Published online on the page : https://ejournal.uinbukittinggi.ac.id/index.php/ijokid/

Knowbase : International Journal of Knowledge in Database

| ISSN (Print) 2798-0758 | ISSN (Online) 2797-7501 |

Data Mining Analysis to Predict Student Skills Using Naïve Bayes Method

Yaslinda Lizar[1], Alya Sahira Firrizqi[1], Asriwan Guci[2], Joko Sunadi[3]*

[1]*Universitas Islam Negeri Imam Bonjol, Padang, Indonesia*
[2]*STIKes MERCUBAKTIJAYA, Padang, Indonesia*
[3]*Universitas Islam Negeri Imam Bonjol, Padang, Indonesia*

Article Information

Article History:
Accepted by the Editor: December 12, 2023
Final Revision: December 22, 2023
Published Online: December 31, 2023

Keywords

Data Mining
Naïve Bayes Method
Data Classification
Students
Skills

Correspondence

E-mail: yaslinda@uinib.ac.id*

A B S T R A C T

The possession of specific skills by students not only has a positive impact on the students themselves but also on the Study Program within a Faculty and the University as a whole. However, Study Programs sometimes face difficulties in determining the skills of numerous students even after they have completed 7 semesters of study. Therefore, a method to extract available data in order to determine student skills quickly and accurately is essential. This research aims to apply a data mining method to predict student skills in the Information Systems Study Program at UIN Imam Bonjol Padang. The study focuses solely on predicting student skills in the fields of data processing and programming. The method employed in this data mining analysis is the Naïve Bayes method. Data will be collected from student course grades related to data processing and programming. The data will be processed using an application and subsequently tested using a Confusion Matrix. The research results indicate that predicting the determination of student skills in the Information Systems Study Program at UIN Imam Bonjol can be achieved using the Naïve Bayes algorithm, which yielded a Naïve Bayes model accuracy of 93%, precision of 81%, and recall of 81%. The obtained model can be implemented in the form of an application to determine decision-making strategies for students.

This is an open access article under the CC–BY-SA license

1. Introduction

Students who possess above-average potential are not only academically excellent but also consistently demonstrate creativity, innovation, and creative thinking. They also exhibit special skills or abilities [1]. Currently, globalization is deeply integrating into human life, influencing nearly all aspects of human existence. Moreover, [2] It describes the world today as akin to someone who goes to sleep and wakes up in the morning to find that everything has changed. With the ever-increasing changes in the era, this also impacts the demands within the working world, which are becoming more rigorous.

Individuals belonging to the generation living in the digital age not only work within office settings but due to the rapid advancement of technology, work can now be carried out from anywhere [1]. Essentially, it means that due to the changing work styles – previously confined to office spaces but now adaptable to any location thanks to technology allowing remote work and international client meetings through teleconferencing – everyone is now required or compelled to improvise and possess specific skills that can stand out in today's job market. It's no longer solely reliant on academic achievements; instead, having a set of distinct skills is highly crucial.

Therefore, to determine the skills possessed by students, a method that can address the aforementioned question is required. At present, data mining is employed to aid in problem-solving and decision-making. Predicting the aforementioned aspect cannot be achieved without identifying and employing a data mining algorithm [3]. Several studies related to data mining have been utilized to enhance the accreditation of a higher education institution, Universitas Dian Nuswantoro, in 2012. These studies were based on the abundance of student data and graduation figures. Data mining was employed to extract insights, allowing for valuable information mining processes beneficial for the university [4].

This research was conducted on students of the Information Systems Study Program at UIN Imam Bonjol Padang to assist the program in guiding students towards their research directions. Furthermore, it allows the program to identify which courses are selected as supporting electives for the students' skills.

Data mining is a technique employed for data extraction to create a model. From this model, data patterns are identified, and information is extracted [5]. Among the numerous techniques in data mining, one of them is classification. This classification technique is a learning method used to predict the value of a target variable [6]. As this research involves two variables: students with data processing skills and students possessing coding skills, it will utilize the classification technique in data mining.

2. Method

2.1. Data Mining

Data Mining is a process that encompasses various computer learning techniques, also known as machine learning, used to analyze and automatically extract or input knowledge. It is often associated with a method used to uncover discoveries and search for information and knowledge within a database, data warehouse, and big data. In research, the representation of data mining involves the application in data excavation, machine learning, statistics, and the generation of resulting information [7].

From the definition itself, data mining involves 'mining', where it is a process of digging or extracting knowledge from a vast amount or large volume of data [1]. The aim of data mining is to uncover related relationships or patterns that offer valuable insights to its users.

Data mining is a component of KDD, an acronym for Knowledge Discovery in Databases, which is tasked with extracting models from data using specific algorithms [8]. "The process of KDD is as follows: (a) Data selection. It involves the selection of data from a dataset before the information excavation process takes place; (b) Preprocessing. It is the process where, before data mining is initiated, there is a need for a cleaning process aimed at eliminating redundant data, examining inconsistent data, and rectifying errors such as printing mistakes or typos; (c) Transformation. It is the coding process applied to selected data, which significantly determines the type of data or information patterns sought within the database; (d) Data mining. It is the process of mining data or searching for information using specific techniques or methods; and (e) Interpretation/Evaluation. It is the process of evaluating the discovered data to determine its alignment with the existing facts or hypotheses.

2.2. Classification Techniques

Classification is one of many techniques within data mining. This classification technique can be used to predict or forecast trends that will occur in the future [4]. In this classification, the process will categorize data into several variables with the aim of forming a model and considering influential attributes [9].

The components of this Classification Technique are as follows: (a) Class. This represents the label determined from the classification result; (b) Predictor. This denotes the independent variable determined based on the results of the attributes' characteristics that have been classified; (c) Training data. This encompasses a large set of collected data containing classes and predictors. Its purpose is to train to correctly and accurately classify into respective classes; and (d) Testing data. This refers to a set of new data that will be categorized to determine the accuracy level of the previously created model. Typically, the testing data set is smaller in size compared to the training data.

2.3. *Naïve Bayes Algorithm*

The Naïve Bayes algorithm is a classification method used for predicting probabilities. Naïve Bayes is based on Bayes' theorem and has the ability to classify similar to a decision tree. It exhibits high accuracy when applied to databases [9]. The algorithm demonstrates a high level of accuracy in classifying data [10]. Here is the formula for the Naïve Bayes theorem :

$$P(H|E) = \frac{P(E|H)\ x\ P(H)}{P(E)}$$

Explanation:

E : Data with an unknown class

H : Hypothesis that data E belongs to a specific class

P(H|E) : Probability of hypothesis H given condition E

P(H) : Probability of hypothesis H

P(E|H) : Probability of E given condition H

P(E) : Probability of hypothesis E

2.4. *Confusion Matrix*

The confusion matrix is a commonly used tool in data mining utilized to evaluate the classification model for predicting correct or incorrect objects [11]. The confusion matrix can also be defined as a table that provides comprehensive data that has been tested to determine its accuracy level [12]. The matrix of predictions will be compared with the class of testing and training data containing actual and predicted values in the classification. Evaluating using this confusion matrix will yield an accuracy value [13]. Table 1 below provides information for the confusion matrix.

Table 1. Confusion Matrix

Classification	**Predicted Class**	
	Class = Yes	**Class = No**
Class = Yes	a (true positive-TP)	b (false negative-FN)
Class = No	c (false positive-FP)	d (true negative-TN)

And here is the formula for calculating the level of accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{A+D}{A+B+C+D}$$

True Positive (TP) is the count of positive records present in the dataset classified as positive. True Negative (TN) is the count of negative records in the dataset classified as negative. False Positive (FP) is the count of negative records in the dataset classified as positive. False Negative (FN) is the count of positive records in the dataset classified as negative [13].

$$Precision = \frac{TP}{FP+TP}$$

Precision is utilized to measure the extent of correctly predicted positive class data proportionately from the overall predicted positive class outcomes.

$$Recall = \frac{TP}{FN+TP}$$

Recall is used to indicate the percentage of positive class data correctly predicted from the entirety of the positive class data.

2.5. *Research Methodology*

The research methodology employed in this research involves the use of quantitative research methods, employing data analysis techniques. The tools utilized include the data mining application Orange, while the testing model employs the Confusion Matrix.

2.5.1. *Method of Collecting Data*

The interview method involves direct engagement related to the research topic to adjust the data process that will be subsequently analyzed. The questionnaire dissemination method involves distributing questionnaires to collect research data. Literature review, in this method, comprises literature related to the conducted research. The literature review encompasses relevant journals, scientific articles, and referenced books.

2.5.2. *Data Grouping Methods*

Larose states that data mining is categorized into several groups based on the tasks that can be executed, namely: (a) Explanation/Description. At times, researchers and analysts seek ways to explain patterns or trends existing within the data; (b) Estimation. Estimation is akin to Classification, except that the target variable for estimation is numerical, not categorical; (c) Prediction. Prediction resembles Classification and Estimation, but the values produced from Prediction are for the future; (d) Classification. Classification involves targeting categorical variable(s). (e) Clustering. Clustering is a method to discover and group data with similar characteristics among themselves. Clustering is an unsupervised data mining technique; and (f) Association. The task of association in data mining is to find attributes occurring at specific points in time.

2.5.3. *Research Stages*

The stages conducted in this research as in Figure 1.

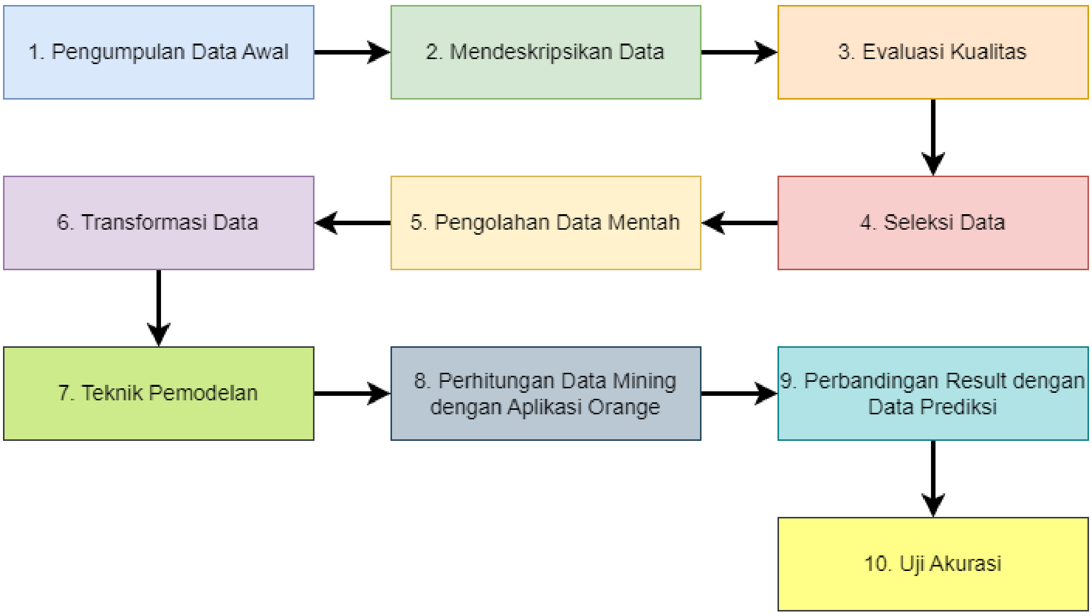


Figure 1. Research Stages

3. Results and Discussion

The data collected for this research was obtained from students enrolled in the Information Systems program at the Faculty of Science and Technology, UIN Imam Bonjol Padang. The dataset comprises 70 records. Figure 2 is research data.

1	Jenis Kelamin	Umur	Sistem Basis Data	Prak. Sistem Basis Data	Data Warehouse	Prak. Data Warehouse	Object Oriented	Prak. Object Oriented	Pemrograman Web	Prak. Pemrograman web	Prediksi Skill
2	P	21	A	A	A	B	B	B	B	A	Pengolahan Data
3	L	21	B	B	A	B	B	B	B	B	Pengolahan Data
4	P	20	B	A	A	A	A	A	B	B	Pengolahan Data
5	P	21	A	A	A	A	B	B	B	B	Pengolahan Data
6	L	19	C	C	B	B	B	B	B	B	Coding
7	L	19	B	B	C	B	B	B	B	B	Coding
8	P	19	B	B	C	B	C	C	B	B	Pengolahan Data
9	P	21	B	B	A	C	B	B	B	B	Coding
10	P	20	A	A	A	A	B	B	B	B	Pengolahan Data
11	P	20	B	B	A	B	A	A	B	A	Coding
12	L	21	B	B	A	A	A	A	B	A	Coding
13	L	20	A	A	A	A	A	A	A	A	Coding
14	P	20	B	B	B	B	C	B	B	B	Pengolahan Data
15	P	20	A	B	A	B	B	B	B	B	Pengolahan Data
16	L	20	B	A	A	A	A	A	A	A	Coding
17	P	21	B	A	A	B	B	B	A	B	Pengolahan Data
18	L	20	C	B	C	B	B	B	B	A	Coding
19	P	21	A	A	A	B	B	B	B	A	Pengolahan Data
20	L	21	B	B	A	B	B	B	B	B	Pengolahan Data
21	P	20	B	A	A	A	A	A	B	B	Pengolahan Data
22	P	21	A	A	A	A	B	B	B	B	Pengolahan Data
23	L	19	C	C	B	B	B	B	B	B	Coding
24	L	19	B	B	C	B	B	B	B	B	Coding
25	P	19	B	B	C	B	C	C	B	B	Pengolahan Data
26	P	21	B	B	A	C	B	B	B	B	Coding
27	...										
71	L	21	B	B	A	B	B	B	B	B	Pengolahan Data

Figure 2. Dataset

The dataset in this study comprises several attributes: gender, age, grades for the following courses - database systems, database systems lab, data warehouse, data warehouse lab, object-oriented programming, object-oriented programming lab, web programming, and web programming lab. The total dataset size for this research is 70 records.

The evaluation of the data utilized in the study did not uncover any null values or commonly referred to as missing values within the student dataset used. The dataset consisting of 70 data will be taken for testing data for processing. In this research, 30 testing data from 70 datasets were used. Data processing is the stage where one ensures that all selected datasets are suitable for further analysis or what is commonly referred to as Data Preprocessing.

Data transformation, this stage involves initializing the data that will be the target of analysis. In this research, the selected target data focuses on the predictive attribute of skills: (a) Students categorized with the skill prediction initiated as "YES" are those who possess skills in data processing; and (2) Students categorized with the skill prediction as "NO" are those who possess skills in coding. Figure 3 below is the initialized dataset.

1	Jenis Kelamin	Umur	Sistem Basis Data	Prak. Sistem Basis Data	Data Warehouse	Prak. Data Warehouse	Object Oriented	Prak. Object Oriented	Pemrograman Web	Prak. Pemrograman web	Target
2	P	21	A	A	A	B	B	B	B	A	YES
3	L	21	B	B	A	B	B	B	B	B	YES
4	P	20	B	A	A	A	A	A	B	B	YES
5	P	21	A	A	A	A	B	B	B	B	YES
6	L	19	C	C	B	B	B	B	B	B	NO
7	L	19	B	B	C	B	B	B	B	B	NO
8	P	19	B	B	C	B	C	C	B	B	YES
9	P	21	B	B	A	C	B	B	B	B	NO
10	P	20	A	A	A	A	B	B	B	B	YES
11	P	20	B	B	A	B	A	A	B	A	NO
12	L	21	B	B	A	A	A	A	B	A	NO
13	L	20	A	A	A	A	A	A	A	A	NO
14	P	20	B	B	B	B	C	B	B	B	YES
15	P	20	A	B	A	B	B	B	B	B	YES
16	L	20	B	A	A	A	A	A	A	A	NO
17	P	21	B	A	A	B	B	B	A	B	YES
18	L	20	C	B	C	B	B	B	B	A	NO
19	P	21	A	A	A	B	B	B	B	A	YES
20	L	21	B	B	A	B	B	B	B	B	YES
21	P	20	B	A	A	A	A	A	B	B	YES
22	P	21	A	A	A	A	B	B	B	B	YES
23	L	19	C	C	B	B	B	B	B	B	NO
24	L	19	B	B	C	B	B	B	B	B	NO
25	P	19	B	B	C	B	C	C	B	B	YES
26	P	21	B	B	A	C	B	B	B	B	NO
27	...										
71	L	21	B	B	A	B	B	B	B	B	YES

Figure 3. Dataset After Initialization

The technique adopted for this research involves utilizing the Naïve Bayes algorithm, and the tool employed for this purpose is the data mining application called Orange. Figure 4 is the data processing carried out in the Orange data mining application using the Naïve Bayes algorithm. At this stage, we determine the target of the training data table. In this research, the target is the target table.

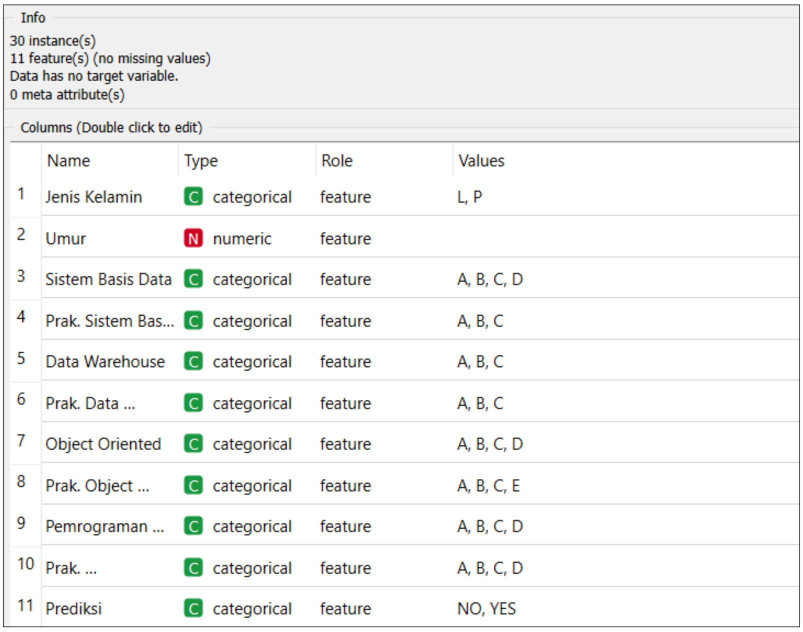
Info
70 instance(s)
11 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	Jenis Kelamin	C categorical	feature	L, P
2	Umur	N numeric	feature	
3	Sistem Basis Data	C categorical	feature	A, B, C
4	Prak. Sistem Bas...	C categorical	feature	A, B, C
5	Data Warehouse	C categorical	feature	A, B, C
6	Prak. Data ...	C categorical	feature	A, B, C
7	Object Oriented	C categorical	feature	A, B, C
8	Prak. Object ...	C categorical	feature	A, B, C
9	Pemrograman ...	C categorical	feature	A, B
10	Prak. ...	C categorical	feature	A, B
11	**Target**	C **categorical**	**target**	**NO, YES**

Figure 4. Columns Data Training

Figure 5 is a column of testing data in the orange application.



Info
30 instance(s)
11 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	Jenis Kelamin	C categorical	feature	L, P
2	Umur	N numeric	feature	
3	Sistem Basis Data	C categorical	feature	A, B, C, D
4	Prak. Sistem Bas...	C categorical	feature	A, B, C
5	Data Warehouse	C categorical	feature	A, B, C
6	Prak. Data ...	C categorical	feature	A, B, C
7	Object Oriented	C categorical	feature	A, B, C, D
8	Prak. Object ...	C categorical	feature	A, B, C, E
9	Pemrograman ...	C categorical	feature	A, B, C, D
10	Prak. ...	C categorical	feature	A, B, C, D
11	Prediksi	C categorical	feature	NO, YES

Figure 5. Data Testing

Figure 6 is a process carried out in a human application using the Naïve Bayes algorithm.

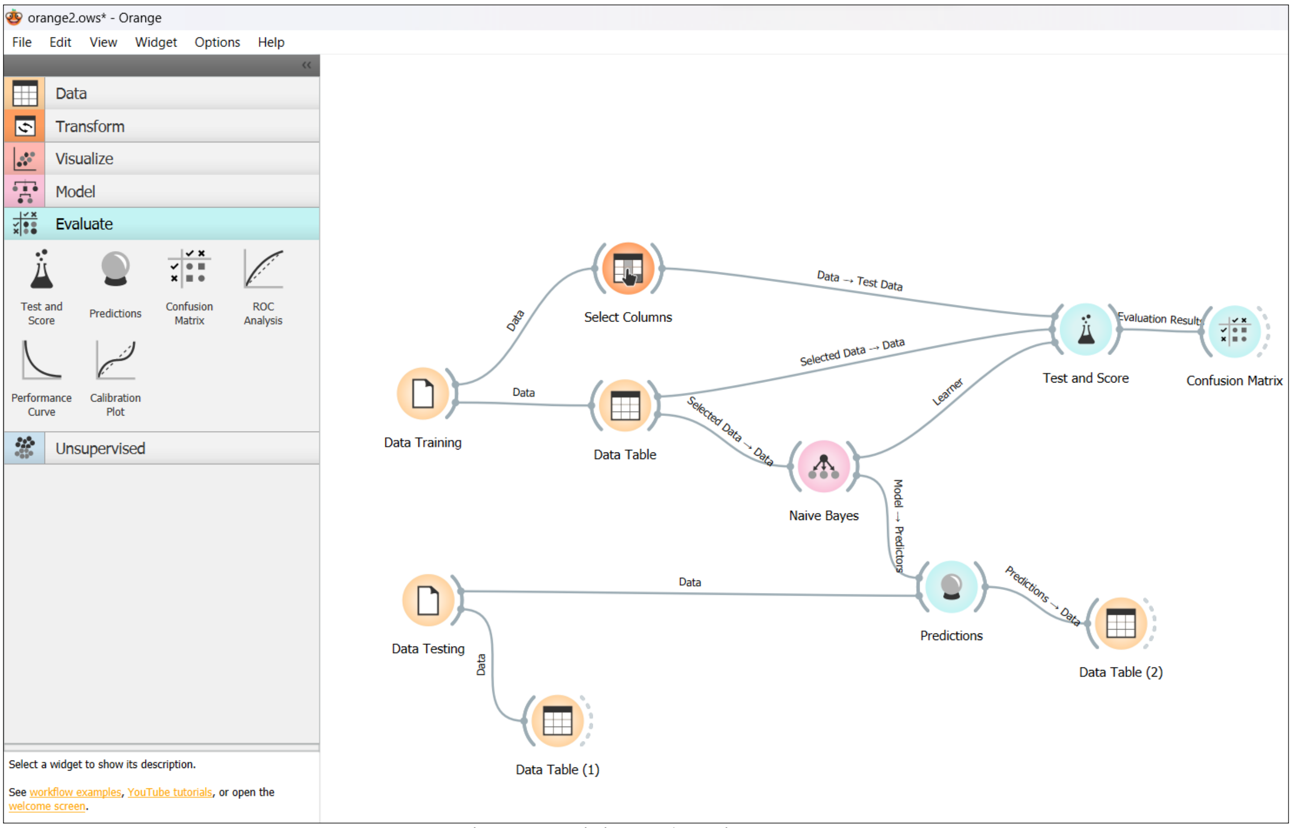


Figure 6. Training and Testing Process

Figure 7 represents the result or outcome of the process that has been conducted. The results can be observed on the left-hand side under the heading “Naïve Bayes”. From the obtained results using the Orange application, in the first process of data testing, the conclusion was “NO”; in the second process, the result was “NO”; in the third process, the result was “NO”, and in the fourth process, the result was “YES”.

Predictions - Orange

Show probabilities for (None)

	Naive Bayes	Jenis Kelamin	Umur	Sistem Basis Data	ak. Sistem Basis Da	Data Warehouse	ak. Data Warehou	Object Oriented	ak. Object Oriente	Pemrograman Wel	k. Pemrograman v	Prediksi
1	YES	P	20	A	A	A	A	B	B	B	B	YES
2	NO	L	19	B	C	B	A	B	B	B	C	NO
3	NO	P	20	B	B	C	B	B	A	A	B	YES
4	YES	L	20	A	B	A	C	C	B	B	B	YES
5	NO	P	19	B	A	A	C	B	A	A	C	YES
6	NO	L	20	B	B	C	A	A	B	B	A	YES
7	NO	P	20	B	A	A	B	B	A	A	A	NO
8	NO	L	21	C	B	A	B	B	A	B	B	YES
9	YES	P	21	B	A	A	A	A	C	B	B	YES
10	YES	L	21	B	A	A	A	D	B	B	B	YES
11	NO	L	20	A	C	B	A	B	B	B	C	NO
12	NO	L	19	A	B	C	B	B	B	B	B	NO
13	YES	P	19	A	B	C	B	B	A	B	B	YES
14	YES	P	21	B	B	A	C	C	B	B	C	NO
15	NO	L	20	B	A	A	A	A	B	C	B	YES
16	YES	P	20	A	B	A	B	B	A	D	B	NO
17	NO	P	21	B	B	A	A	A	B	B	A	NO
18	NO	L	20	B	A	C	A	D	B	A	A	NO
19	YES	P	21	B	B	B	B	B	B	B	B	YES
20	YES	L	20	B	B	A	B	B	B	B	B	YES
21	NO	L	19	A	A	A	A	A	A	A	D	NO
22	YES	P	21	B	A	A	B	B	B	A	B	YES
23	NO	L	20	A	B	C	B	B	E	B	A	NO
24	NO	P	19	D	A	A	B	B	B	B	A	YES
25	NO	L	21	C	B	A	B	B	B	B	B	YES
26	YES	P	20	B	A	A	A	A	A	B	C	YES
27	YES	P	21	B	A	A	A	A	B	B	B	YES
28	NO	L	19	C	C	B	B	B	B	B	C	NO
29	YES	L	20	A	B	C	B	B	B	B	B	NO
30	YES	P	19	A	B	C	B	B	C	B	A	YES

Figure 7. Result

Figure 8 displays the comparison between the predictions generated from the processed testing data within the Orange application and the outcomes produced by the Orange application itself.

	Prediksi	Naive Bayes
1	YES	YES
2	NO	NO
3	YES	NO
4	YES	YES
5	YES	NO
6	YES	NO
7	NO	NO
8	YES	NO
9	YES	YES
10	YES	YES
11	NO	NO
12	NO	NO
13	YES	YES
14	NO	YES
15	YES	NO
16	NO	YES
17	NO	NO
18	NO	NO
19	YES	YES
20	YES	YES
21	NO	NO
22	YES	YES
23	NO	NO
24	YES	NO
25	YES	NO
26	YES	YES
27	YES	YES
28	NO	NO
29	NO	YES
30	YES	YES

Figure 8. Comparison of Prediction Data with Results

The subsequent step involves calculating the accuracy by comparing the predicted data with the obtained results using the Confusion Matrix. The results are as in Figure 9 and Figure 10.

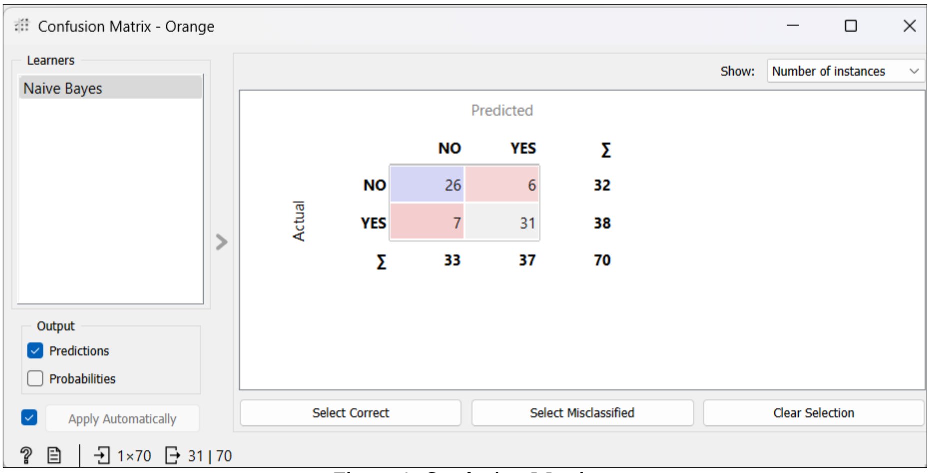


Figure 9. Confusion Matrix

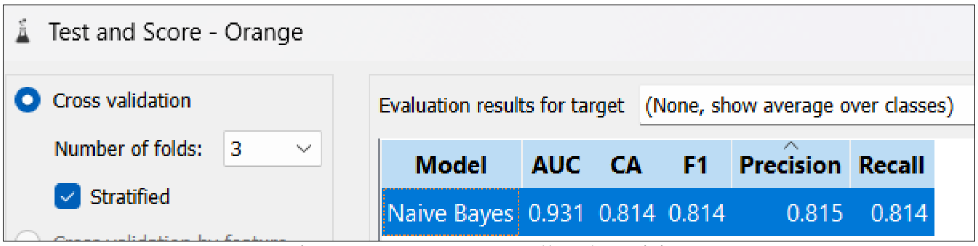


Figure 10. Test Score Recall and Precision

4. Conclusion

Therefore, from the conducted research using the Naïve Bayes algorithm calculated with the Orange application tool, with a total of 70 datasets and 30 data points for testing, the following results were obtained: (a) Based on the data mining calculations using the Orange tool and the Naïve Bayes algorithm, the prediction indicates that among the students, 14 individuals possess data processing skills initialized as YES. Meanwhile, 16 students exhibit coding skills with an initialization of NO. Consequently, it can be inferred that students enrolled in the Information Systems program at UIN Imam Bonjol have a greater proficiency in coding compared to data processing; and (b) Through the findings, discussions, and analyses, it is concluded that predicting the skill determination of students enrolled in the Information Systems program at UIN Imam Bonjol can be accomplished using the Naïve Bayes algorithm. The obtained model's accuracy is 93%, precision is 81%, and recall is 81%. This model can be implemented as an application to strategize decision-making for students.

References

[1] R. Ningsih, "Perubahan Dunia Kerja The Identity Status Profiles of Late Adolescents in Yogyakarta (Indonesia) and Its Microsystem Context Influence View project", doi: 10.13140/RG.2.2.11172.01922.

[2] H. A. R. Tilaar, Perubahan Sosial dan Pendidikan. PT. Gramedia Widiasarana Indonesia, 2002.

[3] J. H. Shindo, M. M. Mjahidi, and M. D. Waziri, "Data Mining Algorithms for Prediction of Student Teachers' Performance in Ict: a Systematic Literature Review," Inf. Technol. Learn. Tools, vol. 96, no. 4, pp. 29–45, 2023, doi: 10.33407/itlt.v96i4.5246.

[4] J. O. Ong, "Implementasi Algotritma K-means clustering untuk menentukan strategi marketing president university," J. Ilm. Tek. Ind., vol. vol.12, no, no. juni, pp. 10–20, 2013.

[5] O. Yuda and S. Nugroho, "Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro."

[6] Kusrini and E. T. Luthfi, Algoritma Data Mining. Yogyakarta: Penerbit Andi, 2009.

[7] B. M. M. Alom and M. Courtney, "Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia," Int. J. Inf. Technol. Comput. Sci., vol. 10, no. 2, pp. 1–9, 2018, doi: 10.5815/ijitcs.2018.02.01.

[8] M. Ardiansyah Sembiring, M. Fitri Larasati Sibuea, A. Sapta, P. Studi Sistem Informasi, and S. Royal, "Analisa

Kinerja Algoritma C.45 Dalam Memprediksi Hasil Belajar," J. Sci. Soc. Res., vol. 1, no. February, pp. 73–79, 2018, doi: http://dx.doi.org/10.54314/jssr.v1i1.110.

[9] R. Kumara and C. Supriyanto, "Klasifikasi Data Mining Untuk Penerimaan Seleksi Calon Pegawai Negeri Sipil 2014 Menggunakan Algoritma Decision Tree C4.5," 2013.

[10] F. Nurhuda, S. W. Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," ITSmart J. Teknol. dan Inf., vol. 2, no. 2, pp. 35–42, 2013, doi: doi:10.20961/its.v2i2.630.

[11] F. Gorunescu, Data Mining Concept, Models and Techniques, 1st ed. Springer Berlin, Heidelberg, 2011. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-19721-5

[12] D. Putra and A. Wibowo, "Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes," Pros. Semin. Nas. Ris. Dan Inf. Sci., vol. 2, pp. 84–92, 2020, doi: DOI:10.30645/SENARIS.V2I0.147.

[13] M. F. Rifai, H. Jatnika, and B. Valentino, "Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS)," Petir, vol. 12, no. 2, pp. 131–144, 2019, doi: 10.33322/petir.v12i2.471.