018 batch. Female students dominated the class attribute. TUTEP dominated the TUTEP value attribute with a score of 425, and the GPA attribute of 3.51-4.00 dominated the class with a confidence value of 60%.

[1]*How to cite this article:*
A. Handayani, N. Satyahadewi, and H. Perdana, "*APPLICATION OF CLASSIFICATION DATA MINING TECHNIQUES FOR POLA ANALYSIS OF STUDENT GRADUATION DATA BY EMERGING PATTERN METHOD*", *Pattimura Int. J. Math. (PIJMATH).*, vol. 02, iss. 01, pp. 01-06, May 2023.

1. Introduction

One indicator of a university's success in teaching and learning is the graduation rate. A high graduation rate can be considered an achievement for a university. Universities that have graduated with high competitiveness certainly have good quality in terms of the timeliness of graduation [1]. Students are said to graduate on time at Tanjungpura University if they can complete the study period within four years. In practice, not all students can complete their studies within four years.

Many factors affect the length of the student study period, both internal and external. The diverse characteristics of students also make the length of the student study period vary [2]. Therefore, universities must be able to evaluate their students to minimize the factors that become obstacles to the timeliness of student graduation [3]. It is a reason for universities to be able to make the right decisions to evaluate the causes of the inaccuracy of students' graduation time.

Making the right decision requires sufficient information to analyze further the factors that inhibit the timeliness of student graduation [4]. The necessary information can be obtained by analyzing the data stored in the academic information system, one of which is student graduation data. After analyzing the student graduation data, universities can find out early on the graduation rate of their students [5]. Therefore, the right data analysis technique is needed to extract hidden information from this student graduation data [6].

Manual analysis techniques certainly cannot extract information from large amounts of data. Therefore, the analysis technique can be used in data mining [6]. Data mining is a branch of science that can help facilitate decision-making in analyzing and extracting data [7]. Using data mining techniques as an analysis technique is expected to provide valuable information and knowledge previously hidden in the data [8]. Based on research [9], data mining has six analysis techniques: description, clustering, classification, prediction, and estimation. The data mining technique used in the research that the researchers conducted was the classification data mining technique. Based on research [10], classification is the process of finding models or functions that can distinguish data classes that aim to predict objects that are not yet known to the class label.

Based on this description, this research used classification techniques with the Emerging Pattern method to find student graduation patterns in the Statistics Study Program at Tanjungpura University. The expected results of this research were to help facilitate management in processing graduation data and analyzing and predicting student graduation. Then, these results can be used in helping the decision-making process when making policies for students of the Statistics study program at Tanjungpura University who are predicted not to graduate on time as a preventive measure against this.

2. Research Methods

2.1 Emerging Pattern (EP)

Based on research [11] Emerging Pattern (EP) is a knowledge discovery from capturing database trends that appear when applied in a database or capturing differences between data classes when applied to data sets and classes. This Emerging Pattern (EP) captures significant changes and differences between datasets. Emerging Pattern (EP), which has a large growth rate (GR) value, can distinguish the characteristics of two datasets and can build very interesting classifiers [12]. According to research [13], the Emerging Pattern method is divided into discriminating between two datasets and classifying more than two datasets. Emerging pattern uses support, growth rate, and confidence values in the analysis process.

Based on research [14], support is also referred to as the support of the dataset. The support value is useful for showing the tendency of an item set, whether it is more dominant in dataset 1 or dataset 2. The support of itemset *X* is contained in dataset *D*, denoted in the support value. $support_D(X)$. The following is the equation for finding the support value in Emerging Pattern [13]:

$$supp_D(X) = \frac{(count_D(X))}{|D|} \quad (1)$$

X is an item set or pattern. *D* is a dataset, and |D| is the total data in the dataset. Furthermore, in **Equation (1)**, where $support_D(X)$ is the support in dataset *D* that contains itemset *X*, $count_D(X)$ is the amount of data in dataset *D* that contains itemset *X*. The support value will be used to find the growth rate (GR) value [13].

The growth rate is the growth of itemset from one set to another (two-ratio support) [13]. The growth rate value is used as a measure of itemset growth in a particular class. Emerging Patterns with a large growth rate (GR) value can be used as a distinguishing characteristic of two datasets and can form an interesting classification [11]. The following is the equation for finding the growth rate value [13]:

$$Growth\ rate\ (X) = \frac{supp_{D_2}(X)}{supp_{D_1}(X)} \quad (2)$$

Based on **Equation (2)**, it is known that the growth rate (GR) is the result of the division between the support value of dataset 2 divided by the support value of dataset 1. Growth rate (GR) has a term called Jumping Emerging Pattern. JEPs (jumping emerging patterns) occur when the support value of dataset 1 or as the denominator in equation (2) is 0 so that the growth rate (GR) will be (∞). Because the growth rate (GR) is worth (∞), itemsets containing Jumping Emerging Patterns have no confidence value [12].

The level of truth of the pattern formed is referred to as confidence. The confidence value shows the relationship between itemsets with the dataset formed. The confidence value is also used as a determinant of interesting patterns in classification with Emerging Patterns [14]. Based on research [13], finding the confidence value uses the following **Equation (3)**:

$$confidence\ (X) = \frac{GrowthRate\ (X)}{GrowthRate\ (X)+1} \tag{3}$$

Based on **Equation (3)**, it is known that the confidence value is the result of dividing the growth rate value divided by the growth rate itself plus 1.

2.2 Research Data

This research data is secondary: student graduation data from the Academic Faculty of MIPA UNTAN. The data used amounted to 186 data from the 2013-2018 batch. The attributes used are gender, batch, study period, predicate, and TUTEP score.

2.3 Preprocessing

Preprocessing removes inconsistent and noisy data, duplicates data, improves data, or can be enriched with relevant external [10]. Preprocessing carried out in this study was to form data that was originally quantitative data into qualitative data in the form of categories. An example of the formation of these data is the attributes of the initial study period. The data is in the form of years, months, and days and then converted into on-time and off-time categories to adjust to this research.

2.4 Descriptive Statistics

Descriptive statistics were carried out to determine the description of the data used. A description is a brief or concise description of the information from the data [15]. The data used amounted to 186, and the attributes used were gender, batch, TUTEP score, and GPA. These attributes were divided into several category classes. These attributes were used to see the pattern of graduation data on the length of the student study period. More details about the student graduation data used can be seen in **Table 1** below:

Table 1. Descriptive Statistics of Student Graduation Data of Statistics Study Program

No.	Characteristics	Classification	Total	Percentage
1	Gender	L	38	20,43%
		P	148	79,56%
2	Batch	2013	20	10,75%
		2014	46	24,73%
		2015	41	22,04%
		2016	34	18,27%
		2017	25	13,44%
		2018	20	10,75%
3	Study Period	TW	65	34,94%
		TTW	121	65,05%
4	Predicate	M	3	1,61%
		SM	120	64,51%
		DP	63	33,87%
5	TUTEP score	< 425	14	7,53%
		≥ 425	172	98,01%

Based on **Table 1**, some information is obtained about the graduation data of the Statistics Study Program students. From the **Table 1**, it is known that for gender characteristics, the most or large number is owned by the female gender, namely 148 people, with a percentage of 70.56%. The characteristics batch with the largest number is the 2014 batch of 46 people, with a percentage of 24.73%. Furthermore, the study period is taken within ≤ 4 years and is said to

be not on time when the study period is taken > 4 years. The largest number for this characteristic belongs to the TTW (not-on-time) class, namely 121 people with a percentage of 65.05%.

3. Results And Discussion

The data used in data mining processing in this study is 186 graduation data. Based on the equations previously described, mining results were obtained for each attribute used in this study. The mining results contain the support value, growth rate (GR), and confidence. Based on research [12], the support value is used to see the tendency of data, whether it is more dominant to dataset 1 or dataset 2. The growth rate (GR) value is used to see class differences. The confidence value is used to conclude whether the patterns found are interesting or not to be used as recommendations in decision-making [13]. This study's mining results for the gender attribute can be seen in **Table 2** Below:

Table 2. The Table of Gender Attribute Mining Result

Itemset	*Class TW*	*Freq TW*	*Class TTW*	*Freq TTW*	*Support D1 (X)*	*Support D2 (X)*	*Growth Rate (X)*	*Confidence*
L	65	13	121	25	0.2	0.207	1.033	0.508
P	65	52	121	96	0.8	0.793	0.992	0.498

Based on **Table 2** It is known that the pattern obtained is that students with the female gender are more dominant in graduating on time compared to male students. Female students have a support value of 80% in the on-time class with a confidence value of 49.79% with a growth rate of 0.992. Male students have a greater support value in the not-on-time class, which is 20.66%, with a growth rate of 1.033 and a confidence value of 50.81%. Based on the confidence value obtained, it is known that the male confidence value is greater than the female. It means that men are more dominant in graduating not on time because data set 2, which is the goal in this study, is not on time. NH mining results for the batch attribute can be seen in **Table 3**, below:

Table 3. The Table of Batch Attributes Mining Result

Itemset	*Class TW*	*Freq TW*	*Class TTW*	*Freq TTW*	*Support D1 (X)*	*Support D2 (X)*	*Growth Rate (X)*	*Confidence*
2013	65	1	121	19	0.015	0.157	10.207	0.911
2014	65	4	121	42	0.062	0.347	5.640	0.849
2015	65	18	121	23	0.277	0.190	0.686	0.407
2016	65	12	121	22	0.185	0.182	0.985	0.496
2017	65	11	121	14	0.169	0.116	0.684	0.406
2018	65	19	121	1	0.292	0.008	0.028	0.027

Based on **Table 3** It is known that the pattern obtained for the batch attribute is that the more dominant on-time batch is the batch of 2018 and the batch of 2015, with a support value in the on-time class above 20%. The not-on-time class is dominated by the batch of 2014, with a support value for the not-on-time class of 34.7% with a growth rate of 5,640. The highest confidence value is in the 2013 batch itemset, which is 91.08% with an untimely class growth rate of 10,207, and the 2018 batch dominates the on-time class. The next mining result is for the TUTEP value attribute. This TUTEP score is categorized into two classes, namely TUTEP scores ≥ 425 and < 425. TUTEP scores were categorized based on the TUTEP graduation standard at Tanjungpura University. Namely, if the score is less than 425, it is declared not to pass TUTEP. The mining results for the TUTEP score attribute can be seen in **Table 4**, Below:

Table 4. The table of TUTEP Value Attributes Mining Result

Itemset	*Class TW*	*Freq TW*	*Class TTW*	*Freq TTW*	*Support D1 (X)*	*Support D2 (X)*	*Growth Rate (X)*	*Confidence*
≥ 425	65	63	121	109	0.969	0.901	0.929	0.482
< 425	65	2	121	12	0.031	0.099	3.223	0.763

Based on **Table 4** The pattern obtained for the TUTEP score itemset is that the on-time class is dominated by students with a TUTEP score ≥ 425 with a support value of 96.9%, a class growth rate of 0.929, and a confidence value of 48.2%. TUTEP scores with scores < 425 dominate the not-on-time class by 9.9%, with a class growth rate of 3.223 and a confidence value of 76.3%. The last mining result is for the predicate attribute. The results of mining the predicate attribute can be seen in **Table 5**, below:

Table 5. The Table of Predicate Attribute *Mining* Result

Itemset	*Class TW*	*Freq TW*	*Class TTW*	*Freq TTW*	*Support D1 (X)*	*Support D2 (X)*	*Growth Rate (X)*	*Confidence*
2,00-2,75	65	0	121	3	0.0	0.025	-	-
2,76-3,50	65	26	121	94	0.4	0.777	1.9 2	0.660
3,51-4,00	65	39	121	24	0.6	0.198	0.331	0.248

Based on this **Table 5**, it is known that the GPA range of 2.00-2.75 is undefined because the support value that becomes a divider is 0, so the GPA range 2.00-2.75 itemset has no confidence value. Because it does not have a confidence value, no conclusion can be drawn from the itemset. Furthermore, the on-time class is dominated by students with a GPA range of 3.51-4.00 with a growth rate of 0.331. The highest confidence value is obtained by the GPA range of 2.76-3.50, which is 66%.

4. Conclusions

Based on the results and discussion, the following conclusions can be drawn:

1. The pattern obtained is that students with the female gender are more dominant in graduating on time compared to male students. Female students have a support value of 80% in the on-time class with a confidence value of 49.79% with a growth rate of 0.992. Male students have a greater support value in the not-on-time class, which is 20.66%, with a growth rate of 1.033 and a confidence value of 50.81%.
2. The more dominant on-time batch is the 2018 and the batch 2015, with a support value in the on-time class above 20%. The not-on-time batch is dominated by the batch of 2014, with a support value of the not-on-time class of 34.7% with a growth rate of 5,640. The highest confidence value is in the 2013 batch itemset, which is 91.08%, with a growth rate of the untimely class of 10,207.
3. The pattern obtained for the TUTEP score itemset is that the on-time class is dominated by students with a TUTEP score ≥ 425 with a support value of 96.9% and a class growth rate of 0.929, a confidence value of 48.2%. TUTEP scores with scores < 425 dominate the untimely class by 9.9%, with a class growth rate of 3.223 and a confidence value of 76.3%.
4. The on-time class is dominated by students who obtain a GPA in the range of 3.51-4.00 with a growth rate of 0.331. The highest confidence value is obtained by the GPA range of 2.76-3.50, which is 66%.

References

[1] G. Dong and J. Li, "Efficient Mining of Emerging Patterns: Discovering Trends and Differences," p. 10.

[2] Asriningtias, Y., and Mardhiyah, R., Aplikasi Data Mining untuk menampilkan Informasi Tingkat Kelulusan Mahasiswa [Data Mining Application to Display Student Graduation Rate Information], Journal of Informatics, 8 (1), 2014.

[3] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[4] M. Irfan, "Analisa Pola Asosiasi Jalur Masuk Terhadap Kelulusan Mahasiswa Dengan Menggunakan Metode Fold-Growth (Studi Kasus Fakultas Sains Dan Teknologi)," [Analysis of Association Patterns of Entry Paths to Student Graduation By Using The Fold Growth Method (Case Study of The Faculty of Science and Technology)] no. 2, p. 19, 2015.

[5] J. Han and M. Kamber, *Data mining: concepts and techniques*, 2nd ed. in The Morgan Kaufmann series in data management systems. Amsterdam ; Boston : San Francisco, CA: Elsevier ; Morgan Kaufmann, 2006.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," p. 169.

[7] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian, "Mathematical Programming for Data Mining: Formulations and Challenges," *Inf. J. Comput.*, vol. 11, no. 3, pp. 217–238, Aug. 1999, doi: 10.1287/ijoc.11.3.217.

[8] Suaidah, Warnars, H., L., H., S., and Damayanti, Implementasi Supervised Emerging Patterns pada Sebuah Atribut: (Studi Kasus Anggaran Pendapatan dan Belanja Daerah (APBD) Perubahan pada Pemerintah DKI Jakarta) [Implementation of Supervised Emerging Patterns on an Attribute: (Case Study of Regional Expenditure Budget (APBD) Changes in the DKI Jakarta Government)]. TINF-015, 2018.

[9] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," [Implementation of Data Mining to Evaluate Student Academic Performance Using the Naïve Bayes Classifier Algorithm]. *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 151, Jan. 2018, doi: 10.24076/citec.2017v4i2.106.

[10] D. A. Yosepta and T. Aprilianto, "Analisa Pola Kelulusan Mahasiswa Pada Sekolah Tinggi Manajemen Informatika & Komputer Asia Malang Dengan Menggunakan Algoritma Iterative Dichotomiser 3 (ID3)," [Analysis of Student Graduation Patterns at the Asian School of Informatics and Computer Management Malang By Using the Iterative Dichotomiser 3 (ID3)] vol. 3, p. 9, 2017.

[11] G. Dong, X. Zhang, L. Wong, and J. Li, “CAEP: Classification by Aggregating Emerging Patterns,” p. 15.

[12] N. Nuruliyani and H. L. H. S. Warnars, “Prototype Data Mining Pola Jabatan Fungsional Dosen Menggunakan Teknik Emerging Pattern: Studi Kasus Universitas Mercu Buana,” [Prototype Data Mining Pattern of Lecturer Functional Position Using Emerging Pattern Technique: Case Study of Mercu Buana University] *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, vol. 7, no. 2, pp. 211–224, Sep. 2019, doi: 10.33558/piksel.v7i2.1842.

[13] I. Farida and S. W. H. L. Hendric, “Prediksi Pola Kelulusan Mahasiswa Menggunakan Teknik Data Mining Classification Emerging Pattern,” [Prediction of Student Graduation Patterns Using Data Mining Classification Emerging Pattern Techniques] *PETIR*, vol. 12, no. 1, Apr. 2019, doi: 10.33322/petir.v12i1.414.

[14] Y. T. Utami, “Penerapan Supervised Emerging Patterns Untuk Multi Atribut Pada Data Online Izin Usaha Pertambangan di Indonesia (Studi Kasus: Eiti Indonesia),” [Application of Supervised Emerging Patterns for Multiattribute on Online Data of Mining Business Licence in Indonesia (Case Study: Eiti Indonesia)] p. 7, 2016.

[15] S. Angriani, S. Neva, and P. Hendra, “Penerapan Data Mining Untuk Memprediksi Status Kelulusan Pada Jalur SNMPTN Menggunakan Algoritma Naïve Bayes Classifier,” 2022.