www.bircu-journal.com/index.php/birci
Budapest International Research and Critics Institute-Journal (BIRCI-Journal)
Humanities and Social Sciences
ISSN 2615-3076 Online)
ISSN 2615-1715 (Print)
Budapest Institute

Prediction of Student Graduation Using Naïve Bayes

Robi Sepriansyah[1], Susan Dian Purnamasari[2]
[1,2]*Faculty of Computer Science, Universitas Bina Darma, Indonesia*
robi.sepriansyah83@gmail.com, susandian@binadarma.ac.id

Abstract

The quality of universities, especially study programs in Indonesia is measured based on an accreditation assessment from the National Accreditation Board for Higher Education (BAN-PT). The quality assessment is measured based on 7 main standards, one of which is students and graduation. Every university must have academic data and biodata of each student based on the initial registration until graduation. Students who are accepted or who enter college are increasing every year, but not all students are able to graduate on time.algorithm Naïve Bayes used study aims to predict student graduation through student academic performance data in semester one to semester four, attributes Nim, Credit and GPA using the Discovery In Database (KDD) This Knowledge model data Testing on the Rapid Miner application.From the results of the tests that have been carried out, it can be concluded that the accuracy value of the prediction results is 95.33%, the results are quite accurate for the data used by testing the testing as many as 120, namely passing in semester 8 as many as 78, passing in semester 9 as many as 24, while 3 who graduated in semester 10, and students who graduated in semester 12 were 15.

Keywords
prediction; classification; naive bayes; rapid miner.

I. Introduction

Student graduation is one of the fields of education that is included in the Internal Quality Assurance Standard (SPMI) at a university. One of the standards in achieving a graduation that has been set in college is to produce timely graduation by taking a maximum of 8 semesters and a total study load that has been covered by a minimum of 144 credits. At Bina Dharma University, the accreditation process is related to the accuracy of the student's graduation, because it is very important in the accreditation assessment process.

Therefore, in order to reduce the number of students who graduate on time, a system that can be used to predict student graduation is needed by using data or information in order to determine student graduation, so that it can be predicted from the start so that parties involved in academics can carry out a policy in order to Minimize the number of graduating students who graduate on time. Development is a systematic and continuous effort made to realize something that is aspired. Development is a change towards improvement. Changes towards improvement require the mobilization of all human resources and reason to realize what is aspired. In addition, development is also very dependent on the availability of natural resource wealth. The availability of natural resources is one of the keys to economic growth in an area. (Shah, M. et al. 2020)

One of the efforts in utilizing student data is to manage data using data mining which is the process of making data processing techniques, data mining techniques, so that certain patterns can be produced which become information based on these methods and

DOI: https://doi.org/10.33258/birci.v5i3.6447

algorithms. The classification method naive Bayes order to produce information in the form of predictions of student graduation. Student data can be mined or commonly known as Knowledge Recovery in Database (KDD), which is an external process of important information from a large database. Therefore, based on available data which can be used to predict the future using a statistical approach, based on complex data can apply the ability to extract the data to be processed into more important information.

In every college, of course, has a database to store all student data in every academic, student data continues to increase every year and accumulates like neglected data because the data is rarely reprocessed, while from student data it can be processed to produce information, namely about the level of graduation of students so as to minimize delays in graduation.

Based on the above problems, predicting student graduation by utilizing data mining algorithm naive Bayes classification method using rapid miner tools where the purpose is to be able to help find information in predicting graduation of students from the Faculty of Engineering, Bina Darma University so that it can provide information for study programs in order to predict student status. and can be used to determine steps and policies for students in targeting their graduation.

II. Review Of Literature

Bina Darma University is a private university, this university is located on Jln A, Yani No 3 Palembang. Academic education at Bina Darma University consists of diploma programs, undergraduate programs, and postgraduate programs, with 7 (seven) faculties including the faculty of computer science, faculty of economics and business, faculty of education, faculty of communication science, vocational faculty, faculty of engineering, and the faculty of psychology. This research was conducted at the engineering faculty by using student alumni data as *training* data and batch student data as *testing* needed to predict the length of graduation for students who will be tested in the *rapid miner* by utilizing data mining using the naive Bayes algorithm.

1) Data

   Data is a collection of information or values that have been obtained based on the results of observations (observations) on an object in the form of numbers, symbols or properties that are already known or considered. Known means that based on the data, it is a fact (evidence) to be able to provide an overview of a situation or problem.

2) *Knowledge Discovery In Database* (KDD)

   Is a structured analysis process to identify valid, useful, and understandable patterns based on complex data sets. Based on these data can be a source of knowledge for an organization, so it can be a useful source of knowledge and used to make decisions.

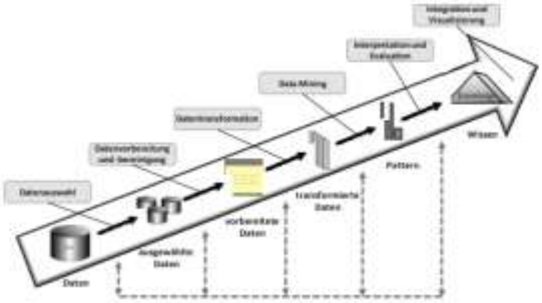


Figure 1. *Knowledge Discovery In Database (KDD)*

a. Data Selection

Selection (selection) is based on a set of operational data that needs to be done before carrying out the information mining stage in the KDD begins. At this stage, select relevant data or attributes so that the data generated from the selection process will be used for the data mining process.

b. Pre-processing/Cleaning

Before the data mining process is carried out, it is necessary to do *data cleaning* which is the focus of *Knowledge Discovery in Database*. By setting aside duplication of data, checking inconsistent/incompatible data, and correcting errors in the data to eliminate *noise,* if there is data that is empty from one of the attributes, the data will be deleted or eliminated.

c. *Transformation*

Data that has been obtained is suitable for data mining processes, which can be used in data processing in *rapid mining applications.*

d. Data mining (data mining)

This stage is the most important stage because it is a process to look for patterns to get interesting information in the data by using methods, techniques, or algorithms that are in accordance with the objectives and process of extracting data.

e. Interpretation/Evaluation

Understanding whether the pattern or information generated contradicts the facts. With the information generated based on the *data mining* in a form that is easy to understand by those who need it.

3) Data Mining

is a series of processes used to extract information that has not been known manually from a database. Data mining is defined as the process of determining the relationship between meaningful new patterns and trends through filtering using pattern recognition techniques such as statistical methods that can be used. to support future decisions.

4) Prediction

Prediction is something that will happen in the future, obtained by the scientific method, or purely subjective. Estimates that can be used for class classification are based on various given attributes, so that they are not only used for time series forecasting, prediction is also a process of evaluating future forecasts based on previous work.

- Classification

  Assuming that the presence or absence of certain characteristics of one class is not related to other classes. Classification is processed by finding a model that looks like describing and characterizing a concept or data class for a particular purpose by using an unknown object class and label model.

- *Naive Bayes Naive Bayes*

  is one of the methods used for classifying data based on Bayes' theorem. By assuming that the data are independent of each other. classification using probability and statistical methods proposed by British scientist Thomas Bayes which can predict future opportunities based on previous experience.

  Theorem *in general, the Naive Bayes* as follows

- . Calculating each class (on time and late)

  Prior Probability:

  $$\boxed{\mathbf{P\ (Ci\ )}}$$

- Calculating the *Posterior probability*

Posterior Probability:

P (XICi)

- Maximizing the Value with**:**

P (XICi) P (Ci)

Description:

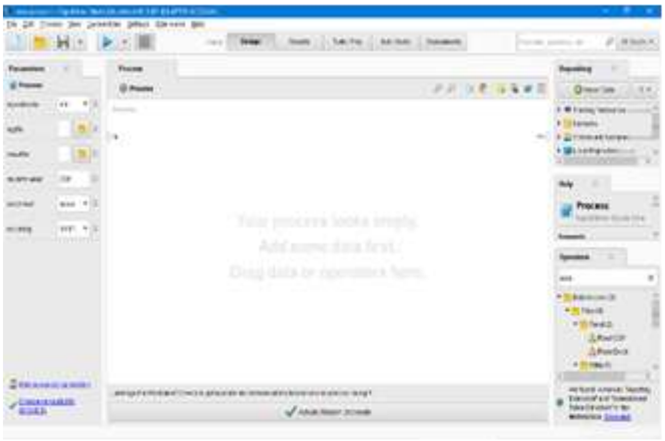
X: Data with *class* unknown

H: Hypothesis of data is a specific class

P (Ci): Probability of *Class*

P (X) : Probability of X

5) Rapid Miner
This study uses Rapid miner which are software tools developed by Markus Hofman at the Institute Blackhardstown and Ralf Klinkenberg technology at rapidi.com which has a GUI (graphical user interface) display. In the use of rapid miner which makes it easier for users to operate the software. Because this software is open source and created using a Java program under a public license GNU and licenses RapidMiner can run on any operating system. This application does not require special programming knowledge to use Rapid Miner because in the rapid miner application there are features that can be used for data mining. in Rapid Miner There are many models, offered, including Naïve Bayes model TreeInduction, Neutral Network modeling, and many more. Rapid Miner also provides many methods, namely classification, grouping, association, etc. In this study, the author uses data mining using the Naïve Bayes in the classification process with training data (Training Data) and testing data (test data) for testing the RapidMiner.



Your process looks empty.
Add some data first.
Drag data or operators here.

Figure 2. *Rapid Miner Interface Display*

III. Research Method

The method used in this study is a qualitative method. And processed using data mining algorithm Naïve Bayes application Rapid Miner to know the prediction results and the level of accuracy of student graduation. The method of data collection was carried out using primary data because the data used were data obtained directly from Bina Darma University where researchers made observations in addition to using literature and literature studies related to the research theme.

3.1 Observation

Stages of collecting information related to research by observing directly an object for a certain time. This is done in the Bina Darma University data processing room to get the data needed for research. The data that has been collected is then stored in excel form so that it can be processed using data mining.

3.2 Literature study

This stage is carried out by taking information based on books, journals and the results of previous research related to the problem, so that it can be used as a source of information related to research.

3.3 Literature

Study Literature study is used to add sources of information related to theories that are in accordance with research problems, this method is used to explore, compare problems based on the literature obtained, and make them the subject of research, this literature study provides a broad orientation and avoids duplication of research which has been done before.

a. Research Phase

In this research, several stages were carried out, as shown below.

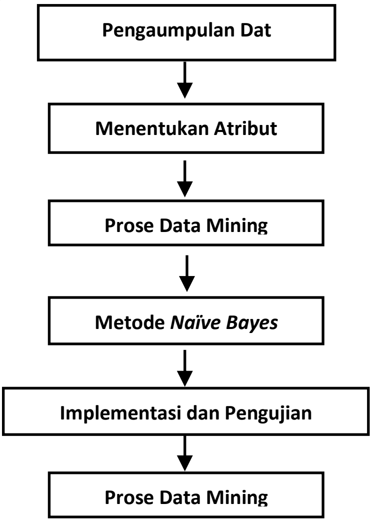


Figure 3. *Stages of*

b. Data Mining Process

Data mining is a technique that uses statistics, mathematics, artificial intelligence, and *machine learning* in order to identify useful information on large data sources. This study aims to apply data mining classification to predict student graduation by using the *Naïve Bayes* so as to produce predictions of where the length of the semester is taken.

c. Algorithm Naive Bayes

Naive Bayes is one of the most effective and efficient inductive learning algorithms in *machine learning* and *data mining*. Naïve Bayes classification uses probability and statistical methods proposed by the British scientist, Thomas Bayes. Can be used to predict future opportunities based on previous experience.

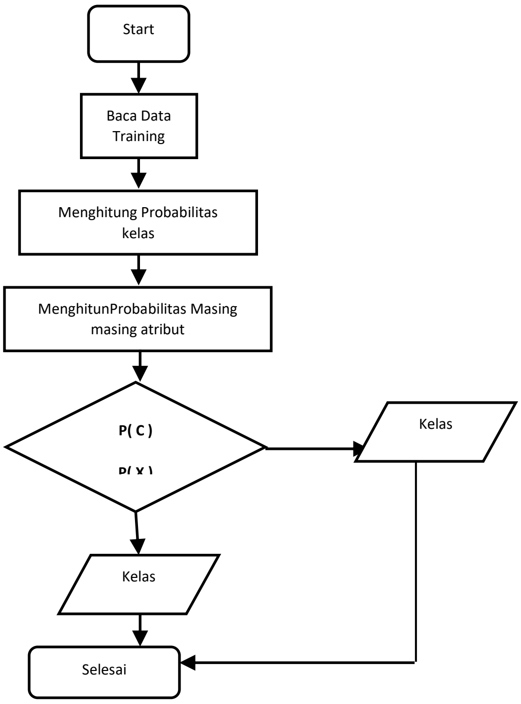

Start
Baca Data Training
Menghitung Probabilitas kelas
MenghitunProbabilitas Masing masing atribut
P(C)
P(X)
Kelas
Kelas
Selesai

Figure 4. *Flow Naive Bayes*

d. Method Testing

In testing the level of accuracy of this study using the *Rapid Miner* because it includes capabilities with the *Naïve Bayes* and can find out the content of patterns from existing data. The following are the stages of testing using the *Rapid Miner*

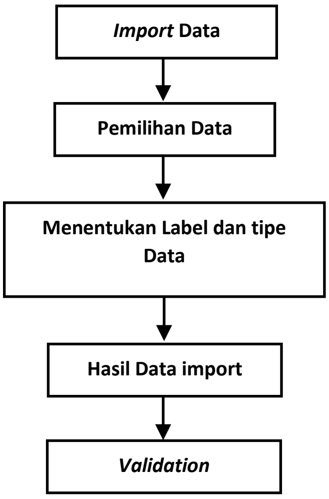


Figure 5. *Stages of using Rapid Miner*

Stage - Stages of implementing the *Training* and *Testing* using *rapid miner application.*

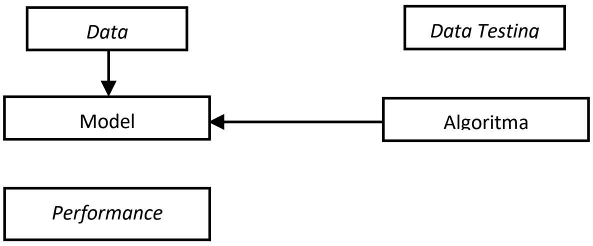


Figure 6. *Stages of using Rapid Miner*

e. Calculation of Naïve Bayes Classification The

algorithm *Bayes* using attributes that can determine student graduation.

Calculate each class (On time and late.)

Prior Probability

P (Ci)

On time I class = 80/151 = 0.26

Late I class = 71/151 = 0.47

Calculates the *Posterior probability*

Posterior Probability

P (XICi)

If X : Sks 60, and Temporary IP 2.95?

Credits 60 I On Time = 0/80 Credits

60 I Late = 5/71

Gp <3 I On Time = 6/80

GPA <3 I Late = 10/71

On Time

: 80/151 * 0/80 * 6/80

: 0.26 * 0 * 0.075

: 0

Late

: 71/151 * 5/71 * 10/71

: 0.47 * 0.070 * 0.140

: 0.004

Maximizing Values on

$$\boxed{\textbf{P (XICi) P (Ci)}}$$

Time

P (XIOn Time : Class)* P (On time : Class)

: 0 * 80/151

: 0

Late

P (XILate : Class)* P (Late : Class)

: 0.0004 * 71/151

: 0.0018

So, based on the calculation results Late > grade On time, then X goes to class **late.**

If X: 74 credits, and temporary IP 3.09?

Credits 74 I On Time = 2/80 Credits

74 I Late = 0/71

GPA >3 I On Time = 73/80

GPA >3 I Late = 43/71

On Time

: 80/151 * 2/80 * 73/80

: 0.26 * 0.025 * 0.91

: 0.0059

Late

: 71/151 * 0/71 * 43/71

: 0.47 * 0 * 0.605

: 0

Maximize Grades with

$$\boxed{\textbf{P (XICi) P (Ci)}}$$

Punctuality

P (XIOn Time : Class)* P (On Time : Class)

: 0.0059 * 80/151

:

Late

P (XILate : Class)* P (Late : Class)

: 0 * 71/151

: 0

So, based on the results of the calculation of the Late value > On time value, then X**.**

IV. Result and Discussion

At this stage looking for patterns or selecting data from a set of operational data that needs to be done before the data mining process in the KKD is carried out, the data generated after the selection process will be used in the *data mining* using classification techniques and *naive Bayes* to be able to determine the results. student graduation prediction.

4.1 Data Mining

Naive Bayes is a classification algorithm using probability and statistical methods that are used to predict future opportunities based on previous experiences and the most effective and efficient learning for *machine learning* and data mining.

In the data mining process that has been discussed previously there are several steps, namely data cleaning is a process to eliminate noise and inconsistent data, data integration is the process of combining data so that data is avoided from duplication of data, data selection is the process of selecting relevant data to use. Data transformation, i.e., data that has been obtained is changed in the form of data that is suitable to be carried out in the data mining process is an essential process where intelligent methods are used to extract data patterns to identify interesting patterns to gain knowledge.

4.2 Data mining process using Rapid Miner

This research uses a rapid miner application to test student graduation predictions by utilizing training data *(training* data) from alumni in 2014 which amounted to 151 data containing 10 attributes including Nim, semester credits 1 to 4, and year of graduation. as labels.data *test* data) from the 2019 batch amounted to 120 with Nim, Credit and GPA attributes. The data was tested using *naive Bayes* on the *rapid miner* 9.10

rapidminer
Studio
Version 9.10
Loading Time Series Extension
Educational Edition registered to Robi Sepriansyah
Copyright (C) 2001-2021 RapidMiner GmbH

Figure 7. *Application Rapid Miner*

As a view for the worksheet menu has several views including *operator* is *view* the most important view. All work steps in *Rapid Miner* displayed in the operator *view*, in the *operator view* there are several groups including the first *Process Control* logic *looping* that regulates the flow of the data analysis process, the second *Utility* which functions as a help operator such as macros, loggin , subprocesses and others, Third *Repository Access* which functions to read or write access to *the repository*, fourth, namely *import* which has a function to read object data from files, databases, fifth, *Export* functions as opposed to *import* where this operator is used to writes data into a certain format, then the *Data Transformation* functions to transform data and metadata. Then the *modeling* in which there are various kinds of data mining methods and techniques to be able to manage data, and the last operator, namely *Evaluation* functions to evaluate the quality of the output produced.

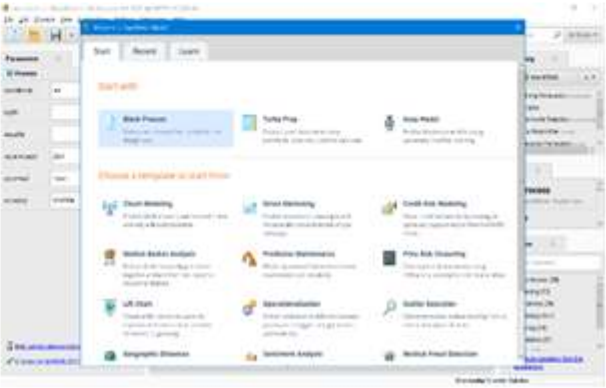


Figure 8. *Welcome Perspective*

4.3 The Model formation

The model is carried out by utilizing the *Rapid Miner* which begins by entering the *Training* data and *Testing* that has been provided previously by pressing on the *operator* section then *reading excel*, because the training and testing data are collected in the form of an excel file. As shown in Figure 9.

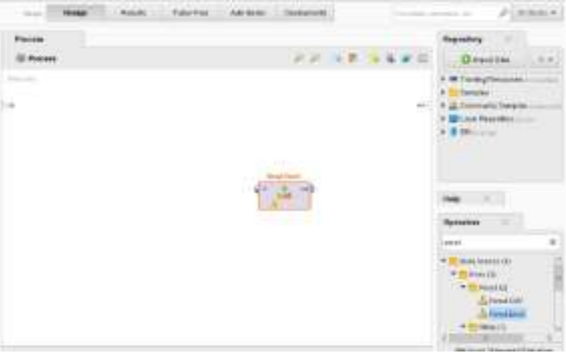
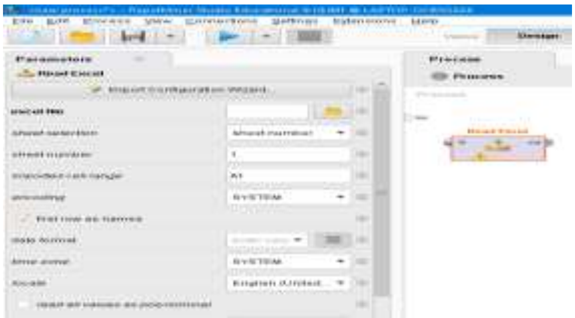


Figure 9. *Adding Read Excel*

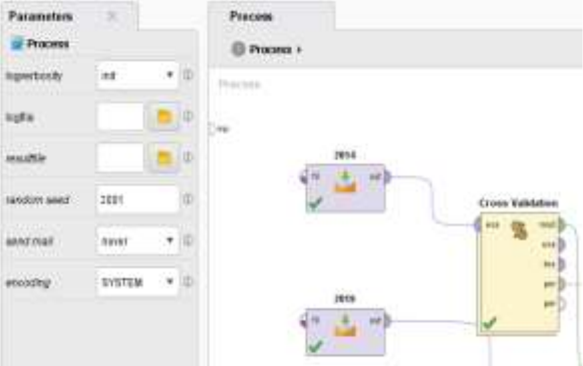
After the *Read Excel* is added, then enter the *training* data and *testing* that will be used. By clicking *import configuration wizard* then selecting the excel file to be used as shown in Figure 10. There are several steps after selecting the data, annotations and attributes on the data used for testing. Because the data used is in accordance with the needs



excel file
sheet selection: sheet number
sheet number: 1
imported cell range: A1
encoding: SYSTEM
☑ first row as names
date format
time zone: SYSTEM
locale: English (United...
□ read all values as polynominal

Figure 10. *Selecting the Training and Testing*

The next step after adding the *Read excel* is adding *cross validation*. As shown in Figure 11.



logverbosity: init
logfile
resultfile
random seed: 2001
send mail: never
encoding: SYSTEM

Figure 11. *Adding cross validation*

In rapid miner, the *cross-validation* operator is a nested operator that has two subprocesses, namely the *training* which is used to train the model and the *testing* to test the model as well as measure the performance of the model, shown as Figure 12.

4.4 Test and Evaluation

Process This test process is carried out by entering *Training Data (Training* Data) and *Training* Data (Test Data) by entering data into the *read excel* because the data to be tested is in excel format, then applied with the *apply model* and connecting between operators as shown in Figure 13, which serves to test the model used.

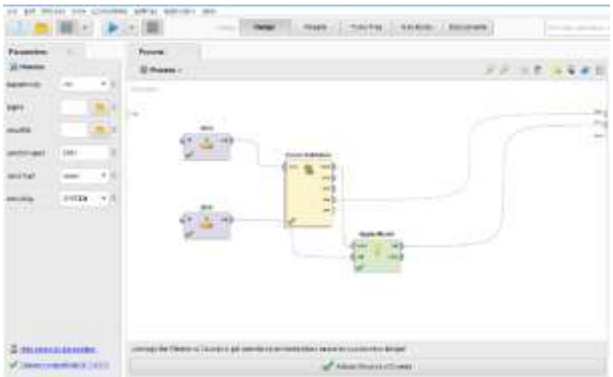


Figure 12. *Testing on the Apply Model*

The prediction results that have been made on the Rapidminer application to predict student graduation by utilizing *training* and *testing data* algorithm model *Naïve Bayes* can be seen in Figure 14. The data used is test data with 120 examples and 8 attributes Among them are Nim, Ip semesters 1 to 4 and GPA semesters 1 to 4 to determine student prediction results with graduation results predicting that those who pass in semester 8 are 78, graduated in semester 9 are 24, while those who pass in semester 10 are 3, and students who graduate in the 12th semester as many as 15 people while the results for testing the two data with perfect accuracy results of 95.33%.

Figure 13. *Prediction Results for Semester Graduation*

V. Conclusion

Based on the results of research that has been carried out and has been described in the application of data mining to predict student graduation at the Faculty of Engineering, Bina Darma University, the authors can draw the following conclusions:

***Figure 14**. Accuracy results from data testing*

1. Application data mining using the *Naive Bayes* that has been carried out can produce information about predicting the graduation of engineering faculty students who graduated in semester 8 as many as 78, graduated in semester 9 as many as 24, while those who graduated in semester 10 were 3, and students who graduated in semester 10 12 as many as 15 people with a prediction accuracy rate of 95.33%. It is accurate enough to be used to predict student graduation.
2. Produce information about student graduation data at the Faculty of Engineering, Bina Darma University. Resulting from the application of data mining in rapid mining applications that have been carried out.
3. Calculations that have been done theoretically and applications produce information on the value of the predetermined classification data.

References

Alim, S. (2021).” Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes Orange Data Mining Implementation For Student Graduation Classification Using K-Nearest Neighbor, Decision Tree And Naive Bayes Models" 6, 12.

Anugrah Putra, D., Kamayani, M., (2020). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naive Bayes di Program Studi Teknik Informatika UHAMKA. Prosid Sem Nas Teknoka 5, 34–40. https://doi.org/10.22236/teknoka.v5i.331

Anwar, F.F., Jaya, A.I., Abu, M., (2022). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree dengan Penerapan Algoritma C4.5. JIMT 19, 19–28. https://doi.org/10.22487/2540766X.2022.v19.i1.15880

Banjarsari, M.A., Budiman, H.I., Farmadi, A., (2015). Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4 02, 15.

Fadillah, A.P., (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi

Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). JuTISI 1. https://doi.org/10.28932/jutisi.v1i3.406

Hananto, V.R., n.d. Analisis Penentuan Metode Data Mining Untuk Prediksi Kelulusan Mahasiswa Sebagai Penunjang Angka Efisiensi Edukasi 11.

Hendra, H., Azis, M.A., Suhardjono, S., (2020). Analisis Prediksi Kelulusan Mahasiswa Menggunakan Decission Tree Berbasis Particle Swarm Optimization. SISFOKOM 9, 102–107. https://doi.org/10.32736/sisfokom.v9i1.756

Larasati, I.D., Supianto, A.A., Furqon, M.T., n.d. Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Metode Modified K-Nearest Neighbor (MK-NN) 6.

Malelak, K.H.L., Ardiada, I.M.D., Feoh, G., (2021). Implementasi Klasifikasi Naive Bayes Dalam Memprediksi Lama Studi Mahasiswa (Studi Kasus : Universitas Dhyana Pura). Sintech Journal 4, 202–209. https://doi.org/10.31598/sintechjournal.v4i2.964.

Maulana, D., Nurjanah, E.L., (2019). Analisa Tingkat Kepuasan Pelanggan Terhadap Penjualan Beauty Produk Pada Online Shop Dengan Menggunakan Metode Naive Bayes 10, 8.

Murtopo, A.A., (2016). Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes. CSRID Journal 7, 145. https://doi.org/10.22303/csrid.7.3.2015.145-154

Nasution, N., Djahara, K., Zamsuri, A., n.d. Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes (Studi Kasus: Fasilkom Unilak) 11 [11] A. Moenir and F. Yuliyanto, "Perancangan Sistem Informasi Penggajian Berbasis Web dengan Metode Waterfall pada PT. Sinar Metrindo Perkasa (Simetri)," J. Inform. Univ. Pamulang, vol. 2, no. 3, pp. 127–137, 2017.

Pambudi, R.D., Supianto, A.A., Setiawan, N.Y., n.d. Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya 7.

Prasetyo, V.R., Lazuardi, H., Mulyono, A.A., Lauw, C., (2021). Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression. TEKNOSI 7, 8–17. https://doi.org/10.25077/TEKNOSI.v7i1.2021.8-17

Rohmawan, E.P., n.d. Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural Network 10

Romadhona, A., Himawan, H., (2017). Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma Decision Tree 13, 15.

Rudy Hendrawan, I.N., Budhi Saputra, I.M.A., Cahya Dewi, G.A.P., Adi Pranata, I.G.S., Wedasari, N.L.N., (2022). Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes. eksplora 11, 50–56. https://doi.org/10.30864/eksplora.v11i1.606

Sabilla, W.I., Putri, T.E., n.d. Prediksi Ketepatan Waktu Lulus Mahasiswa dengan k-Nearest Neighbor dan Naïve Bayes Classifier (Studi Kasus Prodi D3 Sistem Informasi Universitas Airlangga) 8.

Sardi, H.Y., Budayawan, K., (2020). Klasifikasi Tingkat Kelulusan Mahasiswa Elektronika Menggunakan Algoritma Naïve Bayes Classifier 8, 5.

Shah, M. et al. (2020). The Development Impact of PT. Medco E & P Malaka on Economic Aspects in East Aceh Regency. Budapest International Research and Critics Institute-Journal (BIRCI-Journal). P. 276-286.