E-ISSN 2503-1430
ISSN 2541- 3090

Journal of Industrial Engineering Management
(JIEM Special Edition Seminar Nasional Teknik Manajemen Industri (SENTRA))

APPLICATION OF PREDICTION TIME OF GRADUATION USING THE NAÏVE BAYES ALGORITHM WITH THE PYTHON PROGRAM

Priskila Christine Rahayu[(1)], Eric Jobiliong[(2)], Antonny[(3)]
[(1), (2), (3)] Industrial Engineering Study Program, Universitas Pelita Harapan
MH Thamrin Boulevard 1100, Lippo Village, Tangerang
priskila.christine@uph.edu

Article history:
Submit 17 Oktober 2020
Received in from 20 Oktober 2020
Acceted 20 November 2020
Avilable online 4 Maret 2021

ABSTRACT

Accreditation is a process to ensure the quality of a university and study program. There are several factors that determine the quality standard of accreditation. One of them is the time of graduation. However, there is no means that can be used to predict early student graduation time. Therefore, this study aims to create a means that can predict early graduation time. In this study, data mining methods were used, namely the Naïve Bayes algorithm. After that, data processing and application development will be carried out using the Python program. The data used in the data mining process is three years of historical data and the data used for the trial are active student data for the second and third years. There are 5 types of patterns with an accuracy value of 81%, 87%, 92%, 92%, and 95%.

Keywords: Data mining, Naïve Bayes, Python, Application, Prediction Time.

Published By:
Fakultas Teknologi Industri
Universitas Muslim Indonesia
Address :
Jl. UripSumoharjo Km. 5 (Kampus II UMI)
Makassar Sulawesi Selatan.
Email :
Jiem@umi.ac.id
Phone :
+6281341717729
+6281247526640

DOI :http://dx.doi.org/10.33536/jiem.specialedition.773

I. INTRODUCTION

The Ministry of Research, Technology and Higher Education of the Republic of Indonesia (2016) established accreditation regulations as an external quality assurance system to ensure the quality of study programs and tertiary institutions externally both in the academic and non-academic fields to protect the interests of students and society. In accordance with the Accreditation Instrument Policy of the National Higher Education Accreditation Board (2019), that every five years colleges and study programs must be accredited to determine their eligibility based on the criteria in the National Higher Education Standards. The scope of these standards includes 9 standards, namely graduate competence, learning content, learning process, learning assessment, lecturers and education staff, learning facilities and infrastructure, learning management, and learning financing. Graduate competency standards are used as the main reference for the development of the other eight standards. For this reason, every university strives to have good graduate competencies. One way is to produce graduates who are on time.

Through this research, it is hoped that there will be a means that can be used to predict the timeliness of student graduation. The facility is designed in the form of an application that can predict early on the timing of graduation for students, so that the prediction results can be used to make suitable teaching-learning plans for students and lecturers.

The design of this application requires a lot of historical data to form a pattern. Furthermore, this pattern is used to predict future data using current data. Rohmawati (2017) states that prediction is a process of estimating things that can often happen in the future systematically based on information from historical data and current data so that differences or differences in prediction results can be minimized. Prediction is the same as forecasting.

The process of determining patterns for future data using several variables is a data mining process. Septiani (2017) states that the data mining process is an activity consisting of collecting and using historical data used in determining a relationship pattern in large amounts of data. Data mining is also known as Knowledge Discovery in Database (KDD) (Larose, 2005). Meanwhile, Han and Kamber (2007) state data mining as a process to extract or mining several data in the form of knowledge that has not been known manually.

II. METHODOLOGY

The factors that influence students to graduate on time are quantitative data and qualitative data. This study uses quantitative data to measure students' early graduation time.

In this study, the data collected is data from students who have graduated for the last three years (historical data) as initial data in making applications, then testing applications that have been made using second- and third-year student data. The data used includes the value of each student starting from the grade at high school to the IP obtained in each semester starting from the first semester to the seventh semester. Existing data will later be used in the data mining process. After collecting data, determining the attributes that can affect student graduation time such as the majors taken by each student at high school, high school grades related to subjects in the study program, area of origin of the school, gender, GPA, number of credits passed in each semester, specialization while at university.

Han (2011) states 7 steps of the knowledge discovery process, namely:

1. Data Cleaning is a step-in which data that is deemed unimportant and inconsistent data is deleted. Data that is considered unimportant and inconsistent data is discarded to increase the accuracy of the data mining results. Data cleaning affects system performance because the amount and complexity of data is reduced.
2. Data Integration is a step where data sources are combined. Data integration is carried out on attributes that have unique entities such as name, parent number, and so on.
3. Data Selection is a step where data retrieval is carried out related to the purpose of analysis from the database.
4. Data Transformation is a step in changing the form of data into a form suitable for mining through an aggregation process
5. Data Mining is an important process in extracting data using certain methods.
6. Pattern Evaluation is a step to determine a pattern that is considered in accordance with science.
7. Knowledge Presentation is a step in visualizing the techniques used to provide an overview to the user.

Several algorithms that can be used for data

mining processes, such as Decision Tree, Bayes Theorem, Neural Network, and others. According to the results of research by Xhemali et al (2009) states that the Bayes Theorem has a better level of accuracy by using less data than the Decision Tree and Neural Network. One application of the Bayes Theorem in classification is the Naïve Bayes algorithm. The Naive Bayes Algorithm is a simple probability classification method for calculating several probabilities by adding up the frequency and value combinations from a series of existing data. In addition, through research conducted by Rennie et al. (2003), it was found that Naive Bayes is a text classification algorithm that is fast, easy to apply, and quite sophisticated. The Naive Bayes algorithm predicts an opportunity in the future based on historical data such as in Formula (1) and (2) by assuming that each attribute or variable is independent or not interdependent from each condition or event.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

X: Data with unknown class

H: The data hypothesis is a certain class

P (H | X): The probability of hypothesis H with condition X or posteriori probability

P (H): Probability hypothesis H or prior probability

P (X | H): Probability of X based on the conditions in the hypothesis H

P (X): Probability X

$$P(X_i = xi|Y = yi) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(Xi-\mu ij)^2}{2\sigma ij^2}} \quad (2)$$

P: Probability

Xi: Attribute to I

xi: Value of attribute value to i

Y: Class to be defined

yi: Sub class Y to be defined

μ: Average value of all attributes

σ: Standard deviation which is the variant of all attributes

III. RESULTS AND DISCUSSION

The research was carried out sequentially according to the stages described in the methodology, which gave the following results:

At the Data Cleaning stage, data cleaning was carried out by removing data that were deemed not to have complete data such as the scores for each subject in SMA, both at the first and second levels. Each group of values represents the cluster variable (VC) in the study. The results of data acquisition were 201 students, and the results of data cleaning were 130 students.

In the Data Integration stage, data merging was carried out from several data sources, namely the high school student database as the VC1 group and the university student database as the VC2 group. Group VC1 is gender (X1 = male, female), school hometown (X2 = a, b, c, d, e, f, g, h, i, j, k, l), major in high school (X3 = Science, Social Sciences), specialization in study programs (X4 = a, b). Group VC2 is the high school grade (X5 = a, b), the number of credits from semester 1 to semester 7 (X6), GPA from semester 1 to semester 7 (X7), and on time graduation (Y1 = on time, Y2 = not on time).

At the Data Selection stage, it is carried out in accordance with the research objectives. Indirectly, this stage of research has been carried out at the same time during the data cleaning and data integration processes.

At the Data Transformation stage, the data is converted into a frequency form for each VC as in Table 1. The results obtained for the data frequency of Y1 are 70 and the data frequency of Y2 are 60.

Table 1 Results of Data Transformation and Data Mining of Group VC1

		Frequency Yi		Probability	
Frequency Xi		1(on time)	2(not on time)	on time	not on time
1	men	41	50	0.59	0.83
	women	29	10	0.41	0.17
2	a	0	1	0.00	0.02
	b	25	18	0.36	0.30
	c	4	12	0.06	0.20
	d	6	3	0.09	0.05
	h	5	1	0.07	0.02
	i	0	2	0.00	0.03

	j	1	3	0.01	0.05
	k	8	2	0.11	0.03
	l	21	18	0.30	0.30
3	**a**	69	57	0.99	0.95
	b	1	3	0.01	0.05
4	**a**	13	10	0.19	0.17
	b	57	50	0.81	0.83

The next stage is the Data Mining stage. At this stage, the results are obtained in Table 1, namely the probability of each VC1 as P (Xi | H), and the mean and standard deviation values of each VC2 in Table 2.

Table 2 Results of Data Mining of Group VC2

		Value	
Xi		**on time**	**not on time**
5	**mean**	80	76
a	**std dev**	7.45	6.24
	mean	80	76
b	**std dev**	7.16	5.99

		Value	
Xi		**on time**	**not on time**
6	**mean**	15.87	13.90
1	**std dev**	1.94	3.04
	mean	29.24	25.52
2	**std dev**	3.10	4.35
	mean	44.84	39.02
3	**std dev**	3.44	4.26
	mean	61.93	54.18
4	**std dev**	4.24	4.89
	mean	78.94	71.05
5	**std dev**	5.32	6.35
	mean	91.30	81.98
6	**std dev**	4.31	5.34
	mean	109.40	99.33
7	**std dev**	4.65	5.74

		Value	
Xi		**on time**	**not on time**
7	**mean**	2.97	2.81
1	**std dev**	0.25	0.29
	mean	2.96	2.68
2	**std dev**	0.25	0.23
	mean	2.98	2.71
3	**std dev**	0.25	0.19
	mean	3.01	2.68
4	**std dev**	0.26	0.18
	mean	3.01	2.66
5	**std dev**	0.28	0.18
	mean	3.06	2.68
6	**std dev**	0.28	0.18
	mean	3.08	2.68
7	**std dev**	0.27	0.17

At the Pattern Evaluation stage, the data test of one of the active students was carried out by calculating the P value (Xi | H) using Formula (2) and the data in Table 1 and Table 2. The results obtained in Table 3. Then calculated the probability of passing the student on time as P (H1) of 0.54 and the probability of not passing on time as P (H2) of 0.46.

Table 3 Example of Calculating P (Xi | H)

P(Xi|Yi)	**On time**	**Not on time**
Women	0.41	0.17
Hometown i	0.30	0.30
Track Science	0.99	0.95
Lesson Value 1	0.05	0.01
Lesson Value 2	0.06	0.01
sks semester 1	0.15	0.09
sks semester 2	0.02	0.01
sks semester 3	0.20	0.09
IPK semester 1	0.14	0.43

IPK semester 2	0.40	0.83
IPK semester 3	0.69	0.69

Based on several test data, 5 predictive patterns were obtained, namely for students who had passed 3-7 semesters. Each pattern has an accuracy value of 81%, 87%, 92%, 92% and 95%, respectively. Figures 1 to 5 are the results of data processing using Python coding, namely the Integrated Development Environment in the form of Jupyter Notebook.

```
# menghitung nilai akurasi

from sklearn.metrics import classification_report

print(classification_report (y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.93      0.76      0.84        17
           1       0.67      0.89      0.76         9

    accuracy                           0.81        26
   macro avg       0.80      0.83      0.80        26
weighted avg       0.84      0.81      0.81        26
```

Figure 1 Prediction 3 Semester

```
# menghitung nilai akurasi

from sklearn.metrics import classification_report

print(classification_report (y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.91      0.87      0.89        23
           1       0.82      0.88      0.85        16

    accuracy                           0.87        39
   macro avg       0.87      0.87      0.87        39
weighted avg       0.87      0.87      0.87        39
```

Figure 2 Prediction 4 Semester

```
# menghitung nilai akurasi

from sklearn.metrics import classification_report

print(classification_report (y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.92      0.96      0.94        23
           1       0.93      0.88      0.90        16

    accuracy                           0.92        39
   macro avg       0.93      0.92      0.92        39
weighted avg       0.92      0.92      0.92        39
```

Figure 3 Prediction 5 Semester

```
# menghitung nilai akurasi

from sklearn.metrics import classification_report

print(classification_report (y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.92      0.96      0.94        23
           1       0.93      0.88      0.90        16

    accuracy                           0.92        39
   macro avg       0.93      0.92      0.92        39
weighted avg       0.92      0.92      0.92        39
```

Figure 4 Prediction 6 Semester

```
# menghitung nilai akurasi

from sklearn.metrics import classification_report

print(classification_report (y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        23
           1       1.00      0.88      0.93        16

    accuracy                           0.95        39
   macro avg       0.96      0.94      0.95        39
weighted avg       0.95      0.95      0.95        39
```

Figure 5 Prediction 7 Semester

At the Knowledge Presentation stage, an application is made to make the prediction process easier and faster. Applications are made using the Tkinter library found in Python. Applications that have been designed have two uses, namely for lecturers and students. Applications made have a display like Figure 6 to Figure 8.

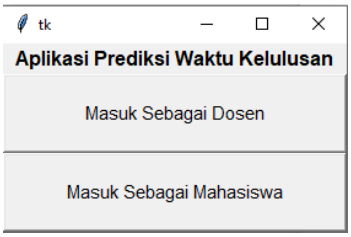

Figure 6 Homepage

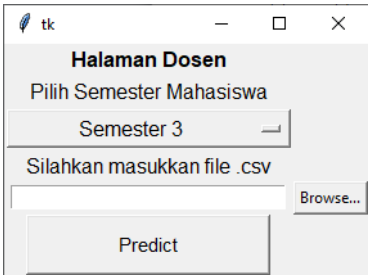

Figure 7 Lecturer Page Views

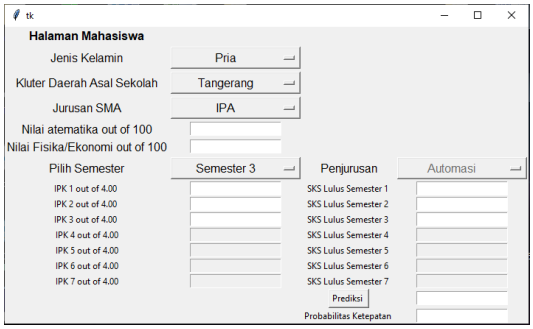

Figure 8 Student Page Views

After making a prediction application, the next stage is the trial phase. By using the application that has been made, out of a total of 33 second year students, 22 are predicted to have the potential to graduate not on time. Meanwhile, only 11 students are predicted to graduate on time. This shows that there are still about 67% of students who are predicted to graduate not on time. In order to reduce this percentage, it is hoped that students can discuss with their academic supervisors to plan teaching and learning activities to catch up.

In the third year of the total 30 students, there are 12 students who are predicted to graduate not on time. Meanwhile, 18 students are predicted to graduate on time. This shows that there are still 40% of students who are predicted to graduate not on time.

IV. CONCLUSIONS AND SUGGESTIONS

Through the research that has been done, five patterns have been obtained that can be used to predict the probability of student graduation timeliness, namely:

1. The first pattern is for students who have completed three semesters of study, with an accuracy rate of 82%.
2. The second pattern is for students who have completed four semesters of study, with an accuracy rate of 87%.
3. The third pattern is for students who have completed five semesters of study, with an accuracy rate of 92%.
4. The fourth pattern is for students who have completed six semesters of study, with an accuracy rate of 92%.
5. Meanwhile, the last pattern is for students who have completed six semesters of lecture. with an accuracy level of 95%.

Based on the results of application trials on active students, the application is quite effective for active students and academic supervisors to plan teaching and learning activities to increase GPA and graduate on time.

REFERENCES

Permenristekdikti RI Nomor 32 Tahun 2016, Akreditasi Program Studi dan Perguruan Tinggi.

Permenristekdikti RI Nomor 44 tahun 2015 dan perubahan Permenristekdikti RI Nomor 50 tahun 2018, Standar Nasional Pendidikan Tinggi.

BAN-PT, 2019, *Kebijakan Instrumen Akreditasi BAN-PT dan LAM Berbasis SN Dikti*.

Rohmawati F.; Rohman M.G.; Mujilahwati S., 2017, Sistem Prediksi Jumlah Pengunjung Wisata Wego Kec.Sugio Kab.Lamongan Menggunakan Metode Fuzzy Time Series, Journal of Informatic UNISLA Vol.2 No.2,

Septiani, W.D., 2017, Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 dan Naive Bayes untuk Prediksi Penyakit Hepatitis, Jurnal Pilar Nusa Mandiri 13, no. 1 hlm. 76-84

Larose, D.T., 2005, Discovering Knowledge in Data: An Introduction to Data Mining, John Willey & Sons, Inc.

Han, J.; Kamber M.; Pei J., 2007, Data Mining: Concepts and Techniques. 3rd edition. San Fransisco: Mofgan Kaufan Publisher.

Han, J., 2011, Data Mining: Concept and Techniques 3rd Edition, Morgan Kaufmann.

Xhemali D.; Christopher J.H.; Stone R., 2009, Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. September 2009. International Journal of Computer Science Issues 4.

Rennie J.; Shih L.; Teevan J.; Karger D.T., 2003, The Poor Assumptions of Naive Bayes Classifiers. In Proceedings of the Twentieth International Conference on Machine Learning (ICML)