BAREKENG: Journal of Mathematics and Its Applications
September 2024 Volume 18 Issue 3 Page 1433–1446
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss3pp1433-1446

PREDICTION OF PROSPECTIVE NEW STUDENTS USING DECISION TREE, RANDOM FOREST, AND NAIVE BAYES

Yulrio Brianorman [1*], Sucipto [2]

[1,2]*Informatics Department, Faculty of Engineering, Universitas Muhammadiyah Pontianak*
Jln. Ahmad Yani, Pontianak, 78123, Indonesia

*Corresponding author's e-mail: * **y.brianorman@unmuhpnk.ac.id***

Article History:

Received: 20th Oct 2023
Revised: 12th Dec 2023
Accepted: 10th June 2024
Published: 1st September 2024

Keywords:

Data Classification Model;
New Student Enrollment;
Decision Tree;
Random Forest;
Naïve-Bayes.

ABSTRACT

Higher education positions new student enrollment as a strategic activity for private universities. The effectiveness of selecting prospective students with a high potential to register and be accepted is crucial. Therefore, this study was conducted to find a data classification model that can determine the potential acceptance of new students, allowing private universities to increase the number of students admitted. This research's data originated from the 2020 new student admissions at a prominent private university in Pontianak city. The three chosen classification methods are the decision tree, random forest, and naïve-bayes. Evaluation results indicate the accuracy rate of the decision tree is 59.1%, random forest at 59.2%, and naïve-Bayes at 58.1%. Despite similar accuracy rates, the random forest method slightly outperformed the others, suggesting it may be the most reliable for predicting student enrollment. Based on these models, the estimated potential of prospective students registering at the university ranges from 72% to 78% of the total student candidates. In conclusion, although the three models have almost similar accuracy rates, all show an optimistic estimate regarding the registration potential of prospective students. Thus, universities can use one or a combination of the three models to enhance efficiency in the student admission process.

How to cite this article:

Y. Brianorman and Sucipto., "PREDICTION OF PROSPECTIVE NEW STUDENTS USING DECISION TREE, RANDOM FOREST, AND NAIVE BAYES," *BAREKENG: J. Math. & App.*, vol. 18, iss. 3, pp. 1433-1446, September, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The growth and reputation of higher education institutions are significantly influenced by the efficacy of their New Student Admissions processes. These processes are not merely procedural requirements but integral to shaping the university's intellectual diversity and demographic composition, thereby impacting its long-term viability and financial well-being [1]. Achieving enrollment targets and fulfilling the set quotas are crucial for effective resource management and economic stability.

At this study's focal point, which is higher education institution, the New Student Admission process is meticulously organized into two distinct phases: the initial online registration and the subsequent phase where registrants must complete their enrollment to become students officially. However, the transition between these stages is marked by significant drop-off rates. For instance, in 2020, out of 1,553 prospective students from West Kalimantan who registered online, only 786 completed the process—leaving 767 individuals who did not finalize their registration. This considerable attrition rate presents a critical challenge in optimizing admission strategies and resource allocation. Marketing communication is a crucial aspect of marketing missions, significantly influencing the success of marketing efforts [2].

The current follow-up procedures involve administrative staff from each study program contacting prospective students to encourage completion of their registration. Despite these initiatives, the current strategy remains predominantly reactive and somewhat unstructured, lacking a data-driven approach to identify and prioritize individuals who are most likely to complete their registration. This deficiency hinders the ability to effectively target marketing and engagement efforts toward the most promising candidates.

This research addresses this gap by proposing a classification approach to predict the potential of prospective students to proceed with the registration process. By implementing and comparing three different classification methods—decision tree, random forest, and naïve Bayes—this study aims to equip administrative staff with a systematic tool to target and engage prospective students who are most likely to enroll more effectively. Through such targeted interventions, the institution can enhance the efficiency of its admissions process, potentially increasing overall student intake and optimizing resource use. Random forest is a powerful ensemble learning technique celebrated for its heightened predictive performance and robustness in handling complex datasets; however, it is criticized for its computational expense and complexity in interpretability [3].

Several studies related to classification and prediction in the context of education have been conducted. Research has been done on the classification of students who took remedial exams [4], student specializations based on the grades from each course they undertook [5], predicting timely student graduation [6], and the prediction of assistance recipients [7]. Additionally, studies have also classified student academic achievements [8].

Furthermore, matters related to singular tuition fees have been classified [9], student admissions at Pamulang University have been reviewed [10], and research has been conducted on the classification determining high school majors [11]. Studies on student academic performance [12], the use of the Naïve Bayes method to determine faculty choices for new students at AMIKOM University, Yogyakarta [13], and predictions related to student graduation [14] have also been undertaken.

Outside of the educational context, research has been conducted on the classification of areas at high risk for the spread of COVID-19 in Indonesia [15], and on classifying families based on their eligibility to receive Village Cash Assistance (BLT DD) [16]. Additionally, research related to the classification of prospective new students has been conducted [17]. They found that the Random Forest algorithm achieved the highest accuracy at 73.61%, followed by k-nearest neighbor at 72.08%, and naïve Bayes at 70.47%. They concluded that Random Forest was particularly effective due to its robustness in handling diverse data types and complex decision boundaries. These findings are particularly relevant as they highlight the effectiveness of Random Forest in similar application contexts, suggesting potential benefits for its use in our study focused on improving predictive models for student enrollment.

Several studies have been conducted on the comparison of classification methods between decision tree, random forest, and Naïve Bayes. These methods have been applied to assess landslide vulnerability in the Longhai region, in China [18], and examined in the context of sentiment analysis [19]. Research has also focused on the performance comparison of classification on a student performance dataset [20], and on determining the most efficient and reliable model to assess flood vulnerability in the Quannan region, China

[21]. Additionally, public sentiment trends on Twitter regarding the Anti-LGBT campaign in Indonesia have been analyzed using a sentiment analysis approach [22].

Studies have also been conducted on soil texture classification in the Kalikonto River Basin using Random Forest and Naïve Bayes machine learning algorithms, with the former achieving higher accuracy based on the Digital Elevation Model [23]. Additionally, Twitter sentiments on climate change from January to June 2022 have been analyzed using the Random Forest and Naïve Bayes methods [24].

Based on a comprehensive review of previous studies and a detailed evaluation of each method's performance in the context of our data, this research has chosen to use the classification methods of decision tree, random forest, and naïve bayes. The decision tree method was selected for its straightforwardness in comprehension and visualization, making it an optimal choice for elucidating the factors that influence student enrollment decisions. However, it is susceptible to overfitting, particularly when applied to complex datasets. The random forest approach was chosen for its exceptional accuracy and efficiency, which prove advantageous when managing large datasets with numerous input variables. Nonetheless, due to its ensemble nature, this method may exhibit reduced transparency and require substantial computational resources. Despite Naïve Bayes assuming feature independence, which can lead to an oversimplification of real-world data interactions, it was incorporated into this study due to its robustness in managing high-dimensional and noisy datasets, thereby ensuring reliable performance amidst the inherent variability characteristic of educational data. These methods are favored in various applications and research contexts due to their distinct advantages and limitations. Employing a quantitative methodology, this study utilizes data from the New Student Admissions of the higher education institution, specifically the 2020 cohort, to assess the efficacy of each method.

2. RESEARCH METHODS

The CRISP-DM method has become a commonly used standard in Data Mining research. CRISP-DM, an acronym for Cross-Industry Standard Process for Data Mining, outlines the essential stages in the data mining process. This initiative was launched by three major companies: Daimler Chrysler (previously Daimler-Benz), SPSS (ISL), and NCR. From 1997 to 1999, CRISP-DM was continuously developed through various workshops, with contributions from over 300 organizations. In 1999, the CRISP-DM 1.0 version was officially released, as described by Wirth & Hipp (2000). The CRISP-DM method consists of six stages, namely Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment, as depicted in Figure 1.

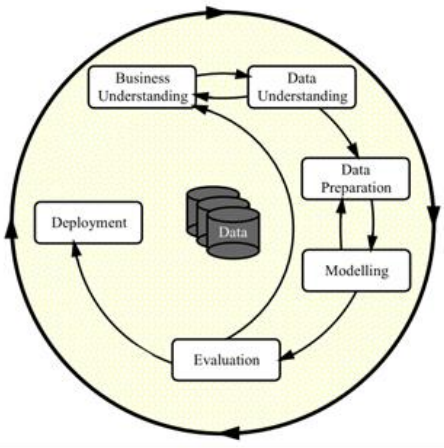

Figure 1. CRISP-DM Phase

Referring to the CRISP-DM method, a research framework has been formulated. This framework illustrates the flow and stages undertaken in the research process. Figure 2 depicts the framework employed in this study.

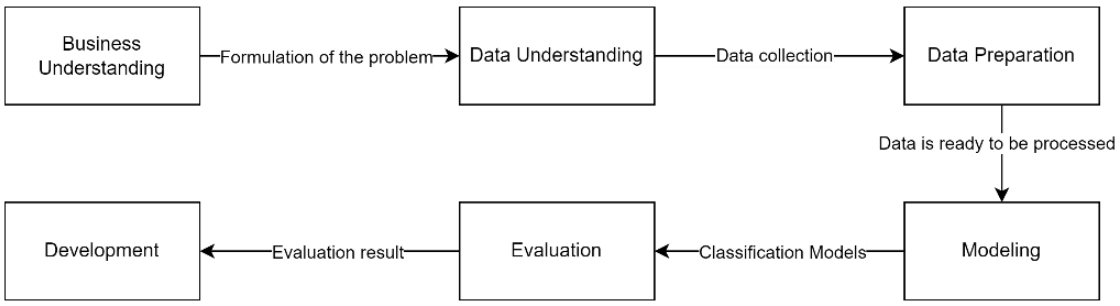

Figure 2. Research Framework

The Data Preparation stage involves three primary steps: data reduction, data cleaning, and data transformation. The data reduction process is designed to enhance the analytical efficiency by selectively eliminating columns that contain either unique identifiers, which typically do not contribute to predictive modeling, or repetitive data, which adds unnecessary redundancy to the dataset. This step is crucial as it not only streamlines the dataset, making it more manageable for analysis but also helps in focusing the model's learning on relevant features that are influential in predicting outcomes. Data cleaning is a vital process that involves the meticulous removal or strategic replacement of missing data with specific values, ensuring the dataset's integrity and consistency. This step may include techniques such as imputation, where missing values are substituted based on the median, mean, or mode of the respective attribute, or using more sophisticated predictive models to estimate missing data accurately. The aim is to create a complete and reliable dataset that accurately reflects the underlying patterns without being skewed by gaps in the data. Meanwhile, data transformation modifies attribute values based on predefined criteria to ensure compatibility with analytical models. This process may include scaling numerical data to prevent disproportionate influence, normalizing to achieve uniform scales, and encoding categorical variables into numerical formats. These steps enhance model accuracy and ensure data adherence to the necessary analytical assumptions.

Three renowned classification techniques (decision trees, random forests, and naïve Bayes) were employed in this study to evaluate their efficacy in predicting the enrollment of new students. Each method was chosen for its unique ability to handle the complexities inherent in educational data. A decision tree is a hierarchical tree structure akin to a flowchart, where each internal node denotes a feature or attribute, each branch represents a decision rule, and each leaf node signifies an outcome. The root node, situated at the apex of the decision tree, initiates the partitioning of a dataset into progressively smaller subsets, while concurrently constructing the decision tree incrementally. The terminal node holds the class label or outcome, and the decision node delineates the choices among multiple alternatives. Owing to their clarity and ease of visualization, decision trees are exceptionally effective for data analysis and facilitating informed decision-making. Random Forest is an ensemble learning method employed for classification, regression, and other analytical tasks. This technique constructs a multitude of decision trees during the training phase and outputs either the mean prediction for regression or the mode of the classes for classification, derived from the individual trees. By mitigating the tendency of decision trees to overfit their training set, Random Forest generates a model with enhanced general applicability. Due to its ensemble nature, which aggregates the predictions of several trees, this method is noted for its robustness and reliability in producing precise predictions. Naïve Bayes is a probabilistic classifier that applies Bayes' theorem with the stringent assumption of independence among predictors. It is particularly effective for high-dimensional data. Despite its simplicity, Naïve Bayes can surpass more complex classification methods. This classifier demonstrates robust performance when the high dimensionality of the input data leads to sparsity, a common scenario in large databases.

The evaluation stage is a critical step in ensuring the effectiveness of the developed model. In this context, the performance of the learning process is measured using a series of performance metrics, namely precision, recall, accuracy, and $F1$ score. Precision assesses the extent to which positive classification results identified by the model are genuinely positive. Conversely, recall provides insight into how much of the actual positive data has been successfully recognized by the model. Accuracy, on the other hand, offers an overall perspective on how accurately the model classifies all the data. Meanwhile, the $F1$ score amalgamates precision and recall providing a harmonized measure of both. **Equation (1)** to **Equation (4)** respectively present the formulas for these metrics, allowing researchers to delve into them in more detail **[25]**.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1\ \text{score} = \frac{\text{Precision } x \text{ Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The explanation regarding the variables TP, FP, FN, and TN is illustrated by the confusion matrix, as depicted in **Figure 3**.

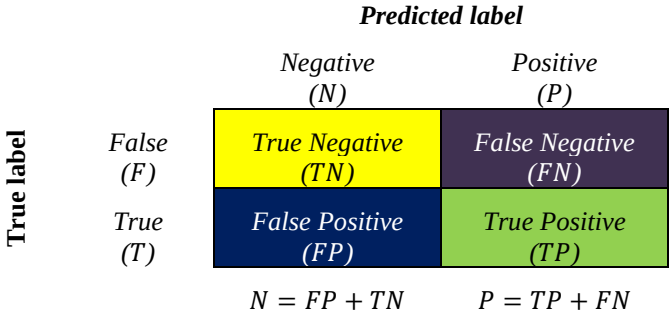

Figure 3. Confusion Matrix

3. RESULTS AND DISCUSSION

The modeling was implemented using the sklearn library in Python through the Google Colab platform. With the computational infrastructure support from Google Colab, the modeling process can run more efficiently and stably. This section will discuss the results obtained from the classification methods of the decision tree, random forest, and naïve bayes. This analysis is crucial to determine which method provides the most accurate prediction based on the available dataset.

3.1 Research Result

3.1.1 Business Understanding

In the initial phase, an interview process was required as a mechanism to delve deeper and identify ongoing issues. From the interview sessions, it was revealed that the academic program administrative staff did not implement a specific methodology when interacting with prospective students; instead, they adopted a random approach. This unsystematic approach has the potential to decrease the effectiveness of the selection process and increase uncertainties in admission outcomes. To address this issue and enhance the quality of selection, it is recommended to integrate a classification method for prospective students. Through this classification approach, prospective students can be grouped based on certain parameters, allowing the staff to engage in more structured interactions and improve efficiency in the selection process.

3.1.2 Data Understanding

The second phase involved data collection which was subsequently analyzed to delve deeper and identify specific characteristics. The procedures undertaken in this stage include:

1. The data collection process was carried out by extracting data from the NSA application. The data obtained represents prospective students from the year 2020 of a private university in Pontianak. The accumulated data can be seen in **Table 1**. The data obtained consists of 1,892 rows encompassing 20 attributes. These attributes include PMB Semester Year, PMB Reg Code, Registration Date, Registration Month, Registration Year, Full Name, Gender, Date of Birth, Age (in years), Religion, Father's Income Range, Mother's Income Range, Registrant's Province of Domicile, Registrant's Domicile District, Name of School of Origin, School Department, School Province, School District, School Type, and Registration.

Table 1. Raw Data

No	PMB Semester Year	PMB Reg Code	...	Registration
1	20201	PMB2020100001	...	1
2	20201	PMB2020100000	...	1
3	20201	PMB2020100003	...	1
⋮	⋮	⋮	⋱	⋮
1892	20201	PMB2020103165	...	1

2. The subsequent step involves the identification of the collected data and the determination of relevant attributes. The types of attributes present in the data can be seen in **Table 2**. The table indicates the presence of 20 attributes in the data. Out of these, 2 attributes are of integer data type, namely PMB Semester Year and Age (in years). Seventeen attributes are of polynomial data type, which includes PMB Reg Code, Registration Date, Registration Month, Registration Year, Full Name, Gender, Date of Birth, Age (in years), Religion, Father's Income Range, Mother's Income Range, Registrant's Province of Domicile, Registrant's Domicile District, Name of School of Origin, School Department, School Province, and School District. Meanwhile, School Type is an attribute with a binomial data type.

Table 2. Data Attribute Type

No	Attribute	Data Type	Parameter
1	PMB Semester Year	Integer	-
2	PMB Reg Code	Polynomial	-
⋮	⋮	⋮	⋮
11	Father's Income Range	Polynomial	> 5,000,000; 3,000,001 – 5,000,000; 2,000,001 – 3,000,000; 1,500,001 – 2,000,000; 1,000,001 – 1,500,000; 500,001 – 1,000,000; 500,001 – 1,000,000; < 500,000
⋮	⋮	⋮	⋮
20	Registration	Binomial	0, 1

3.1.3 Data Preparation

The data preparation stage requires a series of critical steps to ensure the data to be analyzed is of high quality. The following steps were undertaken:

1. Data reduction is an essential step to determine which attributes are relevant for the analysis. Two primary factors that might cause an attribute to be excluded from this analysis are unique data and imbalanced data. Attributes such as PMB Reg Code, Registration Date, Registration Year, Full Name, Date of Birth, and Name of School of Origin are excluded due to their unique values. On the other hand, attributes like PMB Semester Year, Religion, Province of Domicile, School Department, School Province, and School District are overlooked due to data imbalance. Features with imbalanced distribution are expected not to provide sufficient information to aid in classification. Therefore, the attributes focused on in this study include Registration Month, Gender, Father's Income Range, Mother's Income Range, Registrant's Domicile District, and School Type.

2. Data cleaning involves steps to check and correct missing, invalid, or duplicate data. Attributes with missing values will be filled based on the three most frequently occurring attribute values, chosen at random. Meanwhile, duplicate data identified will be eliminated.
3. Data transformation involves modifications to certain attribute parameters. The detailed attribute mapping can be seen in **Table 3**.

Table 3. Attribute Data Transformation Mapping

No	Attribute	Before	After
1.	Registration Month	January	1
		February	2
		March	3
		April	4
		May	5
		June	6
		July	7
		August	8
		September	9
		October	10
		November	11
		December	12
2.	Gender	P	1
		L	0
3.	Father's Income Range	> 5,000,000	1
		3,000,001 – 5,000,000	2
		2,000,001 – 3,000,000	3
		1,500,001 – 2,000,000	4
		1,000,001 – 1,500,000	5
		500,001 – 1,000,000	6
		< 500,000	7
4.	Mother's Income Range	> 5,000,000	1
		3,000,001 – 5,000,000	2
		2,000,001 – 3,000,000	3
		1,500,001 – 2,000,000	4
		1,000,001 – 1,500,000	5
		500,001 – 1,000,000	6
		< 500,000	7
5.	Registrant's Domicile District	*Kota Pontianak*	1
		Kabupaten Kubu Raya	2
		Kabupaten Ketapang	3
		Kabupaten Landak	4
		Kabupaten Kayong Utara	5
		Kabupaten Bengkayang	6
		Kabupaten Sekadau	7
		Kabupaten Sanggau	8
		Kabupaten Kapuas Hulu	9
		Kabupaten Sintang	10
		Kabupaten Sambas	11
		Kabupaten Melawi	12
		Kabupaten Mempawah	13
		Kota Singkawang	14
6.	School Type	*Swasta*	0
		Negeri	1

Subsequently, there is a process of renaming attributes. **Table 4** illustrates the renaming of attributes in the data set, a process aimed at simplifying attribute names for clarity and ease of use in analysis. "Registration Month" is shortened to "Month," streamlining the attribute while preserving its time-related significance. Similarly, "Father's Income Range" and "Mother's Income Range" are condensed to "Father" and "Mother," respectively, focusing solely on the income aspect essential for demographic analysis. "School District" is renamed "Origin," broadening its interpretive use to perhaps emphasize geographical background more generally. Lastly, "School Type" becomes "School," simplifying the

attribute while maintaining its relevance to the type of educational institution. These changes facilitate a more efficient data handling and analysis process, reducing complexity and enhancing the clarity of the dataset for statistical modeling.

Table 4. Attribute Rename Mapping

No	Attribute	
	Before	After
1.	Registration Month	Month
2.	Father's Income Range	Father
3.	Mother's Income Range	Mother
4.	School District	Origin
5.	School Type	School

Table 3 and **Table 5** together illustrate the comprehensive data transformation process applied to the dataset used in this study. **Table 3** details the specific mappings used to convert various categorical and text-based attributes into numerical codes, which are necessary for statistical analysis and machine learning models. The results of the data preparation phase can be seen in **Table 5**.

Table 5. Data Preparation Results

Index	Month	Gender	Age	Father	Mother	Origin	School	Registration
0	2	0	18.0	1.0	1.0	1	0.0	1
1	2	0	20.0	4.0	7.0	1	1.0	1
2	2	0	18.0	4.0	7.0	2	1.0	0
3	2	0	18.0	2.0	7.0	3	1.0	0
4	2	1	19.0	5.0	6.0	4	1.0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.1.4 Modelling

This phase involves the modeling process using the decision tree, random forest, and naïve Bayes methods. Utilizing the sklearn library in the Python programming language and the Google Colab IDE, this process consists of three steps: examining the dependency between feature attributes and the target attribute, dividing the dataset into training and test data, and executing the modeling.

The examination of the dependency between the feature attributes and the target attribute is conducted using the chi-square test. The choice of the chi-square test is grounded in the fact that both tested attributes are categorical. **Figure 5** displays the results of the chi-square test calculations. A higher chi-square test value indicates a stronger dependency between the feature and target attributes. Based on this, the attributes selected for the modeling process include 'Origin', 'Month', dan 'Father'.

The dataset is divided into two parts: training data and test data. The “train_test_split()” function from the sklearn library is employed in this partitioning process. With a proportion of 75% for training data and 25% for test data, the result is 1,166 rows for training data and 389 rows for test data.

Modeling is implemented by leveraging the decision tree, random forest, and naïve Bayes libraries from sklearn. The time taken for the modeling is relatively short. The results of this modeling will be used in a system to classify prospective student data.

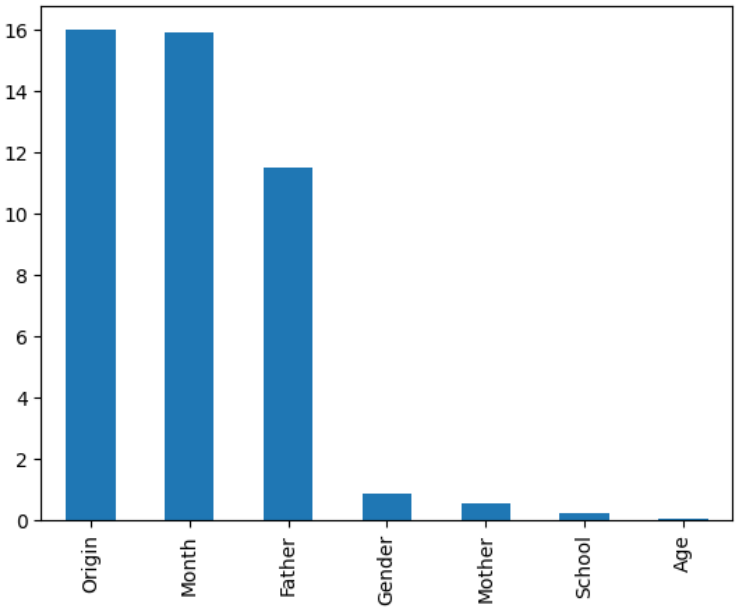

Figure 4. Chi-Square Test Calculation Results

3.1.5 Evaluation

The evaluation of the model is measured based on the confusion matrix. Using the decision tree model, of the 389 predictions, 213 were correctly classified (112 as True Negatives and 101 as True Positives), while 176 were misclassified (105 as False Positives and 71 as False Negatives). This indicates a balance in correctly identifying both positive and negative classes, but a notable number of both types of errors are present. The matrix can be seen in **Figure 6**.

Figure 7 for random forest model, out of the 389 predictions, 223 were correctly classified (95 as True Negatives and 128 as True Positives), while 166 were misclassified (78 as False Positives and 88 as False Negatives). This suggests that the model was better at identifying positive cases than negative ones, but there's still a significant number of misclassifications for both classes.

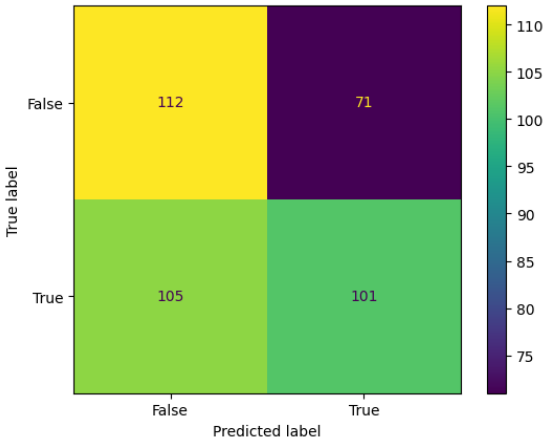

Figure 5. Confusion matrix decision tree

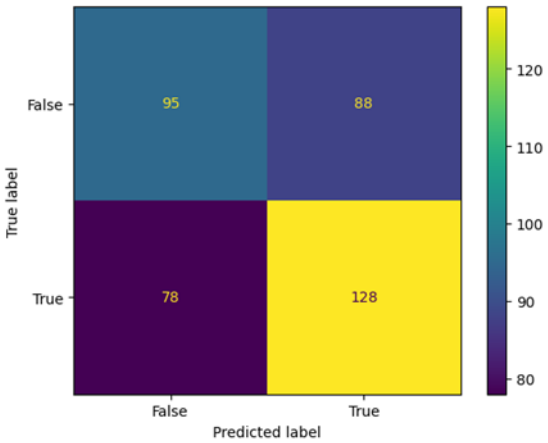

Figure 6. Confusion matrix random forest

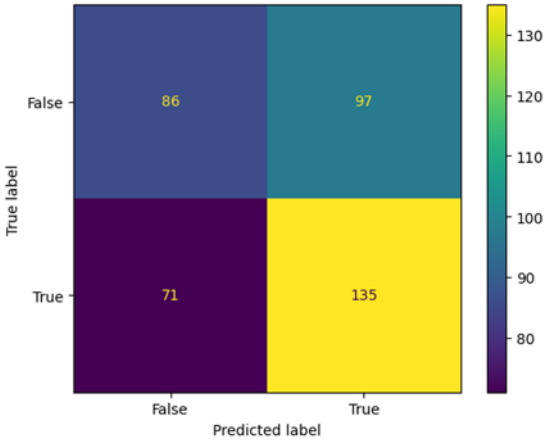

Figure 7. Confusion matrix naïve bayes

Figure 8 for the Naïve Bayes model, out of the 389 predictions, 221 were correctly classified (86 as True Negatives and 135 as True Positives), while 168 were misclassified (71 as False Positives and 97 as False Negatives). This indicates that the model was more adept at identifying positive cases than negative ones, but there remains a notable number of inaccuracies in the predictions for both classes.

The calculations for precision, recall, accuracy, and *F*1 score can be found in Table 6. It shows that Naïve Bayes exhibits the highest precision at 0.655, making it highly reliable for applications where minimizing false positives is crucial. Random Forest, however, demonstrates superior recall at 0.573 and the highest overall accuracy at 0.592, suggesting it effectively identifies relevant instances and maintains a balance between true positive and negative predictions. Furthermore, Random Forest achieves the highest *F*1 score at 0.574, indicating its balanced capability in both precision and recall. Meanwhile, the Decision Tree, with an accuracy of 0.591, though slightly lower, offers a simpler model that may be advantageous in contexts where interpretability and computational efficiency are prioritized.

Table 6. Measurement results for each method

Measurement	Decision Tree	Random Forest	Naïve Bayes
Precision	0.500	0.621	0.655
Recall	0.552	0.573	0.568
Accuracy	0.591	0.592	0.581
***F1* score**	0.553	0.574	0.571

3.1.6 Deployment

Deployment has been successfully implemented in the new student admission system. In this context, "deployment" refers to the application of a developed model or algorithm into a production system for operational purposes or real-world decision-making. With this implementation, the system is now capable of

making predictions on prospective students using the created API. The modeling outcome provides an output of either 1 or 0, where a value of 1 indicates a prediction that the prospective student will register, while a value of 0 indicates otherwise. Thus, educational institutions can enhance the efficiency of the admission process by gaining early insights into the potential behavior of prospective students.

3.2 Discussion of Result

This study produced models based on three applied methods. Although the measurement results are not entirely optimal, the calculations from these three methods have provided substantive guidance in targeting prospective students who have not yet registered.

Based on accuracy measurements, the random forest method showed superior results compared to the decision tree and naïve Bayes. However, the accuracy achieved by the decision tree almost approached the value attained by the random forest. Meanwhile, the naïve Bayes method displayed quite satisfactory results based on the four measurements conducted.

From the confusion matrix, these results can be interpreted to understand the registration potential of prospective students. To estimate this potential, **Equation (5)** is used. The rationale behind the equation is the inclusion of False Positives and False Negatives as indicators of registration potential. The potential calculation results based on the decision tree, random forest, and naïve Bayes methods are 72.4%, 75.5%, and 77.8% respectively. From these results, it can be estimated that the registration potential of prospective students in that educational institution ranges from 72% to 78% of all registrants.

$$\text{Potency} = \frac{TP + FP + FN}{TP + TN + FP + FN} \tag{5}$$

4. CONCLUSIONS

Based on the findings of this study, the Random Forest method exhibited the highest accuracy at 59.2%, closely followed by the Decision Tree with an accuracy of 59.1%, and Naïve Bayes at 58.1%. This slight edge in performance suggests that the Random Forest method is particularly effective in the context of new student admissions systems at educational institutions, making it the preferred choice for integrating into such systems. The classification of prospective students' potential to continue the registration process was effectively conducted using these methods, with Random Forest and Decision Tree emerging as the most reliable for predicting student behavior.

The study's analysis revealed that the registration potential of prospective students—classified by their likelihood to complete the registration process—ranges between 72% to 78% of all registrants. This quantification was facilitated by the Random Forest method, which not only provided the highest accuracy but also the most nuanced insight into the registration behaviors of students. Thus, it can be concluded that the Random Forest method is ideally suited for predicting which students are most likely to proceed from initial interest to formal registration, thereby aiding institutions in targeting interventions and resources more effectively to maximize conversion rates.

To enhance the accuracy of the model, future research could focus on exploring a variety of feature extraction methods and incorporating datasets from more recent years. The adoption of feature fusion techniques could significantly improve data representation. Employing diverse methodologies, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), would broaden the analytical scope. Integrating these approaches would augment the models' understanding and predictive capabilities.

ACKNOWLEDGMENT

We would like to thank the University of Muhammadiyah Pontianak for funding this research.

REFERENCES

[1] H. Noor Alifa and A. Sunarya Sulaeman, "Perguruan Tinggi Negeri BLU di Indonesia; Pengelolaan Anggaran, Karakteristik, dan Peningkatan Kinerja," *Jurnal Riset Akuntansi dan Keuangan*, vol. 11, no. 2, pp. 401–4016, 2023.

[2] R. Rismiatun, "Efektivitas Strategi Komunikasi Pemasaran Universitas Budi Luhur Dalam Penerimaan Mahasiswa Baru 2019," *KOMUNIKASI: Jurnal Komunikasi*, vol. 11, no. 1, pp. 17–22, Feb. 2020.

[3] Y. Manzali, Y. Akhiat, K. Abdoulaye Barry, E. Akachar, and M. El Far, "Prediction of Student Performance Using Random Forest Combined With Naïve Bayes," *Comput J*, May 2024, doi: 10.1093/comjnl/bxae036.

[4] J. JAYA PURNAMA, H. M. NAWAWI, S. ROSYIDA, RIDWANSYAH, and RISNANDAR, "Klasifikasi Mahasiswa Her Berbasis Algortima SVM Dan Decision Tree," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, no. 6, 2020, doi: 10.25126/jtiik.202073080.

[5] I. P. Pradnyana Iswara, F. Farhan, W. Kumara, and A. Afif Supianto, "Rekomendasi Pengambilan Mata Kuliah Pilihan Untuk Mahasiswa Sistem Informasi Menggunakan Algoritme Decision Tree," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 6, no. 3, pp. 341–348, 2019, doi: 10.25126/jtiik.2019.6892.

[6] NGATMARI, M. BISRI MUSTHAFA, C. RAHMAD, R. ANDRIE ASMARA, and F. AHUTOMO, "Pemanfaatan Data Pddikti Sebagai Pendukung Keputusan Manajemen Perguruan Tinggi," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, no. 3, pp. 555–564, 2020, doi: 10.25126/jtiik.202072585.

[7] R. AKBAR and S. UYUN, "Penentuan Bantuan Siswa Miskin Menggunakan Fuzzy Tsukamoto Dengan Perbandingan Rule Pakar Dan Decision Tree (Studi Kasus: SDN37 Bengkulu Selatan)," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 8, no. 4, pp. 651–662, 2021, doi: 10.25126/jtiik.202183307.

[8] S. PRATAMA, ISWANDI, A. SEVTIAN, and T. PUTRI ANJANI, "Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Menggunakan Algoritma C4.5 dengan CRISP-DM," *Journal of Applied Informatics and Computing (JAIC)*, vol. 7, no. 1, p. 20, 2023, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[9] R. SUSETYOKO, W. YUWONO, E. PURWANTINI, and N. RAMADIJANTI, "Perbandingan Metode Random Forest, Regresi Logistik, Naïve Bayes, dan Multilayer Perceptron Pada Klasifikasi Uang Kuliah Tunggal (UKT)," *Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan*, vol. 7, no. 1, pp. 8–16, 2022.

[10] A. SAIFUDIN, "Metode Data Mining Untuk Seleksi Calon Mahasiswa Pada Penerimaan Mahasiswa Baru Di Universitas," *J Teknol*, vol. 10, no. 1, pp. 25–36, 2018, doi: 10.24853/jurtek.10.1.25-36.

[11] S. MUTROFIN, M. M. MACHFUD, D. HERNYKA SATYARENI, R. V. HARI GINARDI, and C. FATICHAH, "Komparasi Kinerja Algoritma C4.5, Gradient Boosting Trees, Random Forests, Dan Deep Learning Pada Kasus Educational Data Mining," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 4, pp. 807–814, 2020, doi: 10.25126/jtiik.2020732665.

[12] A. RAHMAN, "Klasifikasi Performa Akademik Siswa Menggunakan Metode Decision Tree dan Naive Bayes," *Jurnal SAINTEKOM*, vol. 13, no. 1, pp. 22–31, Mar. 2023, doi: 10.33020/saintekom.v13i1.349.

[13] R. MOHAMAD ANDRI K. RASYID, A. RIYANTO, R. WIDYAWATI, and ISTININGSIH, "Implementasi Algoritma Naïve Bayes Untuk Sistem Rekomendasi Pemilihan Fakultas Di Universitas Amikom Yogyakarta," *JIKOM: Jurnal Informatika dan Komputer*, vol. 13, no. 1, pp. 1–9, 2023.

[14] H. LATIFAH and S. MUJIYONO, "Perbandingan Algoritma Naïve Bayes, K-NN, ID3 Dan SVM Dalam Menentukan Prediksi Kelulusan Siswa Di SMK Muhamadiah Majenang," *JAMASTIKA*, vol. 2, no. 1, pp. 38–45, 2023.

[15] AINURROHMAH and D. TRI WIYANTI, "Analisis Performa Algoritma Decision Tree, Naïve Bayes, K-nearest Neighbor Untuk Klasifikasi Zona Daerah Risiko COVID-19 Di Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 10, no. 1, pp. 115–122, 2023, doi: 10.25126/jtiik.2023105935.

[16] D. KURNIADI, F. NURAENI, M. FIRMANSYAH, and P. Korespondensi, "Klasifikasi Masyarakat Penerima Bantuan Langsung Tunai Dana Desa Menggunakan Naïve Bayes Dan Smote," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, pp. 309–320, 2023, doi: 10.25126/jtiik.2023106453.

[17] P. SEJATI, MUNAWAR, M. PILLIANG, and H. AKBAR, "Studi Komparasi Naive Bayes, K-nearest Neighbor, Dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Mundur," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 7, pp. 1341–1348, 2022, doi: 10.25126/jtiik.202296737.

[18] W. Chen, S. Zhang, R. Li, and H. Shahabi, "Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling," *Science of The Total Environment*, vol. 644, pp. 1006–1018, Dec. 2018, doi: 10.1016/j.scitotenv.2018.06.389.

[19] M. Guia, R. Silva, and J. Bernardino, "Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis," in *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, SCITEPRESS - Science and Technology Publications, 2019, pp. 525–531. doi: 10.5220/0008364105250531.

[20] K. Yadav and R. Thareja, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 12, pp. 11–19, Dec. 2019, doi: 10.5815/ijisa.2019.12.02.

[21] W. Chen *et al.*, "Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods," *Science of The Total Environment*, vol. 701, p. 134979, Jan. 2020, doi: 10.1016/j.scitotenv.2019.134979.

[22] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm," *Procedia Comput Sci*, vol. 161, pp. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.

[23] H. Pramoedyo, D. Ariyanto, and N. N. Aini, "COMPARISON OF RANDOM FOREST AND NAÏVE BAYES METHODS FOR CLASSIFYING AND FORECASTING SOIL TEXTURE IN THE AREA AROUND DAS KALIKONTO, EAST JAVA," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 4, pp. 1411–1422, Dec. 2022, doi: 10.30598/barekengvol16iss4pp1411-1422.

[24] F. Fauzi, W. Setiayani, T. W. Utami, E. Yulianto, and I. W. Harmoko, "COMPARISON OF RANDOM FOREST AND NAÏVE BAYES CLASSIFIER METHODS IN SENTIMENT ANALYSIS ON CLIMATE CHANGE ISSUE," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1439–1448, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1439-1448.

[25] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.