Jurnal Mantik, 6 (3) (2022), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Prediction of Students Drop Out With Naïve Bayes

Nariza Wanti Wulan Sari[1], Dedy Mirwansyah[2], Fahrullah[3], Ivan Leonard Tandi[4], Ali Hakim Hadi Sopyani[5]

[1,3,4]Faculty of Computer/Study Program of Information System, Universitas Mulia, Indonesia

[2,5]Faculty of Computer /Study Program of Informatics management, Universitas Mulia, Indonesia

A R T I C L E I N F O

Article history:

Received Nov 03, 2022
Revised Nov 16, 2022
Accepted Nov 30, 2022

Keywords:

Algorithm
Drop out
Naïve Bayes
Prediction
Student

ABSTRACT

The increasing number of non-active students at Mulia University PSDKU Samarinda makes the risk of student drop out increasing. The purpose of this study is to provide predictive results for students dropping out at Mulia University PSDKU Samarinda with the Naïve Bayes algorithm, so that from an early age it can know the characteristics of students dropping out and can reduce the number of students dropping out. 2 data sets are used, namely 1) Student graduation data which has 15 attributes and 290 records including name, study program, GPA, year of entry, year of exit, graduation, year, month, day, total in months, academic year, and predicate graduation. 2) Pulled data Feeder which contains 4 (fours) attributes and 407 records (only counting the retrieved data) spread over several files and merging or integrating the data. The data used for the undergraduate level is student data who entered in 2012, 2013, and 2014 which amounted to 280 records. At the D3 level used, data from students who entered in 2012, 2013, 2014, 2015, 2016, and 2017 amounted to 127 records. The result is that D3 level dropout students have an accuracy rate of 89.47% with 100% precision for drop out students. The S1 level obtained an accuracy of 96.43% with a student dropout precision of 88.46%.

Corresponding Author:

Nariza Wanti Wulan Sari,
Faculty of Computer/Study Program of Information System,
Universitas Mulia,
Letjen Z.A. Maulani Street Number 9, Balikppapan City, East Kalimantan, Indonesia.
Email: nariza.ws@universitasmulia.ac.id

1. INTRODUCTION

The number of graduates and drop out is one measure of the quality of a university. According to Nasrullah (2018), students who have the potential to drop out can be a factor that reduces the quality of education and college accreditation (Ameri et al., 2016). Some of the reasons students drop out include being inactive for 3 (three) academic years, low academic achievement or GPA, and because they have passed the maximum study period.

The increasing number of non-active students at Mulia University PSDKU Samarinda makes the risk of student drop out increasing. This problem is the main topic in this research. The solution proposed for this problem is to predict students who have the potential to drop out (del Bonifro et al., 2020), so that early on they can find out the characteristics of students who drop out (Lee et al., 2021).

Data mining is a branch of science that can dig up important information or patterns from a large amount (Azahari et al., 2020). One application in data mining is classification. Classification is the process of finding a model that describes and distinguishes data classes or concepts that aim to be used to predict the class of objects whose class label is unknown (Azhari et al., 2021). One method of classification is the Naïve Bayes algorithm (Nuraeni et al., 2021). The Naïve Bayes algorithm uses a probability method based on the Bayes theorem which predicts future opportunities based on previous experience(Chen et al., 2020). Syarli and Muin (2016) predict graduation of new college students using the Naïve Bayes method to obtain an accuracy rate of 94%. Anggreani et al. (2018)uses the Naïve Bayes method to predict the length of study for students, obtaining an accuracy rate of 76%. Hartatik uses Naïve Bayes to predict student graduation using 2 (two) models, achieving an accuracy rate of 85% (Hartatik, 2021).

So, in this study, predictions of students dropping out at Mulia University PSDKU Samarinda were made using the Naïve Bayes algorithm. The purpose of this study is to provide predictive results for students dropping out at Mulia University PSDKU Samarinda with the Naïve Bayes algorithm, so that from an early age it can know the characteristics of students dropping out and can reduce the number of students dropping out.

2. RESEARCH METHOD

Based on the source, the data in this study is internal data and based on the method of obtaining it, it is secondary data obtained from the Academic Section and the data Feeder pull. The data collection carried out in this study used a document study. The data used in this study are students who graduated and dropped out of the 2012-2017 class. The attributes used are graduation or dropout status (Status), GPA until the end of study (GPA)(Sari et al., 2019), number of credits taken (SKS), study period (Study_Term), number of leave ever taken (Leave), and gender (Gender). All attributes are classified using the Naïve Bayes algorithm with the Rapidminer software (RapidMiner, 2019) (Dewi et al., 2021). To calculate the accuracy value, the data is divided into testing data and training data and various compositions are tried(Musu et al., 2021). The stages that have been carried out in this research are as follows Figure 1.

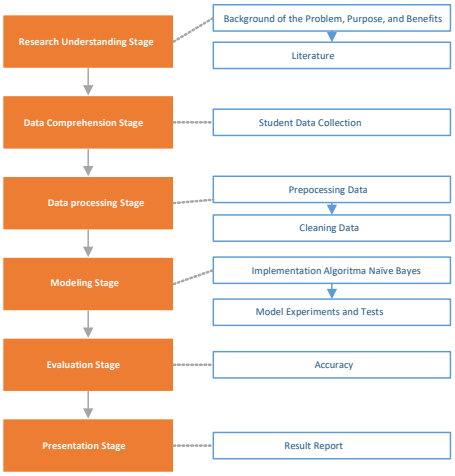

Figure 1. Research Stages

Figure 1 shows the stages of research that have been carried out starting from the stage of understanding research, stages of understanding data, processing data, modeling stages, evaluation stages, to the stage of presenting or reporting research results.

3. RESULTS AND DISCUSSIONS

In this study, 2 (two) data sets were used. First, student graduation data which has 15 attributes and 290 records including name, study program, GPA, year of entry, year of exit, graduation, year, month, day, total in months, academic year, and graduation predicate. The second is the Data Feeder pull which contains 4 attributes and 407 records (only counting the retrieved data) which are spread over several files and are combined or integrated with the data. The data used for the undergraduate level is student data who entered in 2012, 2013, and 2014 which amounted to 280 records. At the Diploma level used is student data who entered in 2012, 2013, 2014, 2015, 2016, and 2017 which amounted to 127 records. Furthermore, data processing is carried out with the following flow:

3.1 Data Cleaning

At this stage, data cleaning is carried out which is also the beginning of the Knowledge Discovery from Data (KDD) process. The data cleaning process is carried out by checking for inconsistent or irrelevant data, repetitive data, and data that has missing values. Data is said to be missing value if one of the attribute values of the record is missing or does not exist, then the record is deleted because it is considered a missing value. The data used in this study is consistent, non-repetitive, and has no missing value. At this stage the data is clean and proceed to the next stage, namely data integration.

3.2 Data Integration

In the data integration stage, data from 2 data sets will be combined into 1 data set. As in the following Figure 2.

NIM	Nama	IPS 7	SKS 7	Semester 8	IPK 8	SKS 8	Semester 9	IPS 9	SKS 9	semester 10	IPS 10	Angkatan	Status	IPK	Lama	Cuti	Non-Aktif	SKS	Jenis Kelamin
2,01E+11	UMAR HADI											2016	DO	0,14	6		0	34	Laki-Laki
2,01E+11	YULIANA REZA SUSANTI		4	Aktif	3	4						2014	lulus	3,28	47		0	120	Perempuan
2,01E+11	DEVID VARIZKY											2014	DO	0	6		0	24	Laki-Laki
2,01E+11	ARWANSYAH											2014	DO	0,27	6		0	34	Laki-Laki
2,01E+11	ASHAR ABDUL AZIZ	2,72	21	Aktif	3,56	9	Aktif	3,81	18	Aktif	3,1	2015	lulus	3,1	59	2		127	Laki-Laki
2,01E+11	ACHMAD WAHYUT											2015	lulus	3,46	55	0		219	Laki-Laki
2,01E+11	AMAT BASSARI											2015	lulus	3,37	35	0		113	Laki-Laki
2,01E+11	[illegible]			Cuti			Aktif		6	Aktif	3,54	2015	lulus	3,56	59	2		113	Laki-Laki
2,01E+11	IFFRARNI SUHARISHNA			Aktif	3,8	6						2015	lulus	3,72	47	0		232	Laki-Laki
2,01E+11	MUHAMAD ARFI ASYQORI											2015	DO	1,01	20	0		20	Laki-Laki
2,01E+11	KUDZI MUAIEIF			Aktif	3,1	6						2015	lulus	3,22	47	0		113	Laki-Laki
2,01E+11	TITIK SURYANI			Cuti			Non-Aktif			Non-Aktif	3,29	2015	DO	3,29	60	5		80	Perempuan
2,01E+11	VITYA AYANDA DEWANTO			Cuti			Non Aktif			Non-Aktif	3,12	2015	DO	3,12	60	5		80	Perempuan
2,01E+11	AL SADIKIN											2015	DO	0	6	0		20	Laki-Laki
2,01E+11	WUTIYA											2015	lulus	3,36	35	0		113	Perempuan
2,01E+11	MUHAMAD RISMAN			Cuti								2015	DO	0	6	0		20	Laki-Laki
2,01E+11	[illegible]						Aktif		6	Aktif	3,42	2015	lulus	3,42	59	2		121	Laki-Laki
2,01E+11	AHMAD YULIANSYAH			Non Aktif	0		Non Aktif			Non-Aktif	2,48	2016	DO	2,48	60	6		71	Laki-Laki
2,01E+11	DIAN WILLY S, PANJAITAN		19	Aktif	3,15	17	Non-Aktif			Non-Aktif	2,86	2016	DO	2,86	60	3		120	Laki-Laki
2,01E+11	HERY YARSANTO			Aktif	0	6	Non-Aktif			Non-Aktif	3,13	2016	DO	3,13	60	3		114	Laki-Laki
2,01E+11	RUSWAN FAHRUROZI											2016	DO	0,58	6	0		20	Laki-Laki
2,01E+11	ACHMAD NURHDI		4	Aktif	3,75	4						2016	lulus	3,55	47	0		120	Laki-Laki
2,01E+11	ADI ABADAN											2016	DO	0,55	6	0		20	Laki-Laki

Figure 2. Data Integration Results

3.3 Data Selection

From the new data set, not all attributes and records are used, only the appropriate attributes and records are taken for analysis, so that the data to be analyzed has 6 attributes and for the S1 level the number of records used is 280. The data has been selected so it can go to the next stage, namely data transformation.

Table 1. Selection Result Data

No	Status	GPA	SKS	Studi Term	Leave	Gender
1	DO	0,66	116	60	1	P
2	DO	1,69	146	78	3	L

No	Status	GPA	SKS	Studi Term	Leave	Gender
3	DO	0,00	20	6	0	L
4	DO	3,35	148	60	0	L
5	Lulus	3,00	148	58	0	P
6	Lulus	3,02	148	58	0	L
7	Lulus	3,23	148	58	0	L
8	DO	3,16	20	6	0	P
⋮	⋮	⋮	⋮	⋮	⋮	⋮
277	DO	2.45	21	6	0	L
278	DO	0,00	21	0	0	P
279	DO	0,00	21	0	0	L
280	DO	2,68	161	84	4	L

3.4 Data Transformation

The data transformation stage is that the data is changed or combined into an appropriate format for processing in data mining. The data set used is already in a formal excel form so there is no need for data transformation. Furthermore, the mining process can be carried out.

3.5 Implementation of Naïve Bayes

The data were analyzed using the Naïve Bayes method, the following is an image of the model application in rapid miner for the drop out student data set.

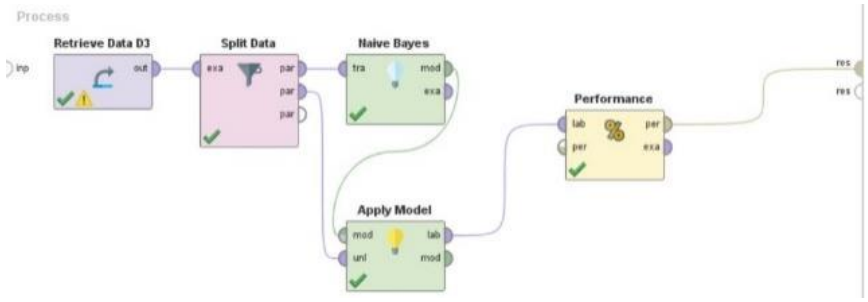

Figure 3. Naïve Bayes model in Rapidminer

In the data split, the composition of the testing and training data is adjusted, the composition of 70:30, 80:20, and 90:10 is used and gives different accuracy results.

3.6 Evaluation

Accuracy results using the composition of 70:30, 80:20, and 90:10 are shown in the following Table 2.

Table 2. Diploma Data Set Accuracy Results

Composition	Accuracy
70:30	89,47%
80:20	84%
90:10	84,62%

In the Diploma data set with 127 records the composition of 70:30 for training and testing data has a higher accuracy rate than the composition of 80:20 and 90:10, which is 89.47%. So, for the prediction of students dropping out of Diploma level, the composition of 70:30 is used.

Table 3. Undergraduate Data Set Accuracy Results

Composition	Accuracy
70:30	94,05%
80:20	92,86%
90:10	96,43%

In the undergraduate data set with 290 records, the composition of 90:10 for training and testing data has a higher accuracy rate than the composition of 70:30 and 80:20, which is 96.43%. So for the prediction of dropout students at undergraduate level, the composition of 90:10 is used. Predictions were also made for 2018 undergraduate students, using previous training data and using 2018 class data as testing, it was found that from 122 students, 2 students were confirmed to have dropped out and 34 of them were predicted to drop out until the term for the undergraduate level ended.

4. CONCLUSION

Based on the research that has been done, it is found that using the Naïve Bayes method can predict the dropout student data set and has a good level of accuracy. The result is that Diploma dropout students have an accuracy rate of 89.47% with 100% precision for dropout students. The Undergraduated level obtained an accuracy of 96.43% with a student dropout precision of 88.46%. From the research, it is also known that the selection of the number of records and the composition of training/testing affect the level of accuracy obtained. Some suggestions put forward by the authors include the Head of the Mulia University PSDKU Samarinda Office and the Head of the Study Program, to give more attention to 34 students who are predicted to drop out so as to reduce the number of dropout students. Furthermore, research can add other student characteristics attributes so that they can provide even richer information such as age (when entering college), classification of the city of origin of high school, father's occupation, class (regular/executive), and number of dependents in the family (number of siblings) and others.

ACKNOWLEDGEMENTS

Thank you very much for the support provided by LPPM Mulia University so that this research was completed smoothly.

REFERENCES

Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. *International Conference on Information and Knowledge Management, Proceedings*, *24-28-October-2016*. https://doi.org/10.1145/2983323.2983351

Anggreani, D., Herman, & Astuti, W. (2018). Kinerja Metode Naïve Bayes dalam Prediksi Lama Studi Mahasiswa Fakultas Ilmu Komputer. *Seminar Nasional Ilmu Komputer Dan Teknologi Informasi*, *3*(2).

Azahari, A., Yulindawati, Y., Rosita, D., & Mallala, S. (2020). Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3). https://doi.org/10.25126/jtiik.2020732093

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes.

JURNAL MEDIA INFORMATIKA BUDIDARMA, *5*(2). https://doi.org/10.30865/mib.v5i2.2937

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*. https://doi.org/10.1016/j.knosys.2019.105361

del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12163 LNAI*. https://doi.org/10.1007/978-3-030-52237-7_11

Dewi, P. S., Sastradipraja, C. K., & Gustian, D. (2021). Sistem Pendukung Keputusan Kenaikan Jabatan Menggunakan Metode Algoritma Naïve Bayes Classifier. *Jurnal Teknologi Dan Informasi*, *11*(1). https://doi.org/10.34010/jati.v11i1.3593

Hartatik, H. (2021). Optimasi Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes. *Indonesian Journal of Applied Informatics*, *5*(1). https://doi.org/10.20961/ijai.v5i1.44379

Lee, J. H., Kim, M., Kim, D., & Gil, J. M. (2021). Evaluation of Predictive Models for Early Identification of Dropout Students. *Journal of Information Processing Systems*, *17*(3). https://doi.org/10.3745/JIPS.04.0218

Musu, W., Ibrahim, A., & Heriadi. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4 . 5. *Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *X*(1).

Nasrullah, A. H. (2018). Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out. *ILKOM Jurnal Ilmiah*, *10*(2). https://doi.org/10.33096/ilkom.v10i2.300.244-250

Nuraeni, F., Agustin, Y. H., Rahayu, S., Kurniadi, D., Septiana, Y., & Lestari, S. M. (2021). Student Study Timeline Prediction Model Using Naïve Bayes Based Forward Selection Feature. *8th International Conference on ICT for Smart Society: Digital Twin for Smart Society, ICISS 2021 - Proceeding*. https://doi.org/10.1109/ICISS53185.2021.9532502

RapidMiner. (2019). *Naive Bayes - RapidMiner Documentation*. RapidMiner.

Sari, N. W. W., Suyitno, S., & Mirwansyah, D. (2019). Faktor-Faktor Yang Mempengaruhi Ipk Lulusan Stmik Sentra Pendidikan Bisnis. *Prosiding Seminar Nasional Matematika Dan Statistika*, 199–205.

Syarli, & Muin, A. A. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan. *Jurnal Ilmiah Ilmu Komputer*, *2*(1).