Prediction of Interest in The D3 Nautical and D3 Port and Shipping Management Study Program for New Student Admissions at State Polytechnic of Bengkalis Using Naïve Bayes

Zulyani[1,a)], Robyansyah[1,b)], Nur Rahmani[1], Handro Okta Prianus[1)]

[1]*Maritime Department, Polytechnic of Bengkalis, Bengkalis, Indonesia*

[a)]Corresponding author: zulyani@polbeng.ac.id
[b)]robyansyah147@gmail.com

Abstract. The selection of an academic program at a higher education institution is a crucial decision that can significantly impact a student's academic journey. By choosing an appropriate program, students are more likely to be immersed in an academic environment with peers, faculty, and learning conditions that align with their goals and abilities. This fosters a positive and supportive atmosphere, enables students to excel academically and acquires the necessary knowledge to succeed in the workforce post-graduation. The Maritime Department at State Polytechnic of Bengkalis offers two study programs: D3 Nautical and D3 Port and Shipping Management. In recent years, there has been a noticeable decline in the interest of prospective students in the Nautical program, while the Port and Shipping Management program has experienced a steady increase in interest. This research employs the Naïve Bayes algorithm as a predictive model to classify student interest in the Maritime Department's programs for the 2024 new student admissions. The study focuses on predicting enrollment trends for the D3 Nautical and D3 Shipping Management programs. The results demonstrate that the Naïve Bayes algorithm is capable of predicting program interest for new student admissions in the Maritime Department with an accuracy of 54%, a precision of 77%, and a recall of 58%.

Keywords : Interest, Student, Predictive Model

INTRODUCTION

Every year, Politeknik Negeri Bengkalis conducts new student admissions, including re-registration at the New Student Admission Center of Politeknik Negeri Bengkalis. Based on the data of student registrations in the Maritime Department for the period of 2019-2023, there is a trend indicating an increase in students completing the re-registration process, while the number of students not completing the re-registration in the D3 nautical program has shown a decreasing trend. However, this number remains relatively high. This phenomenon is intriguing for further research to reduce the number of students not completing re-registration. The D3 nautical program conducted an investigation, which found that students who registered but did not re-register often chose other higher education institutions.

One alternative solution is to search for insights from the dataset of the first wave of new student admissions for the academic year 2024/2025. These insights are expected to provide a distribution map of the students' geographical origins who chose the D3 nautical and D3 Port and Shipping Management programs. Subsequently, a model will be developed from this distribution data to predict the interest in study programs for the next wave of new student admissions.

Based on the phenomenon and background mentioned above, a strategy is required by the D3 nautical program to further expand its market share. Therefore, the researcher is interested in conducting a study titled "Prediction of Interest in the D3 nautical program in the Admissions of Politeknik Negeri Bengkalis Students." The algorithm used in this research is Naïve Bayes. If the prediction indicates interest in the D3 nautical program, prospective students

will be given special rewards, which is expected to increase their interest in choosing this program in the future.

METHODS

This study applies the Naïve Bayes model, a classification algorithm in data mining, to predict the interest in the Nautical and Port and Shipping Management diploma programs during student admissions at Politeknik Negeri Bengkalis. The research design uses a self-controlled experimental method to explore causal relationships. The aim is to develop a model that can be used to predict interest in the Nautical diploma program under the Department of Maritime Affairs at Politeknik Negeri Bengkalis. The study adheres to established guidelines by following the six stages of the Cross Industry Standard Process for Data Mining (CRISP-DM). The stages are as follows:

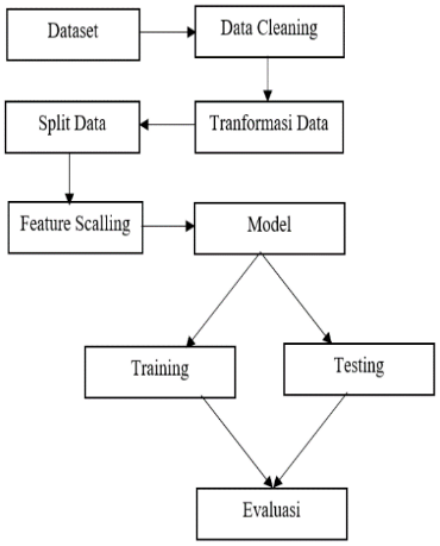

FIGURE 1. Data Processing Stages

1. Dataset
   The dataset used in this study is from the first wave of student admissions for the Academic Year 2024/2025, in the form of an Excel file (.xlsx)
2. Data Cleaning
   Data Cleaning is the stage where missing values are addressed, unnecessary columns are removed, and data imbalances are handled.
3. Data Transformation
   Data Transformation is a preprocessing stage where the data is modified to ensure it is machine-readable.
4. Data Splitting
   Split Data splitting involves dividing the dataset into two parts: training data and testing data.
5. Feature scalling
   Feature scalling is the process of normalizing or standardizing the values of feature variables.
6. Modelling
   The modeling stage is used to create a model that processes the data.
7. Evaluation
   Evaluation is the stage used to measure the quality and performance of the model.
8. Deployment
   This stage involves generating a report on the analysis results, including insights into data patterns.

RESULTS AND DISCUSSION

Naïve Bayes Manual Calculation

Naïve Bayes is a classification method based on probability and statistics proposed by the British scientist Thomas Bayes, which predicts future probabilities based on past experiences. Naïve Bayes operates on the simplifying assumption that the values of attributes are conditionally independent given the output value. The advantage of using manual Naïve Bayes calculations is that this method requires only a small amount of training data to estimate the parameters needed in the classification process. The following are the steps carried out to make predictions using Naïve Bayes in this study:

1. **Data Collection**

   In the data collection stage, 322 data points were obtained, which were then divided into two parts: 151 data points for training (comprising 90 for Port and Shipping Management and 61 for Nautical) and 171 data points for testing (comprising 98 for Port and Shipping Management and 73 for Nautical). This study applies the Naïve Bayes Algorithm with six attributes: gender, registration pathway, registration wave, religion, district, and province. There are two possible outcomes: Port and Shipping Management and Nautical.

2. **Calculating Class Probability**

   Based on the predetermined training data, the following are the probability values from the training data. There are two classes: Port and Shipping Management and Nautical.

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

TABLE 1. Class Probability

CLASS PROBABILITY	
CLASS	VALUE
Port and Shipping Management	0,6
Nautical	0,4

a. Probability **"PSM"**

$$\frac{number\ of\ SPM}{number\ of\ data} = \frac{90}{151} = 0{,}6$$

b. Probability **"Nautical"**

$$\frac{number\ of\ nautical}{number\ of\ data} = \frac{61}{151} = 0{,}4$$

3. **Calculating the Probability of Each Attribute**

   For binomial/discrete attributes, there are six attributes: Gender, Registration Pathway, Registration Wave, Religion, District, and Province. In each step of calculating the probability for each attribute, the denominator is the class weight, with 90 for the D3 Port and Shipping Management class and 61 for the D3 Nautical class.

   The following are the probability calculations for each attribute, using the class weight denominators for the D3 Nautical and D3 Port and Shipping Management programs.

a) Results of Calculation of Gender Attribute Probability

TABLE 2. Gender Probability

Entry Route	Total	Port and Shipping Management	Nautical
SNBP	19	0,2	0,2
	10		
SNBT	20	0,2	0,3
	20		
Mandiri	51	0,6	0,5
	29		
Mandiri-CBT	0	0,0	0,0
	2		

b) Results of Calculation of Probability of Entry Route Attributes

TABLE 3. Probability of Entry Route

Entry Route	Total	Port and Shipping Management	Nautical
SNBP	19	0,2	0,2
	10		
SNBT	20	0,2	0,3
	20		
Mandiri	51	0,6	0,5
	29		
Mandiri-CBT	0	0,0	0,0
	2		

c) Results of Registration Wave Attribute Probability Calculations

TABLE 4. Registration Wave Probability

Registration	Total	Port and Shipping Management	Nautical
Registration 1	74	0,8	0,8
	51		
Registration 2	16	0,2	0,2
	10		
Registration 3	0	0,0	0,0
	0		

d) Results of Calculation of Religious Attribute Probability

TABLE 5. Religious Attribute Probability

Religion	Total	Port and Shipping Management	Nautical
Islam	69	0,8	0,8
	46		
Kristen	19	0,2	0,2
	12		
Katolik	2	0,0	0,0
	3		

e) Results of Calculation of Regency Attribute Probability

TABLE 6. Regency Probability

Regency	Total	Port and Shipping Management	Nautical
Bengkalis	41 24	0,5	0,4
Siak	2 5	0,0	0,1
Kepulauan Mentawai	11 10	0,1	0,2
Sidikalang	0 1	0,0	0,0
Toba	0 1	0,0	0,0
Kepulauan Meranti	6 9	0,1	0,1
Karo	3 1	0,0	0,0
Simalungun	1 1	0,0	0,0
Pakpak Bharat	1 0	0,0	0,0
Tebing Tinggi	1 1	0,0	0,0
Pekanbaru	2 1	0,0	0,0
Rokan Hilir	3 0	0,0	0,0
Labuhan Batu Selatan	1 0	0,0	0,0
Tapanuli Selatan	1 0	0,0	0,0
Dairi	2 0	0,0	0,0
Rokan Hulu	0 0	0,0	0,0
Dumai	11 1	0,1	0,0
Semarang	1 1	0,0	0,0
Tapanuli Utara	1 0	0,0	0,0
Purwakarta	0 0	0,0	0,0
Padang Pariaman	0 0	0,0	0,0
Batam	1	0,0	0,0

	0		
Tapanuli	1	0,0	0,0
	0		
Lahat	0	0,0	0,0
	0		
Karimun	0	0,0	0,0
	0		
Agam	0	0,0	0,0
	0		
Kampar	0	0,0	0,0
	0		
Batu Bara	0	0,0	0,0
	0		
Medan	0	0,0	0,0
	0		
Solok	0	0,0	0,0
	0		
Palembang	0	0,0	0,0
	0		
Humbang Hasundutan	0	0,0	0,0
	0		

f) Results of Provincial Attribute Probability Calculations

TABLE 7. Province Probability

Province	Total	Port and Shipping Management	Nautical
Riau	65	0,7	0,7
	41		
Sumetera Barat	11	0,1	0,2
	11		
Sumatera Utara	12	0,1	0,1
	5		
Jawa Tengah	1	0,0	0,0
	1		
Jawa Barat	0	0,0	0,0
	0		
Kepulauan Riau	1	0,0	0,0
	1		
Sumatera Selatan	0	0,0	0,0
	1		

4. Comparing the Probability Values of Each Class

At this stage, multiplication is performed for each class to be classified. Once the value of each class is determined, the data will be compared, and the class with the highest value will be the predicted result.

TABLE 8. Comparison of Values of Each Class

Name	ZASQIA SAVIANA
Gender	Female
Entry Route	Mandiri
Registration	Registration 2
Religion	Islam
Regency	Bengkalis
Province	Riau
Class	?

To test the data that has been obtained, a data testing process is needed to determine the prediction. Here is a manual calculation of the values that have been obtained.

a. Shipping Management Class
$0{,}3 \times 0{,}6 \times 0{,}2 \times 0{,}8 \times 0{,}5 \times 0{,}7 \times 0{,}6 = 0{,}006048$
b. Nautical Class
$0{,}1 \times 0{,}5 \times 0{,}2 \times 0{,}8 \times 0{,}4 \times 0{,}7 \times 0{,}4 = 0{,}000896$

5. Determining the Final Probability Value for Each Class (Label)

At this stage, the final results of each class are compared. The calculation shows that the data falls into the "Port and Shipping Management" category, as the value for Port and Shipping Management is the highest when compared to the other class.

6. Calculating accuracy

After processing the training data, the accuracy of the training process is obtained. To calculate accuracy, a Confusion Matrix is used with the provided testing data. Accuracy measurement is done by comparing the predicted training data results, based on the predetermined attributes, with the actual data. The following is the resulting Confusion Matrix.

TABLE 9. *Confusion Matrix*

Confusion Matrix		
	Prediction	
Actual	Port and Shipping Management	Nautical
Shipping Management	75	23
Nautical	55	18

True Positive (TP) refers to the number of positive records in the dataset that are classified as positive. True Negative (TN) refers to the number of negative records in the dataset that are classified as negative. False Positive (FP) refers to the number of negative records in the dataset that are incorrectly classified as positive. False Negative (FN) refers to the number of positive records in the dataset that are incorrectly classified as negative.

TABLE 10. *Performance Vector*

Performance Vector		
Formula	**Variabel**	**Result**
TP + TN / Jumlah Data	Akurasi	54%
TP / (TP + FP)	Presisi	77%
TP / (TP + FN)	Recall	58%

Thus, the calculations for accuracy, precision, and recall can be performed as follows:

a. Calculating Accuracy

Accuracy is the ratio of the number of correct predictions to the total number of data points. To calculate accuracy using the Confusion Matrix, the sum of True Positives and True Negatives is divided by the total number of data points. The calculation is as follows:

$$Accuracy = \frac{75 + 18}{171} = \frac{93}{171} = 0{,}54 \times 100\% = 54\%$$

b. Calculating Precision

Precision measures the proportion of positive class data that has been correctly predicted out of all predicted positive results. The precision value is obtained by dividing True Positive by the sum of False Positive and True Positive. The calculation is as follows:

$$Precision = \frac{75}{75 + 23} = \frac{75}{98} = 0{,}77 \times 100\% = 77\%$$

c. Calculating Recall

Recall indicates the percentage of positive class data that has been correctly predicted out of the total positive class data. The recall value is calculated by dividing True Positive by the sum of True Positive and False Negative. The calculation is as follows:

$$Recall = \frac{75}{75 + 55} = \frac{75}{130} = 0{,}58 \times 100\% = 58\%$$

CONCLUSIONS

1. The research and testing conducted conclude that the Naive Bayes algorithm can predict the interest in the D3 Nautical and D3 Port and Shipping Management study programs for new student admissions in the Maritime Department of Politeknik Negeri Bengkalis, achieving an accuracy level of 54%.
2. The implementation of the Naive Bayes algorithm, which has undergone training and testing, yielded an accuracy of 54%, precision of 77%, and recall of 58%.
3. The factors influencing students' interest in the D3 Nautical and D3 Port and Shipping Management study programs in the Maritime Department of Politeknik Negeri Bengkalis are primarily dominated by the choice of location or the proximity of the campus to the students' original schools or residences, particularly for high school students in the Bengkalis District. Additionally, the interest in the Maritime program is also predominantly from students in the Riau Province.

ACKNOWLEDGMENTS

1. This study was supported and for financial support of this research under Contract No: 79/KT-PN/P3M-PB/2024, by a Research and Community Service the State Polytechnic of Bengkalis, Indonesia
2. The authors would like to thank you so much for my super team Prediction Of Interest In The D3 Nautical And D3 Port And Shipping Management Study Program For New Student Admissions At State Polytechnic Of Bengkalis Using Naïve Bayes
3. The authors would like to thank the lecturers and supervisors cadets of the Maritime Department and Informatics Engineering of the State Polytechnic of Bengkalis

REFERENCES

[1] Ary, M. (2014). Identifikasi Perilaku Calon Mahasiswa Baru Dalam Memilih Program Studi Menggunakan Analisis Faktor. *Paradigma Jurnal Komputer dan Informatika Akademi Bina Sarana Informatika*, 35-45.

[2] Iranita. (2012). Analisis Faktor Yang Mempengaruhi Pilihan Mahasiswa Fakultas Ekonomi Universitas Maritim Raja Ali Haji Dalam Menciptakan Keunggulan Kompetitif (Competitif Advantage). *JEMI*, 77-88.

[3] Lianda, D., & Atmaja, N. S. (2021). Prediksi Data Buku Favorit Menggunakan Metode Naïve Bayes (Studi Kasus: Universitas Dehasen Bengkulu). *Pseudocode*, *8*(1), 27– 37.https://doi.org/10.33369/pseudocode.8.1.27-37

[4] Mafakhir, A. Z., & Solichin, A. (2020). Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta. *Fountain of Informatics Journal*, *5*(1), 21. https://doi.org/10.21111/fij.v5i1.4007

[5] Suryadi, A., & Harahap, E. (2018). Sistem Rekomendasi Penerimaan Mahasiswa Baru Menggunakan Naive Bayes Classifier Di Institut Pendidikan Indonesia. *Joutica*, *3*(2),171. https://doi.org/10.30736/jti.v3i2.231