Data Mining Classification to Predict Student Graduation Using the Naive Bayes Method

Jefri[a1], Zaehol Fatah[a2]
[a] Information Systems Study Program, Faculty of Science and Technology, Ibrahimy University, East Java, Indonesia - 68374
[1]renorisma903@gmail.com, [2]zaeholfatah@gmail.com

Abstrak

Data mining membantu memberikan keputusan yang tepat dan cermat. Kelulusan mahasiswa tepat waktu merupakan salah satu point penilaian dalam proses akreditasi perguruan tinggi. Namun kelulusan mahasiswa tidak selalu dapat dideteksi secara cepat sehingga dapat mengurangi penilaian suatu perguruan tinggi dalam proses akreditasi. Permasalah inilah yang muncul untuk mengetahui mahasiswa nantinya bisa lulus tepat waktu atau tidak. metode klasifikasi untuk prediksi lulusan mahasiswa menggunakan algoritma Naïve Bayes. kelulusan mahasiswa tepat waktu atau tidak, yang diharapkan hasilnya dapat memberikan informasi dan masukan bagi pihak perguruan tinggi dalam membuat kebijakan kedepannya. Dari hasil pengujian ini didapatkan hasil bahwa dengan menerapkan algoritma Naïve Bayes sistem dapat memprediksi kelulusan mahasiswa dengan tepat waktu, Setelah membandingkan beberapa literatur, dapat disimpulkan bahwa metode ini dapat digunakan untuk prediksi tersebut dengan tingkat keakuratan 90%. Kajian literatur ini penting sebagai faktor pendukung bagi penelitian.

Kata kunci: *Data Mining, Naive Bayes, Prediction, Student Graduation*

Abstract

Data mining helps provide precise and careful decisions. Student graduation on time is one of the assessment points in the higher education accreditation process. However, student graduation cannot always be detected quickly, which can reduce the assessment of a university in the accreditation process. This problem arises to find out whether students will be able to graduate on time or not. Classification method for predicting student graduates using the Naïve Bayes algorithm. Whether a student graduates on time or not, it is hoped that the results will provide information and input for the university in making future policies. From the results of this test, it was found that by applying the Naïve Bayes algorithm the system can predict student graduation in a timely manner. After comparing several literatures, it can be concluded that this method can be used for this prediction with an accuracy rate of 90%. This literature review is important as a supporting factor for research.

Keywords : *Data Mining, Naive Bayes, Prediction, Student Graduation*

1. Introduction

Higher education is an educational institution that provides high-level learning services, which is an optional final stage in formal education. These higher education institutions are generally in the form of universities, academies, institutes or high schools. The types of higher education include vocational, academic and professional. Based on the level, higher education provides diploma, bachelor's, master's, specialist and doctoral programs[1].

Universities are expected to provide quality education for students so as to produce students who have competence in their fields. In Indonesia, the quality of a higher education institution is determined by the accreditation grade issued by the National Accreditation Board for Higher Education (BAN-PT). There are many aspects that can be used as a benchmark for the quality of higher education. One of them can be seen from the number of students who can complete their studies within the specified time or in other words, students graduate on time.

The more students who graduate on time, the better the performance of the higher education institution, so that the student graduation rate on time becomes one of the assessment criteria. accreditation for a university.

Many factors can support an increase in the number of students graduating on time, including:

1. Effective Academic Management
2. Relevant Curriculum
3. Academic Guidance Support

By implementing effective academic management, a relevant curriculum, and strong academic guidance support, universities can significantly increase the number of students graduating on time. This not only benefits students, but also increases the reputation and credibility of the educational institution[2]. Graduation data can be analyzed and processed in detail using data mining techniques.

Data mining is a process of extracting valuable knowledge or information from large and complex datasets. The main goal of data mining is to identify patterns, relationships, or information that may not be directly visible in the data, so that it can provide deeper and more valid information[3]. Data mining in another sense is a term used to describe the discovery of knowledge in databases, which uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and knowledge assembled from various large databases[4].

Student graduation rates are predicted using a system. However, some universities do not yet have a system to predict delays in student graduation, so universities cannot prevent this.

How can the Naïve Bayes algorithm be applied to predict student graduation on time? The Nave Bayes algorithm is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. Nave Bayes performance is competitive in the classification process even though it uses the assumption of attribute independence (no relationship between attributes). Naive Bayes is used for data classification techniques using probability and statistical methods that predict future opportunities based on previous experience, so it is known as Bayes' Theorem. This theorem is combined with Naive where it is assumed that the conditions between one attribute and another attribute are mutually independent. Naive Bayes classification assumes that the presence or absence of certain characteristics of one class has nothing to do with the characteristics of other classes[5].

From several explanations regarding graduate predictions, there are several differences, including using the Naive Bayes algorithm with a classification method using the attributes of student name, NPM, entry route, student GPA and gender. With these attributes and criteria data, the algorithm can predict student graduation rates. So, with this student graduation prediction system, it can help to find out whether students graduate on time or not, with the hope that the results can provide information and input for universities in making future policies.

Based on the existing problems, it is necessary to have a system to predict student graduation rates based on existing variables. Apart from that, a suitable algorithm is needed so that it can produce good accuracy values. With the system created, it is hoped that universities can make policies so that students can graduate on time.

2. Research Method / Proposed Method

This research began with problem formulation and literature study. A certain amount of data is required obtained through observation and documentation methods to be able to solve the problem. Next, the required student data was collected, with the acquisition of 146 data sets of students from the class of 2018 who had completed their studies in 15 attributes. Then in the second stage, data pre-processing is carried out to obtain good data before the data is processed using Naïve Bayes. If data pre-processing has been completed, 146 student data in 15 attributes will be used to carry out the mining process. Third, the mining process by applying the naïve Bayes algorithm. The fourth stage is experimentation and model testing. The fifth stage is evaluation and validation of test results. the results of the accuracy values obtained from applying the Naive Bayes algorithm. To calculate the accuracy of the patterns obtained, test the accuracy method in the concept of data mining. Accuracy calculations can be assisted by applying the Naive Bayes calculation formula. The results of applying the Naive Bayes

formula create very good accuracy based on the calculations of applying the Naive Bayes model [6].

3. Literature Study

3.1 Data collection

The techniques used to collect research data consist of:

1. Observation method

Efforts to obtain or collect data directly by observing in the field, which will then be useful in calculations when predicting the performance of systems running in the company[7].

2. Library method

Collect data by studying books, scientific journals, and information from the internet that supports research[8].

3.2 Data analysis

Student data in this research is data from Kaggle in the form of gender, student status, married status, age, Semester Achievement Index (IPS) from semester 1 to semester 8, and Cumulative Achievement Index (GPA). With the target class classification, namely graduation status which includes whether it is on time or late[9].

3.3 Naive Bayes algorithm

Naive Bayes is the simplest form of Bayesian network classifier. In naive Bayes, each feature node has a class node as its parent, but does not have the parents of any other feature nodes. However, a large body of work in supervised learning has shown that such a simple naive Bayes classifier can compete with state-of-the-art classifiers such as C4.5 and is still one of the top 10 data docking algorithms in the world[10].

3.4 Rapidminer

Rapidminer is a software for data science created by Rapidminer, Inc which provides integrated data, text mining, machine learning, and predictive analysis. As well as assisting with steps in a machine learning process such as visualization results, optimization, data preparation and model validation. Rapidminer brings artificial intelligence to the enterprise through an open and extensible data science platform. Built for analytics teams, Rapidminer brings together the entire data science cycle from data preparation to machine learning to predictive deployment (Hofmann & Klinkenberg, 2016)[11].

In simple terms, Rapidminer is an application that is used to process data using various techniques and methods in data mining, so that it can become useful information. Rapidminer is an open-source software that is useful for processing mining data. Rapidminer processes it by extracting patterns from datasets and combining them with statistical methods, artificial intelligence and databases. The aim of using Rapidminer is to obtain high quality information from the processed text. Rapidminer is used in various studies for examining researched data[12].

4. Result and Discussion

This section contains the results and discussion of the research topic, which can be created first as a research methodology. This section also represents explanations in the form of explanations, pictures, tables and others. The number of words in this section ranges.

Data collection

The data collection process was carried out by taking it from Kaggle. The data used is in the form of a Microsoft Excel file with .xlsx format. The data comes from the 2018 incoming class of students who have completed their studies, with a total of 146 data and 14 attributes. An example of the data used can be seen in the following image. This data was taken with the aim of further analysis in research projects, specifically for the purposes of Naive Bayes classification using RapidMiner. An example of the data used is in the picture.

NAMA	JENIS KELAMIN	STATUS MAHASISWA	UMUR	STATUS NIKAH	IPS 1	IPS 2	IPS 3	IPS 4	IPS 5	IPS 6	IPS 7	IPS 8	IPK	STATUS KELULUSAN
UNAMA	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.17	2.7	3.23	2.41	3	2.47	1.75	0	2.75	TEPAT
LEYLA TRIYANA PRATIWI	PEREMPUAN	MAHASISWA	26	BELUM MENIKAH	3.6	3.5	3.42	2.85	3.31	2.95	2.18		3.39	TEPAT
VERIS SOFIYAN PRAYOGA	LAKI - LAKI	MAHASISWA	29	BELUM MENIKAH	2.67	2.66	2.93	3.14	2.92	2.64	2.88	0.5	2.81	TEPAT
ADITYA AKBAR NUGRAHA	LAKI - LAKI	MAHASISWA	27	BELUM MENIKAH	2.48	2.86	2.09	2.55	2.55	2.43	2.55	2.17	2.82	TEPAT
ERNA EKA RIYANTI	PEREMPUAN	MAHASISWA	25	BELUM MENIKAH	3.19	3.08	3.31	2.83	3.36	2.73	3.06	0	3.09	TEPAT
FARID DWI NORYANTO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.1	2.98	3.17	3.25	3.41	3.08	3.43	3	3.23	TEPAT
DAFIK HADI WINOTO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	2.98	2.68	2.23	2.86	2.25	2.64	1.52	2.1	2.54	TEPAT
WAHYU FITRIYANTO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.45	3.15	3.54	3.78	3.42	3.88	2.5	4	3.56	TEPAT
IMAM SURYO SUSANTO	LAKI - LAKI	MAHASISWA	26	BELUM MENIKAH	3.31	3.02	3.48	3.7	3.19	3.21	2.58	4	3.4	TEPAT
AHMAD SUTOPO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	2.62	2.73	2.11	3.36	2.79	3.17	2.45	0	2.97	TEPAT
HERI SUFA'AT	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3.24	3.06	2.79	2.73	3.02	2.59	3.08	0	3.08	TEPAT
MURYA AMIEN NUR PRABOWO	LAKI - LAKI	MAHASISWA	26	BELUM MENIKAH	2.83	3.02	3.02	3.63	3.21	3.09	2.57	3	3.33	TEPAT
EDY PURNOMO	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	2.83	3	2.54	3.36	2.79	3.28	3.28	3	3.15	TEPAT
DIDIK KURNIAWAN	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	3	3.08	3.02	3.07	2.88	2.76	3	2.5	2.92	TEPAT
LYDIA EVITA SANDRA DEWI	PEREMPUAN	MAHASISWA	25	BELUM MENIKAH	3.52	3.52	3.31	3.33	3.3	2.94	3.46	0	3.26	TEPAT
DAVID KURNIAWAN	LAKI - LAKI	MAHASISWA	36	BELUM MENIKAH	2.29	0.5	0.16	0.32	1.13	0.37	0.28	0.11	1.02	TEPAT
DEVI KISTIANI	PEREMPUAN	MAHASISWA	25	BELUM MENIKAH	3.52	3.46	3.69	3.41	3.4	0	0		3.43	TEPAT
ADI PURWADI	LAKI - LAKI	MAHASISWA	27	BELUM MENIKAH	2.6	2.91	2.66	2.91	3.25	2.88	2.83	0	2.86	TEPAT
ARIF WIBOWO	LAKI - LAKI	MAHASISWA	30	BELUM MENIKAH	2.52	2.59	2.11	3	2.73	2.5	2.48	0	2.68	TEPAT
NOVAZAEN DWI SAPUTRO	LAKI - LAKI	MAHASISWA	28	BELUM MENIKAH	2.79	3.05	3.21	2.92	3.41	2.86	2.5	0	3.04	TEPAT
MOHAMAD ADE FADHORI	LAKI - LAKI	MAHASISWA	30	BELUM MENIKAH	2.93	2.52	1.95	2.37	3.07	2.91	1.73	1.6	2.81	TEPAT
SUBAEDAH	PEREMPUAN	MAHASISWA	23	BELUM MENIKAH	3.52	3.48	3.71	3.83	3.25	3.65	2.33	3.5	3.56	TEPAT
DANI PRIMAYANTI	PEREMPUAN	MAHASISWA	24	BELUM MENIKAH	3.48	3.73	3.6	3.52	2.9	3.32	3.21	0	3.45	TEPAT
DERA BAHTIAR WIDIYANTO	LAKI - LAKI	MAHASISWA	23	BELUM MENIKAH	2.52	3.3	3.79	3.71	3.65	2.1	3.59	4	3.52	TEPAT
MUKHAMAD YAHYA WICAKSON	LAKI - LAKI	MAHASISWA	24	BELUM MENIKAH	2.45	2.66	2.93	3.08	3.29	3.18	2.1	3	3.1	TEPAT
WAHYU DWI UTOMO	LAKI - LAKI	MAHASISWA	28	BELUM MENIKAH	2.62	1.75	1.95	2.17	1.95	2.18	1.45	1.08	2.34	TEPAT
AHMAD FIRDAUS MABRURI	LAKI - LAKI	MAHASISWA	25	BELUM MENIKAH	2.33	3	2.19	0.44	2.17	0.59	0.26		1.72	TEPAT
HARYOKO ABDUL HAMID	LAKI - LAKI	MAHASISWA	26	BELUM MENIKAH	3.24	3.46	3.56	3.83	3.73	3.74	2.33	3.5	3.58	TEPAT
ERSA GILANG ARDHANI	LAKI - LAKI	MAHASISWA	25	BELUM MENIKAH	2.95	2.91	3.44	3.38	3.33	3.36	2.08	0	3.1	TEPAT
NEORITA NUR FATIMAH	PEREMPUAN	MAHASISWA	25	BELUM MENIKAH	3.48	3.5	2.77	3.59	3.5	3.35	3.27	0	3.31	TEPAT
ULIL ALBAB	LAKI - LAKI	MAHASISWA	23	BELUM MENIKAH	2.93	3.18	3.31	3.25	3.44	3.22	3.07	3.5	3.33	TEPAT
DEWI KHOIRUN NISA'	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	3	3.33	3.27	2.96	3.16	3.12	2.61	0	3.08	TEPAT
AGUS NUGROHO	LAKI - LAKI	MAHASISWA	23	BELUM MENIKAH	2.83	3.11	3.1	2.96	3.2	3.53	2.48	3	3.27	TEPAT
MUKHAMAD LUQMAN HABIBI	LAKI - LAKI	MAHASISWA	25	BELUM MENIKAH	2.88	3.25	3.42	3.29	3.33	2.75	3	0	3.13	TEPAT
AMBAR SETIYANI	PEREMPUAN	MAHASISWA	25	BELUM MENIKAH	3.71	3.79	3.96	3.91	3.75	3.94	2.58	4	3.85	TEPAT
AVINA SEPTIANA	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	3.07	2.85	3.2	3	3.21	2.47	2.55	2.75	3.1	TEPAT
MARIA ULFA	PEREMPUAN	MAHASISWA	24	BELUM MENIKAH	3.19	2.9	2.59	2.77	3.09	2.83	3.11	2.9	3.08	TEPAT
IVANDHY SETYARACHMAN	LAKI - LAKI	MAHASISWA	31	BELUM MENIKAH	2.86	2.98	3.39	3.27	2.98	2.85	2.12		3.21	TEPAT
CHASAN MUBAROK	LAKI - LAKI	MAHASISWA	25	BELUM MENIKAH	2.98	3.18	3.73	3.42	3.5	3.47	2.58	3.5	3.52	TEPAT

Figure 2. 1 Sample Data

RapidMiner Studio 10.3 provides an efficient and effective environment for performing data testing and predictive analysis. The structured testing process and advanced features offered ensure that users can obtain accurate and reliable results. Here is the data testing process and display of RapidMiner Studio 10.3

Figure 2. 2 Display of Rapidminer Studio 10.3

Open the Rapidminer application, after it appears as shown, the start with display will appear then select blank process to open a new worksheet/open a file that has been saved as in Figure 2.3

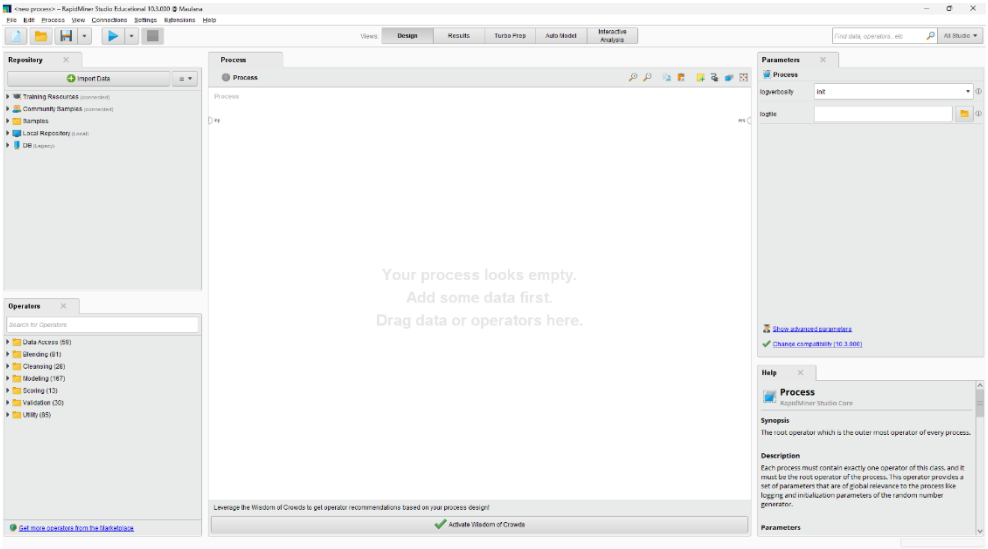

Figure 2. 3 Worksheet views

Next is the data input process and requires the read excel operator by dragging and dropping into the worksheet like the data retrieval process using double click on the read excel operator

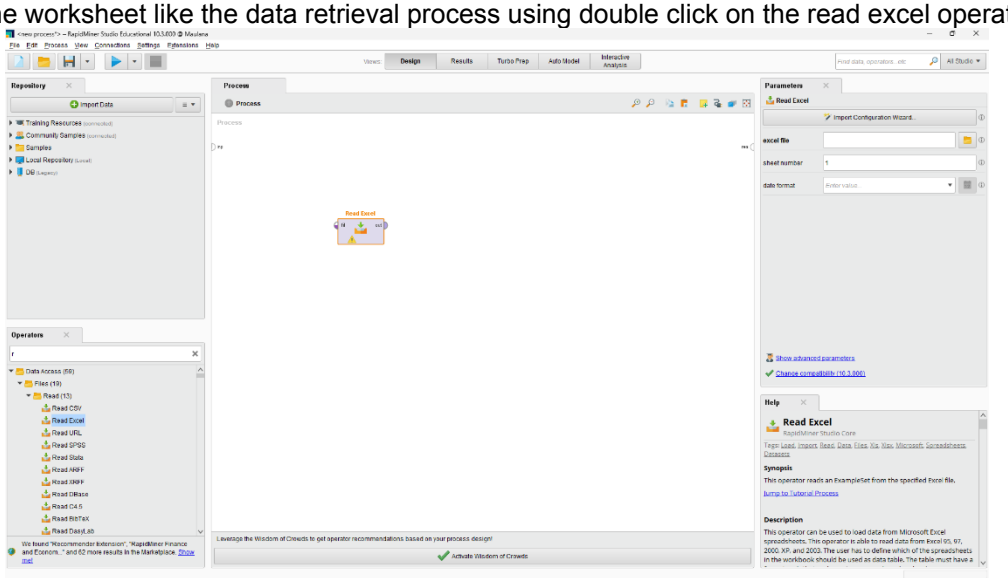

Figure 2. 4 Excel Read Operators

Click on the Read Excel module to open the file selection window. In this window, you will be given the option to search for and select the data file you want to import. Use the 'Browse' or 'Select File' button to open the file selection dialog. Navigate to the directory where the Excel file you want to use is stored, select the data file that suits your analysis needs. Make sure that the selected file has the correct format and that the required data is complete.

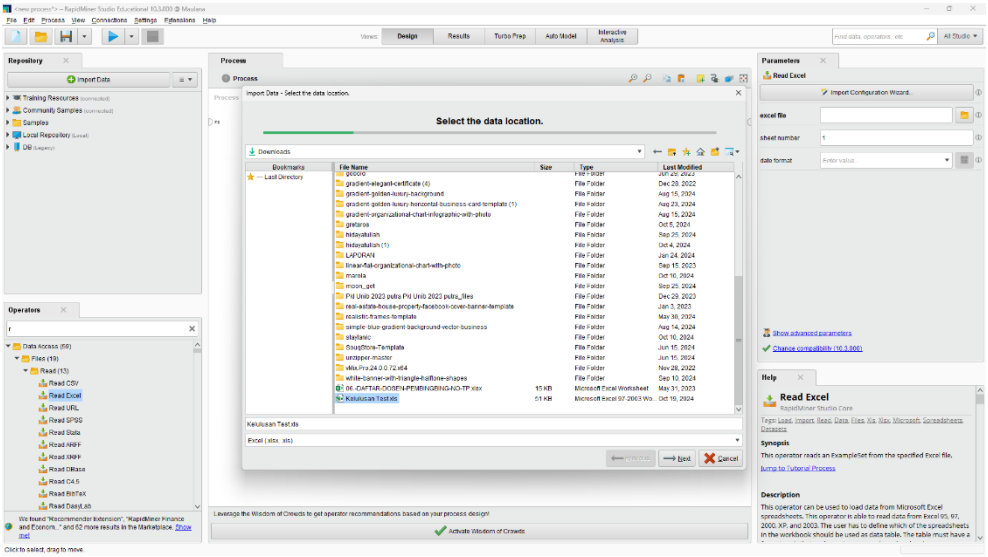

Figure 2. 5 Student data import

After selecting the file, the next step is to verify that the file has been properly loaded into RapidMiner. The Read Excel module will preview the data of the selected file, check the column headers, data format and ensure there are no errors in the imported data. If necessary, make additional settings such as setting the delimiter, sheet to be read, and cell range.

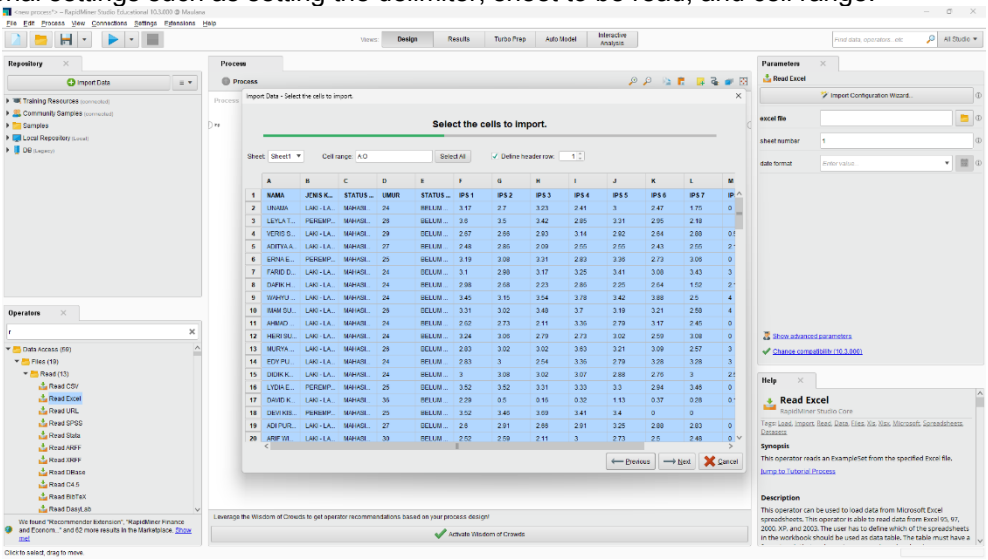

Figure 2. 6 file verification

Next, click next on the selected data. After that, determine the data type, then the data class is given a label attribute with the change role option and click finish, then it will look like Figures 7 and 8.

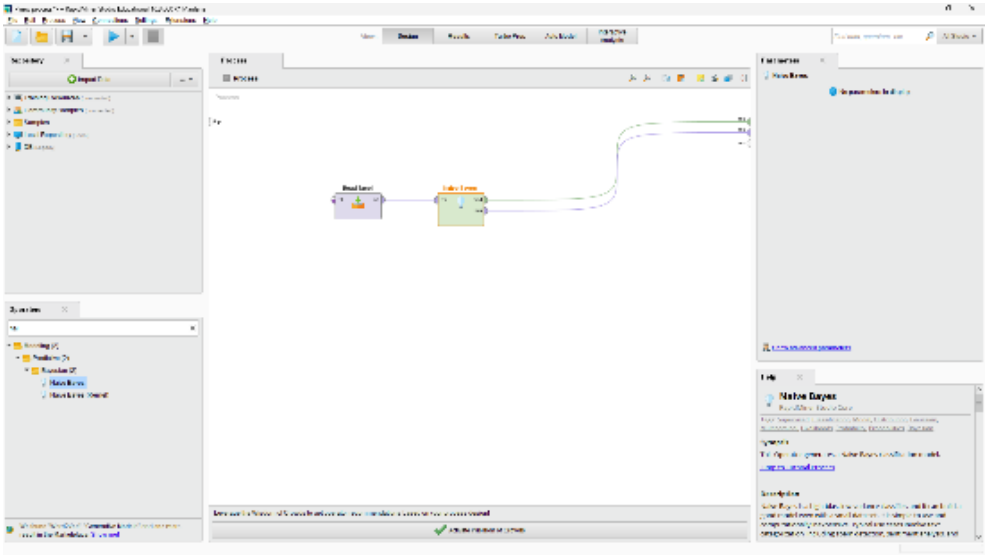

Figure 2. 7 Data import

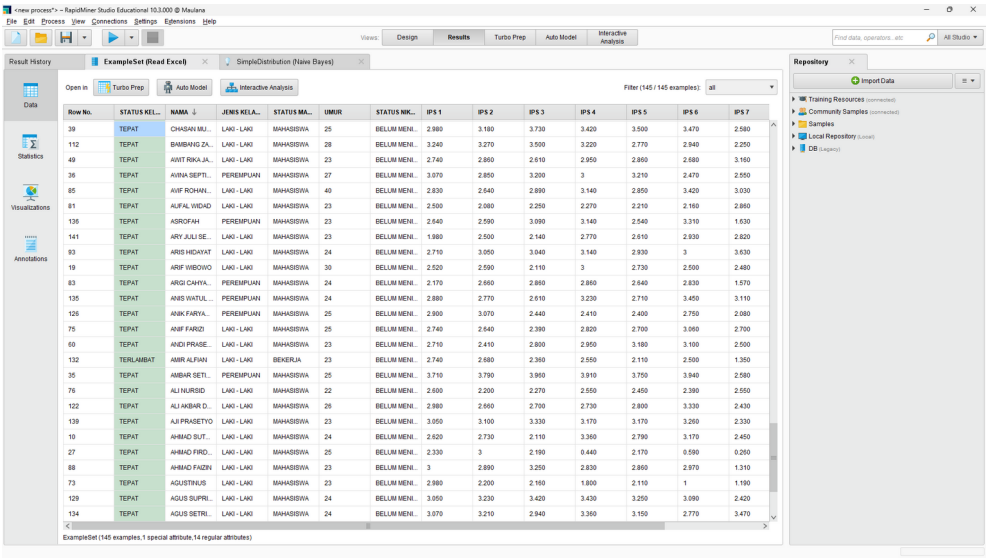

Figure 2. 8 Student Data Test Prediction Results

The number of testing data used was 146 students using the Naïve Bayes method. The prediction results were that 144 students graduated on time.

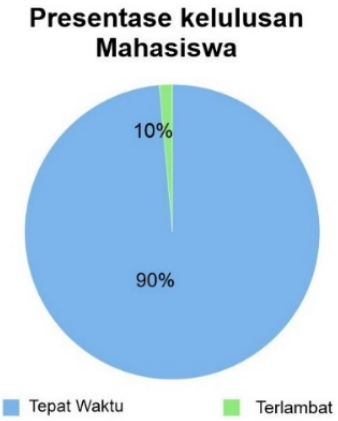

Figure 2. 9 Graphics of Student Graduation Percentage

Prediction results of data testing. The number of testing data used was 146 students using the Naïve Bayes method. The prediction results showed that 144 students graduated on time.
cleaned then used. The attributes used are gender, student status, age, cumulative achievement index (GPA). The data taken is divided into two groups, which have two data classes, namely on time and late. Next, testing with RapidMiner Studio 10.3 as a tool to obtain the level of accuracy of the Naïve Bayes method. The following is a calculation to get the accuracy value with the following equation:

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} = \text{X100\%}$$

$$= \frac{50+40}{50+5+5+40} \text{ x } 100\%$$

$$= 90\%$$

Predictions were made using training data using Naïve Bayes, resulting in 144 students predicted being on time, 2 students predicted being late.

5. Conclusion

This research shows that the Naïve Bayes method is effective in predicting student graduation on time with an accuracy rate of up to 90%. By utilizing attributes such as cumulative grade point average (GPA) and other academic data, this algorithm is able to identify relevant patterns to support policy decisions that increase on-time graduation. This implementation provides a practical solution for universities in managing student data to improve accreditation and institutional reputation.

References

[1] M. P. Dr. Achmad Fathoni Rodli and S. E. M. S. M. Nikma Yucha, *Manajemen Kinerja Institusi Perguruan Tinggi*. CV Rey Media Grafika, 2022. [Online]. Available: https://books.google.co.id/books?id=S6p-EAAAQBAJ

[2] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, “Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review,” *Fakt. Exacta*, vol. 13, no. 1, p. 35, 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[3] A. Wasik *et al.*, “Implementasi data mining untuk memprediksi penjualan accessoris handphone dan handphone terlaris menggunakan metode k-nearest neighbor (k-nn) 1,” vol. 1, no. 2, pp. 469–479, 2024.

[4] E. luthfi and U. Amikom, *Algoritma Data Mining*. Penerbit Andi, 2009. [Online]. Available: https://books.google.co.id/books?id=-Ojclag73O8C

[5] A. S. Huda, R. M. Awangga, and R. N. S. Fathonah, *Prediksi Penerimaan Pegawai Baru Dengan Metode Naive Bayes*. in Data Science. Kreatif, 2020. [Online]. Available: https://books.google.co.id/books?id=aawGEAAAQBAJ

[6] P. Studi, T. Informatika, A. A. Muin, P. Studi, and S. Informasi, “Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus : Data Mahasiswa Baru Perguruan Tinggi),” vol. 2, no. 1, 2016.

[7] T. P. Hamakonda, *Pengantar klasifikasi persepuluhan Dewey*. Gunung Mulia, 1978. [Online]. Available: https://books.google.co.id/books?id=l4yxjMA8rQcC

[8] W. I. Rahayu, M. H. K. Saputra, R. M. Awangga, and R. Habibi, *Penerapan Metode Naive Bayes dan Skala Likert Pada Aplikasi Prediksi Kelulusan Mahasiswa*. Kreatif, 2020. [Online]. Available: https://books.google.co.id/books?id=CZv9DwAAQBAJ

[9] S. P. M. S. Ni’matuzahroh and M. P. Susanti Prasetyaningrum, *OBSERVASI: TEORI DAN APLIKASI DALAM PSIKOLOGI*. in 1. UMMPress, 2018. [Online]. Available: https://books.google.co.id/books?id=CMh9DwAAQBAJ

[10] M. Zed, *Metode Penelitian Kepustakaan*. Yayasan Pustaka Obor Indonesia, 2008. [Online]. Available: https://books.google.co.id/books?id=zG9sDAAAQBAJ

[11] L. J. Moleong and T. Surjaman, *Metodologi penelitian kualitatif*. Remadja Karya, 1989. [Online]. Available: https://books.google.co.id/books?id=YXsknQEACAAJ

[12] M. Hasanah, N. H. Harani, N. Riza, R. M. Awangga, and R. Habibi, *Implementasi Barcode Dan Algoritma Regresi Linear Untuk Memprediksi Data Persediaan Barang*. Kreatif. [Online]. Available: https://books.google.co.id/books?id=db38DwAAQBAJ