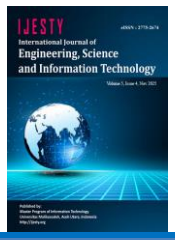
International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 44-54
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.982
Research Paper, Short Communication, Review, Technical Paper



Predictive Analysis of Potential Fraud in the Distribution of The Program Indonesia Pintar (PIP) Funds Using the Naïve Bayes and SVM Methods

Rizki Izandi Gumay*, Sajarwo Anggai, Tukiyat

[1,2,3] Master of Informatics Engineering, Pamulang University, Indonesia

Corresponding author Email: rigi.nazki@gmail.com

The manuscript was received on 20 February 2025, revised on 28 April 2025, and accepted on 21 August 2025, date of publication 2 November 2025

Abstract

The distribution of funds for The Indonesia Smart Program (Program Indonesia Pintar, or PIP), as a national education assistance program, faces serious challenges related to the potential for fraud that can harm the state and hinder the goal of equitable access to education. This study aims to develop a machine learning-based predictive model to detect potential fraud in the distribution of PIP funds by comparing two main algorithms, Naive Bayes and Support Vector Machine (SVM). The dataset used is the result of the integration of PIP and DAPODIK data in 2023, as well as additional features of engineering results based on the pattern of audit findings. All data, through preprocessing, normalization, and balancing processes, uses SMOTE to overcome class imbalances. The model was evaluated using accuracy, precision, recall, and F1-score metrics, both on internal and external test data from Banten Province. The results showed that SVMs with RBF kernel and optimal parameter tuning provided the best performance with an accuracy of up to 98.5% on test data. At the same time, Naive Bayes tended to be more sensitive to changes in data distribution in new data. Features such as recipient differences, budget checks, and stakeholder proposals have proven to be the leading indicators in detecting fraud. This study emphasizes the importance of external validation and regular model updates so that fraud detection systems remain adaptive to data dynamics in the field. The resulting model can be used as a tool for supervision and decision-making to prevent fraud in distributing education funds.

Keywords: *Smart Indonesia Program, Fraud, Naive Bayes, Support Vector Machine, Machine Learning, SMOTE.*

1. Introduction

Education is essential for nation-building because it produces quality human resources, encourages innovation and economic growth, and reduces social inequality. However, financial limitations often hinder access to education for children from underprivileged families [1]. To overcome this, the government, through Presidential Instruction No. 7 of 2014, launched the Productive Family Program, including the The Indonesia Smart Program (Program Indonesia Pintar, or PIP), run by the Ministry of Education and Culture since 2015 [2]. PIP aims to improve access to education for children aged 6–21 years, prevent school dropouts, and encourage students who do not continue their education to return to school [3].

However, implementing educational assistance programs such as PIP is inseparable from the risk of abuse and fraud. Fraud in the distribution of education funds can be in the form of falsifying student data, inflating the number of aid recipients, and misappropriating funds by irresponsible individuals [4]. These fraudulent practices hurt state finances and hinder the achievement of the program's main objectives. Data from the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) shows that in 2024, the budget for PIP will reach IDR 17.9 trillion, with a target of 17.9 million students. Large amounts of funds increase the potential for fraud. Therefore, effective and efficient efforts are needed to prevent and detect fraud in distributing PIP funds.

Previous research has identified some factors that cause fraud in educational assistance programs, such as students' economic conditions, home-to-school distance, and parental education [5], as well as the use of data mining to detect anomalous patterns [6]. However, most studies are still limited to descriptive analysis and have not developed accurate and comprehensive predictive models, thus creating a research gap in detecting fraud in the distribution of PIP funds [7].

The urgency of this research is based on the importance of the PIP program in improving access to education and potential losses due to fraud. This study aims to develop a predictive model that can identify potential fraud in the distribution of PIP funds using the Naive Bayes and SVM methods [8], [9], [10]. This method was chosen because it is effective in classifying and detecting anomalies. The study also considers specific contextual factors, such as demographic, geographical, and socio-economic characteristics. Classification

algorithms such as Naive Bayes and SVM can predict potential fraud risks based on historical data. Naive Bayes algorithms, with their simplicity and computational efficiency, are suitable for handling large datasets. Meanwhile, SVMs can model non-linear relationships between input and output features, improving prediction accuracy. Combining these two methods is expected to provide more robust and accurate results.

In addition, handling imbalanced data through the SMOTE (Synthetic Minority Oversampling Technique) technique is an important focus in this study [11]. Data tends to be unbalanced in the context of fraud detection because the number of fraud cases is usually much smaller than typical cases. Therefore, SMOTE is used to generate synthetic samples from minority classes (fraud) so that the model can learn better and reduce bias against the majority class [12].

By developing an accurate predictive model, this research is expected to significantly contribute to preventing and detecting fraud in the distribution of PIP funds. This model can be used by local governments and related parties to identify schools that have the potential to commit fraud, so that more effective prevention and supervision measures can be taken. In addition, this research is also expected to provide new insights into the factors that contribute to potential fraud, so that more targeted policies can be formulated to increase the effectiveness and transparency of the PIP program.

2. Literature Review

2.1. Indonesia Smart Program

The Indonesia Smart Program (Program Indonesia Pintar, or PIP) is an educational assistance from the government to increase access to education for children aged 6-21 years from underprivileged families to prevent school dropouts. This program is part of Presidential Instruction Number 7 of 2014 concerning improving family welfare [13].

The PIP distribution process begins with data collection of prospective recipients by schools and education offices, followed by verification by the Ministry of Education and Culture. Funds are channeled through bank accounts or distribution agencies for transparency and accountability.

The amount of assistance in 2024 is adjusted to the level of education: elementary school/equivalent: IDR 450,000/year, junior high school/equivalent: IDR 750,000/year, high school/equivalent: IDR 1,000,000/year. These funds are used for educational needs such as books, stationery, uniforms, transportation, pocket money, and additional tutoring [14].

2.2. Fraud Detection

Fraud detection is identifying and uncovering fraudulent activities in a system or organization. In educational assistance programs such as PIP, fraud detection is critical to ensure that aid funds are distributed to eligible recipients and used according to their designation. Fraud detection involves a series of methods, techniques, and approaches that aim to identify suspicious patterns, anomalies, or indicators that indicate the presence of potential fraud.

Fraud, in the context of education, can take various forms, ranging from falsification of student data, inflating the number of aid recipients, to misuse of funds by irresponsible parties. According to a report from Transparency International, the education sector is vulnerable to Fraud due to the large flow of public funds and the complexity of the aid distribution system. Therefore, developing and implementing effective detection systems for Fraud is a top priority for educational institutions and the government [15].

2.3. Predictive Analysis

Predictive analytics is a method that uses statistics, modeling, and machine learning to analyze historical data and predict future events. This technique is important in identifying suspicious patterns and preventing potential fraud, including in the PIP. The main components of predictive analysis include quality data, statistical and machine learning algorithms such as logistic regression, random forests, gradient boosting, deep learning, and adequate technological infrastructure. The data can include recipient demographic information, disbursement history, school characteristics, and regional socio-economic indicators. This analysis process includes stages such as problem identification, data collection and cleaning, model selection and training, model evaluation and validation, and continuous implementation and monitoring in the operational system [16].

2.4. Data Mining for Fraud Detection

Data mining for fraud detection is the application of data analysis techniques to identify abnormal patterns or anomalies that can indicate fraudulent activities in a dataset. With the help of machine learning algorithms and computational capabilities, this approach can analyze large and complex amounts of data that are difficult to uncover by traditional methods. In the context of aid distribution, data mining can detect suspicious or irregular patterns of fund disbursement. The process includes collecting data from various sources such as transaction records and recipient profiles, followed by data preprocessing, including cleanup, missing value handling, and normalization. After that, feature extraction is done to identify important variables, followed by applying algorithms such as classification, clustering, or anomaly detection. The final results are evaluated and interpreted to produce information that can be used as the basis for fraud prevention or handling measures [17].

2.5. Classification in Machine Learning

Classification is the main task in machine learning that aims to predict the label or category of a dataset based on its features. Included in supervised learning, classification involves training models with pre-labeled data. The model then maps the inputs to discrete outputs through mathematical and logical functions. There are two main classifications: binary classifications that predict two classes, and multi-class classifications that involve more than two classes. The classification process usually includes several stages: data preprocessing (cleaning, missing value handling, and normalization), selecting relevant features, division of datasets into training and test data, model training, and model performance evaluation using test data [18].

2.6. Naïve Bayes Algorithm

The Naive Bayes algorithm is a probabilistic classification method based on the application of Bayes' Theorem, assuming a strong (naïve) independence between features. Although these assumptions are often unrealistic, Naive Bayes has proven effective in various

applications, including spam detection, sentiment analysis, and fraud detection. The main advantages of Naive Bayes lie in its simplicity, computational efficiency, and ability to handle high-dimensional datasets, which makes it suitable for applications such as fraud detection in the distribution of education aid funds. Bayes' theorem, on which this algorithm is based, is expressed as:

$$.P(C \mid X) = (P(X \mid C)P(C))/P(X) \quad (1)$$

Where:
X = Unknown Class (Label) sample data
C = Hypothesis that X is data with class C (Known class)
P(C) = Probability of Hypothesis C
P(X) = Qualified Sample data opportunities
P(X | C) = Chance of sample data X, when it is assumed that the Hypothesis is true (Valid)
In detecting Fraud in the distribution of PIP funds, C can represent the class "Fraud" or "Non-Fraud". At the same time, X is a feature vector that describes the characteristics of a transaction or beneficiary [19].

2.7. Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is a popular and effective machine learning method for classification and regression tasks, first introduced by Vapnik in 1995. SVM looks for the best hyperplane that separates two data classes, using the maximum margin and support vector, i.e., the data closest to the hyperplane. To address data that cannot be separated linearly, SVM uses a kernel function that transforms the data to a higher-dimensional space to make separation possible. This capability makes SVM well-suited for handling complex and non-linear classification problems, such as fraud detection. In this context, SVM can identify fraud patterns that involve non-linear relationships between variables, such as the amount of funds, the frequency of disbursement, and the characteristics of the beneficiaries [20].
Mathematically, for the case of binary classification, SVM seeks to find a hyperplane that meets the equation:

$$w.x + b = 0 \quad (2)$$

Where w is the normal weight vector against the hyperplane, x is the input vector, and b is the bias. The goal is to maximize the margin between the two classes.

2.8. Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method used to solve the problem of class imbalance in datasets. This method was first introduced by Chawla in 2002 as a solution to improve classification performance on unbalanced datasets. SMOTE creates new synthetic samples from minority classes, so that the number of minority class samples increases and is more balanced with the majority class.
In fraud detection, SMOTE can be an instrumental technique. Fraud cases are usually much fewer than everyday transactions, so the dataset tends to be unbalanced. This imbalance can cause classification models to tend to be biased towards the majority class and less sensitive to the minority class (fraud). Applying SMOTE can synthetically increase the number of fraud samples so that the model can learn better from both classes [21].

3. Method

3.1. Needs Analysis

The data used in this study are data from the PIP and Basic Education Data (DAPODIK). The data includes information on PIP recipients and the number of elementary to high school/vocational school students throughout Indonesia for the 2023 Fiscal Year. Details of the data used include:

1. Identity data of PIP recipients.
2. Data on PIP recipients.
3. Data on the value of the aid distributed.
4. Data on the number of students of the Education unit in Dapodik.

The data is sourced from the Education Financing Service Center (Puslapdik) and the Data and Information Center (Pusdatin) of the Ministry of Education, Culture, Research and Technology. PIP data is a database with SQL extensions. At the same time, Dapodik data is Excel data. The data includes information on the distribution of the PIP for the 2023 fiscal year, and data on the total number of students in each educational unit, covering elementary to high school/vocational levels throughout Indonesia.

3.2. Population and Sample

The population in this study is all educational units in Indonesia registered in the Puslapdik and Pusdatin databases, covering elementary to high school/vocational schools that receive the PIP and are registered in DAPODIK. The research sample is a part of this population that is selected to be used as a research object. This study uses the stratification sample method to select the sample. Strata Identification of PIP recipient school population is divided by type of school (elementary, junior high, high school/vocational school) and geographical location (province), ensuring that the sample represents different schools and locations in Indonesia.
Furthermore, sampling is taken from the most significant number of PIP recipients. The sample method is based on size (proportional sampling), where each stratum can be selected proportionally to the size or number of PIP recipients. The data is sourced from Puslapdik and Pusdatin, including PIP distribution information and the number of students. This data ensures that the selected sample has complete and relevant information. PIP recipients in West Java, Central Java, East Java, and DKI Jakarta are the top 4 (four) PIP recipients with a percentage size compared to the total population of 39.64% of PIP recipients, which were chosen as the object of research.

3.3. Data Collection Procedure

The data collection process is carried out through several methods. First, data extraction from the Puslapdik database uses SQL queries to meet the research needs. Second, collecting the number of students in all Education units on the DAPODIK application from Pusdatin. Furthermore, the data was iterated by combining PIP recipient data with data on the number of DAPODIK students. The data iteration process, which includes extracting PIP data from the Puslapdik and Pusdatin databases of the Ministry of Education and Culture, is then combined with PIP and DAPODIK data with the help of the SQL Server 2022 database application to extract the two databases to produce a new dataset.

After the dataset is formed, the data results are combined. Obtain additional information from the Ministry of Education and Culture and the Auditor General regarding the patterns often found in the problem of misuse of PIP funds. Next, data cleansing is carried out to remove invalid or incomplete data, transform the data to suit the needs of the analysis, and select relevant features for classification.

3.4. Technical Analysis

3.4.1. Data Preprocessing

The pre-processing stage of data is an important part of the predictive analysis stage. These activities include column name cleanup, numeric data type conversion, categorical variable encoding, and engineering new features relevant for potential fraud detection. This process is done with special functions, such as pra_proses_data (df), which cleans up the column names and converts numeric data using regex to be ready for further processing.

3.4.2. Feature Selection, Data Split, and Data Normalization

Feature selection is done based on domain knowledge and statistical analysis, for example, with SelectKBest and the f_classif function to select the most relevant features to the target. Then all numerical features are normalized using the Standard Scaler to have a zero mean distribution and a standard deviation of one.

3.4.3. Data Imbalance Handling (SMOTE)

SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique that creates synthetic samples from minority classes using linear interpolation between existing samples. For each minority sample, SMOTE generates a new synthetic sample with the formula:$x\_i x\_new$

$$x\_new = x\_i + \lambda . (x\_zi - x\_i) \quad (3)$$

Where:

x_i = Original minority sample.

x_zi = One of the k-nearest neighbors (nearest neighbor to z) from.x_i

λ = Random value $\in$

3.4.4. Development of the Naïve Bayes Model and SVM

Naive Bayes was chosen in this study because of its efficient ability to handle high-dimensional data and its advantages in real-time prediction scenarios. This algorithm works based on Bayes' Theorem, assuming independence between features. The selection of Naive Bayes as a baseline model is based on high computational speed, suitable for data with large sample sizes, good interpretability through posterior probability, easy analysis of the dominant factors causing fraud, and performance stability despite noise in the PIP-DAPODIC integration data.

The Naive Bayes model is GaussianNB, assuming the features are usually distributed in each class. This model will calculate the probability of a school being categorized as "fraud" based on historical patterns and indicators such as the difference between recipients and the percentage of unliquidated students.

SVM is one of the most popular classification algorithms in machine learning, especially for anomaly and fraud detection problems. Using the kernel concept, SVM is known for building optimal decision boundaries, even on non-linearly separable data. In this study, SVM was chosen as a comparative model that can capture complex patterns in PIP distribution data, so it is expected to increase the accuracy and sensitivity of fraud case detection compared to baseline models such as Naive Bayes. SVM is used with hyperparameter optimization via RandomizedSearchCV.

3.5. Model Evaluation

Evaluation of the performance of classification models is a crucial stage in developing and applying machine learning models, particularly in fraud detection research where accurate identification of potential fraud is essential. This process aims to measure how well a model can predict a data class that has never been seen before, ensuring that the algorithms developed can accurately and reliably identify potential fraud on unseen data. One of the fundamental concepts in evaluating classification models is the Confusion Matrix, which is a table that shows the model's performance in predicting a class by comparing the prediction results with the actual values.

Table 1. Confusion Matrix Table

Class		Classification Results	
		P	**N**
Target	T	TP	TN
	F	FP	FN

Table 1 describes For binary classification cases such as fraud detection (Fraud vs. Non-Fraud), the Confusion Matrix consists of four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), where TP and TN indicate correct

predictions, while FP and FN represent prediction errors [22]. In addition to measuring prediction accuracy, the evaluation stage in machine learning-based research also assesses the balance between fraud case detection (recall), overall prediction accuracy, and the model's ability to distinguish between fraud and non-fraud classes using metrics such as the F1-score.

4. Results and Discussion

4.1. Dataset Results

After the merger of the SIPINTAR PIP data with DAPODIK, additional information from the Auditor General of the Ministry of Education and Culture, research, and technology on the patterns often found in the problem of misuse of PIP funds, This dataset consists of more than 93,658 data records of Education units that received PIP assistance, with the following dataset information as in Figure 1 below:

```
RangeIndex: 93658 entries, 0 to 93657
Data columns (total 34 columns):
#  Column                Non-Null Count  Dtype    #  Column                   Non-Null Count   Dtype
-- ------                --------------  -----    -- ------                   --------------   -----
0  ï»¿year               93658 non-null  int64    17 proposal_service         93658 non-null  int64
1  prov                  93658 non-null  int64    18 risk                     93658 non-null
2  province              93658 non-null  object   object
3  kab                   93658 non-null  int64    19 worthy                   93658 non-null
4  regency               93658 non-null  object   int64
5  kec                   93658 non-null  int64    20 not_feasible             93658 non-null
6  subdistrict           93658 non-null  object   int64
7  school_uid            93658 non-null  object   21 new_savings_status       93658 non-null
8  school_name           93658 non-null  object   int64
9  npsn                  93658 non-null  object   22 old_savings_status       93658 non-null
10 level                 93658 non-null  int64    int64
11 level.1               93658 non-null  object   23  anomalous_savings_status  93658  non-null
12 amount_recipient_pip 93658 non-null  int64    int64
13 amount_total_student 93658 non-null  int64    24 already_activated        93658 non-null
14     percentage_difference93658     non-null   int64
float64                                          25 not_activated            93658 non-null
15 proposal_dkts         93658 non-null  int64    int64
16 proposal_stakeholders 93658 non-null  int64    26 not_yet_liquid           93658 non-null
                                                  int64
                                                  27 liquid                   93658 non-null
                                                  int64
                                                  28 potential                93658 non-null
                                                  object
                                                  29 total_number            93658 non-null  int64
                                                  30 check_budget             93658 non-null
                                                  int64
                                                  31 verification_proposal    93658 non-null
                                                  int64
                                                  32 check_eligibility        93658 non-null
                                                  int64
                                                  33 specific_audit           93658 non-null
                                                  int64
dtypes: float64(1), int64(24), object(9)
```

Fig 1. Dataset information

4.2. Data Preprocessing Results

Data preprocessing is carried out to improve the precision and performance of the data so that it becomes data ready for data training and testing, making the data processing process important in this research. The data is then cleaned of empty values, duplications, and inconsistencies. Furthermore, numerical data transformation is carried out using StandardScaler normalization to ensure that each feature is on a uniform scale, so that the model training process is not biased towards features with an extensive range of values. Category data is processed with a one-hot encoding technique so that it can be accepted by Naïve Bayes and SVM algorithms. Then, pre-processing of the data results from data transformation after the cleaning process, data type conversion, feature engineering, and label mapping to obtain empty data in the dataset.

The results of checking the blank data show that all columns in the dataset have been filled in completely, without a single missing value in each main attribute, such as year, province, district, sub-district, school name, NPSN, level, number of PIP recipients, and total number of students. This condition indicates that the data quality used in the study is excellent, so there is no need for an imputation or deletion process due to missing values. Furthermore, a feature engineering process was carried out by adding new columns for the difference in recipients and the percentage that had not been disbursed, so that the dataset amounted to 36 columns. Adding two columns to the selection dataset aims to enrich the available information and improve the quality of the analysis. With these two additional features, the modeling and fraud detection process is expected to be more accurate because the model can take advantage of information that was not previously available. All columns in both datasets have also been checked and confirmed to have no blank values, so the data quality is maintained. Here are the top 20 lines of data for educational units that have been feature engineered:

Table 2. Top data lines of Education units that received PIP after feature engineering

	Year	Prov	Province	Regency	Regency	Subdistrict	Check _Budget	Verification _Proposal	Cheque_ Feasibility	Audit_ Certain	Difference_ Recipient	Percentage _Not yet _Liquid
0	2023	2	West Java	205	Bogor	20521	1	1	0	0	-45	0.0
1	2023	3	Central Java	316	Blora	31616	1	0	0	0	-16	0.0
2	2023	3	Central Java	318	Pati	31803	0	0	0	0	-152	0.0
3	2023	3	Central Java	310	Klaten	31001	1	0	0	0	-47	0.0
4	2023	2	West Java	212	Tasikmalaya	21224	1	0	0	0	-38	0.0
5	2023	3	Central Java	314	Sragen	31403	1	1	0	0	-8	0.0
6	2023	3	Central Java	309	Boyolali	30902	1	0	0	0	-37	0.0
7	2023	2	West Java	208	Bandung	20812	1	0	0	0	-30	0.0
8	2023	2	West Java	215	Kuningan	21513	1	0	0	0	-127	0.0
9	2023	5	East Java	560	Surabaya	56017	0	0	0	0	-65	0.0
10	2023	3	Central Java	301	Cilacap	30120	1	0	0	0	-44	0.0
11	2023	5	East Java	524	Jember	52418	0	0	0	0	-278	0.0
12	2023	2	West Java	261	Bogor	26103	1	0	0	0	-135	0.0
13	2023	5	East Java	514	Nganjuk	51407	1	0	0	0	-22	0.0
14	2023	3	Central Java	301	Cilacap	30120	1	0	0	0	-31	0.0
15	2023	5	East Java	524	Jember	52419	1	0	0	0	-58	0.0
16	2023	3	Central Java	327	Pemalang	32707	0	0	0	0	-99	0.0
17	2023	5	East Java	518	Malang	51813	1	0	0	0	-82	0.0
18	2023	3	Central Java	306	Purworejo	30609	1	1	0	0	-7	0.0
19	2023	2	West Java	219	Subang	21906	0	0	0	0	-134	0.0

The selection of features aims to select the most relevant variables that affect the target to be predicted, so that the resulting model becomes more straightforward, more efficient, and has better accuracy. Reducing non-essential or redundant features makes the model training process faster and the results easier to interpret. One commonly used way to assess feature relevance is to examine the degree of correlation between each feature and the target. This correlation analysis helps identify which features strongly relate to the target, making them worth considering in the later modeling process. The data before normalization on each feature has a very different range of values. Features such as usulan_pemangku have values ranging from 0 to hundreds, jumlah_total features range from millions to hundreds of millions, features persentase_belum_cair are in the range of 0 to tens, categorical features such as risk, tidak_layak, and periksa_anggaran are worth 0 or 1. This difference in scale causes scale-sensitive machine learning models such as SVM to be biased towards large-value features, so features with small ranges become less influential in the model training process.

After normalization using the StandardScaler, all numerical features are changed to have a mean of 0 and a standard deviation of 1. As a result, all features are on a relatively equal scale, around -3 to +3. The extreme value (outlier) remains, but has been anticipated in standard units of deviation from the mean. Categorical features converted to numeric have also transformed, so the values of 0 and 1 change to negative and positive values according to the data distribution. The following are the results of data normalization as shown in Figure 2::

```
The first five lines of the normalization result of the feature:
[[-0.2858565  -0.42851415 -0.218973    0.5302298  -0.3717361   0.64643094
  -0.23624977 -0.19228205  0.87416193 -0.41533    -0.30298211 -0.67478308]
 [-0.38470501 -0.42851415 -0.27693965 -0.58137725  1.39714608 -0.46387952
   0.62320182  0.17963221 -1.14395281 -0.41533    -0.30298211  1.48195773]
 [-0.38470501 -0.42851415 -0.27693965  0.67661427  0.86877868 -0.74145714
  -0.05930385 -0.63283585  0.87416193 -0.41533    -0.30298211 -0.67478308]
 [-0.38470501 -0.42851415 -0.27693965  0.29692956 -0.3717361   0.6968996
  -0.23624977  0.13865861  0.87416193 -0.41533    -0.30298211  1.48195773]
 [-0.38470501 -0.42851415 -0.27693965  0.50735723 -0.3717361  -0.28723922
  -0.23624977 -0.49485636  0.87416193 -0.41533    -0.30298211 -0.67478308]]

The last five lines of the normalization result of the feature:
[[ 0.62025492 -0.42851415  0.56357677 -1.30415055 -0.3717361  -0.56481683
  -0.23624977  2.24407154 -1.14395281  2.40772399 -0.30298211  1.48195773]
 [-0.38470501 -0.42851415 -0.27693965  0.39756888 -0.3717361  -0.26200489
  -0.23624977  0.44753655  0.87416193 -0.41533    -0.30298211  1.48195773]
 [-0.38470501 -0.42851415 -0.27693965  0.3746963  -0.3717361  -0.11059892
  -0.23624977 -0.06620942  0.87416193 -0.41533    -0.30298211 -0.67478308]
 [-0.38470501 -0.42851415 -0.27693965  0.63544364  4.17681807 -0.74145714
  -0.21097178 -0.73789638  0.87416193 -0.41533    -0.30298211 -0.67478308]
 [-0.38470501 -0.42851415 -0.27693965  0.53480431 -0.3717361   1.68103842
```

```
-0.23624977  0.37819661  0.87416193 -0.41533    -0.30298211  1.48195773]]
```

Fig 2. results of data normalization

The main problem in this dataset is class imbalance, where the proportion of fraud data is much less than that of non-fraud data. To overcome this, the SMOTE (Synthetic Minority Over-sampling Technique) technique with a parameter of k_neighbors=5 is used to oversample the minority class. As a result, the class distribution is balanced (50:50) without causing overfitting, which is shown by the stability of descriptive statistics before and after SMOTE. This process ensures that the built model can learn optimally from both classes.

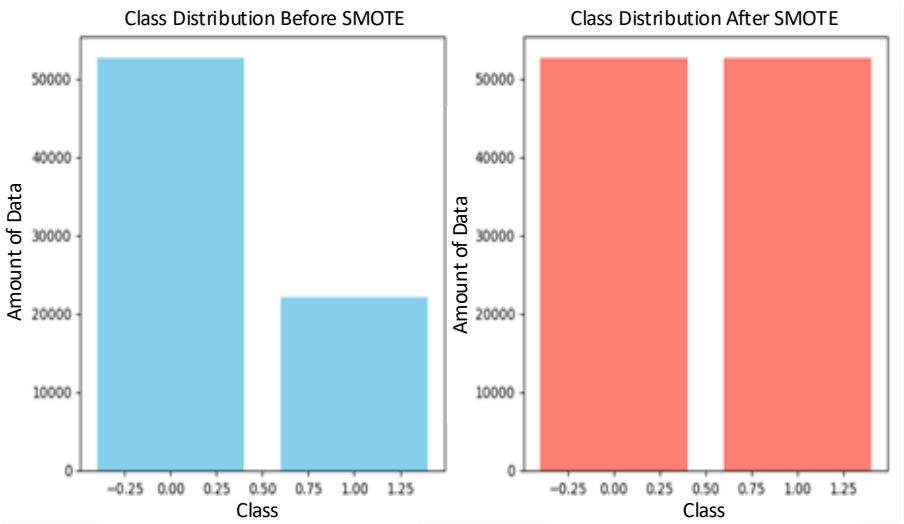


Fig 3. Comparison of results before and after SMOTE

The class distribution before applying SMOTE shows a significant imbalance, with Class 1 having only 22,180 instances compared to 52,746 instances for Class 0. After applying SMOTE, the data becomes imbalanced, with both Class 1 and Class 0 containing 52,746 instances each. This balancing, as illustrated in Figure 3, demonstrates that SMOTE effectively addresses class imbalance by generating synthetic samples for the minority class, resulting in equal representation of both classes and supporting more reliable model training and evaluation

4.3. Results of the Development of the Naïve Bayes Model

The Naive Bayes model is one of the main approaches to classify educational units based on potential fraud in the distribution of PIP funds. Naive Bayes was chosen because this algorithm is simple, fast, and effective in handling extensive data, and can provide probability estimates for each class based on the available features. Assuming that each feature is independent of the others, the model statistically calculates the chances of fraud occurring in each school. The results of the prediction and performance evaluation of the Naive Bayes model will be described below to assess how well this model distinguishes between schools at risk of fraud and those not. The parameters of the naïve Bayes model with the prior probabilities (classes) of fraud and normal are [0.5, 0.5], and the mean and variance of each feature for each class. The following are the results of the prediction from the Naïve Bayes model of the status of the fraud category that has been tested:

Table 3. Prediction Results of the Naïve Bayes Model

School_Name	Province	Regency	Current	Predicted	Category
SD NEGERI PENGGALANG 04 ADI PALA	Central Java	Cilacap	0	0	Usual
SD 2 KUTUK	Central Java	Kudus	1	0	Usual
SD AL ISLAM MOROWUDI	East Java	Gresik	1	1	Fraud
SD NEGERI 2 WATUREJO	East Java	Malang	0	0	Usual
SDISLAM PLUS AL MUKHTARIYYAH	Central Java	Blora	0	0	Usual
SMKS BANDUNG SELATAN 1 BANDUNG	West Java	Bandung	1	1	Fraud
SMAN 1 PUJER	East Java	Bondowoso	0	0	Usual
SDN CARIWUH	West Java	Tasikmalaya	0	0	Usual
SMP IT AR-RUDHO	D.K.I. Jakarta	Jakarta Timur	0	0	Usual
SD NEGERI WONOREJO 03	Central Java	Semarang	0	0	Usual
SLB AL-AZAMI	West Java	Cianjur	1	0	Usual
SMAS MATHA UL ANWAR	West Java	Karawang	1	1	Fraud
SLB YAKALIMU	West Java	Purwakarta	1	0	Usual
SMP PGRI 01 BANTUR	East Java	Malang	1	1	Fraud
SMPS HIDAYATULLAH WEDUNG	Central Java	Demank	1	0	Usual
SMPN 4 KLARI	West Java	Karawang	0	0	Usual
PKBM KUSUMA WIJAYA	East Java	Surabaya	1	0	Usual
PKBM SEJAHTERA	Central Java	Rembang	1	0	Usual
PKBM BANGUN BANGSA	Central Java	Semarang	1	0	Usual

Table 3 shows the results of the Naive Bayes model prediction of fraud status in several educational units from various provinces and districts in Indonesia. Each row represents one school, with information ranging from the name of the school, location (province and district), to actual results (Actual) and model prediction results (Predicted). In addition, this table also presents the prediction

probabilities for each class, namely Probability_Normal and Probability_Fraud, as well as the final category based on the prediction results.

In most cases, the Naive Bayes model can distinguish between schools classified as usual and fraudulent by providing a fairly strict probability. The results of this prediction also show a tendency for the model to give a high probability to the majority (standard) class, which is reflected in the number of Probability_Normal close to 1 in schools with the actual standard label. Meanwhile, some fraud cases identified by the model are also supported by very high Probability_Fraud, such as at SMP PGRI 01 BANTUR. These findings indicate that the Naive Bayes model effectively identifies fraud cases with very different feature patterns from regular classes. However, there is still the potential for misclassification in cases with similar characteristics between the two classes.

4.4. Results of the Development of the Support Vector Machine (SVM) Model

The next stage is to build and test a Support Vector Machine (SVM) model to classify potential fraud in PIP recipients' education units. SVM was chosen because this algorithm is known to be very effective in handling data with linear and non-linear complex patterns. By leveraging the RBF kernel, SVM can establish an optimal separator boundary between the fraud and regular classes, even when data characteristics are difficult to separate.

The value of parameter C is set at 10, which means that the model provides a considerable penalty for misclassification to focus more on separating fraud and non-fraud data with optimal margins. The gamma parameter is set to "auto, " meaning that the gamma value is automatically adjusted based on the number of features in the data, allowing the model to adjust its sensitivity to local patterns.

Using the RBF kernel allows SVM to capture complex non-linear patterns in the data, so that the model not only distinguishes data based on straight lines, but can also recognize more complicated relationships between features. The resulting intercept value of -0.06309 indicates the position of the separating hyperplane with respect to the point of origin in the feature space. These parameters are combined through a tuning process to obtain optimal fraud detection performance on PIP distribution data. With this configuration, the SVM model is expected to provide more accurate and reliable predictions in distinguishing education units at risk of fraud and those not.

After obtaining the parameter tuning results, using the RBF kernel with a value of C=10 and gamma='auto' or 'scale' consistently resulted in the highest accuracy score, with a mean_score value close to 0.99. This shows that the model with the RBF kernel can capture the nonlinear patterns that exist in data fraud very well. In contrast, linear kernels tend to result in slightly lower scores, although the training time is relatively shorter. The following are the results of the prediction from the Naïve Bayes model of the status of the fraud category that has been tested:

Table 4. SVM Model Prediction Results

School Name	Regency/City	Province	Current	Predicted	Category
SD NEGERI PENGGALANG 04 ADIPALA	Cilacap	Central Java	0	0	Usual
SD 2 KUTUK	Kudus	Central Java	1	1	Fraud
SD AL ISLAM MOROWUDI	Gresik	East Java	1	1	Fraud
SD NEGERI 2 WATUREJO	Malang	East Java	0	0	Usual
SD ISLAM PLUSAL MUKHTARIYYAH	Blora	Central Java	0	0	Usual
SMKS BANDUNG SELATAN 1 BANDUNG	Bandung	West Java	1	1	Fraud
SMAN 1 PUJER	Bondowoso	East Java	0	0	Usual
SD N CARIWUH	Tasikmalaya	West Java	0	0	Usual
SMP IT AR-RUDHO	East Jakarta	D.K.I. Jakarta	0	0	Usual
SD NEGERI WONOREJO 03	Semarang	Central Java	0	0	Usual
SLB AL-AZAMI	Cianjur	West Java	1	1	Fraud
SMAS MATHLAUL ANWAR	Karawang	West Java	1	1	Fraud
SLB YAKALIMU	Purwakarta	West Java	1	1	Fraud
SMP PGRI 01 BNATUR	Malang	East Java	1	1	Fraud
SMPS HIDAYATULLAH WEDUNG	Demak	Central Java	1	1	Fraud
SMPN 4 KLARI	Karawang	West Java	0	0	Usual
PKBM KUSUMA WIJAYA	Surabaya	East Java	1	1	Fraud
PKBM SEJAHTERA	Rembang	Central Java	1	1	Fraud
PKBM BANGUN BANGSA	Semarang	Central Java	1	1	Fraud

Table 4 shows the predictions of the Support Vector Machine (SVM) model on fraud status in various educational units from several provinces and districts. Each row displays the school name, location, actual label, model prediction results (Predicted), the probabilities of each category (Normal Probability and Fraud Probability), and the final category is assigned based on that prediction. From the results of this prediction, it can be seen that SVM can provide substantial predictions for most cases. The SVM model can identify obvious fraud cases and is quite sensitive to borderline cases. For example, at SMKS BANDUNG SELATAN 1 BANDUNG, the probabilities of normal and fraud are not both extreme. However, the model still classifies the school as fraudulent because the likelihood of fraud is higher than usual. A similar pattern is also seen in several other schools, such as PKBM KUSUMA WIJAYA and PKBM SEJAHTERA, which have a very high probability of fraud and are eventually categorized as fraud.

4.5. Implementation Results on New Data

After the model is developed and evaluated using training and test data, the next step is to test the model's performance on a new dataset from PIP recipient schools in Banten Province. This test aims to find out the extent to which the model that has been built can recognize fraud patterns in data that has never really been analyzed before. Using data on PIP recipient schools in Banten Province, this evaluation is also a benchmark for the model's generalization ability in detecting potential fraud in areas with different characteristics from the training data. The results of implementing this new data will serve as a basis for assessing the model's reliability in real applications and provide an overview of the challenges that may be faced when the model is used in the field. The following is an example of the prediction results of the Naïve Bayes and SVM models using a new dataset in Banten Province.

Table 5. Comparison of Naïve Bayes and SVM Prediction Results Using the New Dataset

	School_Name	District	Regency	Potential	Predicted_ Risk_nb	Predicted_ Risk_nb
0	SDN 2 KRIMATJAYA	Mount Kencana	Lebak	1.0	1	1
1	SD NEGERI CEMPAKA PUTIH 02	East Ciputat	South Tangerang	0.0	1	0
2	SD NEGERI JOMBANG 02	Ciputat	South Tangerang	0.0	1	0
3	SKH KEKERABATAN	Maja	Lebak	1.0	0	1
4	SMAN 17 KABUPATEN TANGERANG	Legok	Tangerang	0.0	1	0
5	SKh. AL KHAIRIYAH	Citangkil	Cilegon	1.0	0	1
6	SKh. MUSTIKA TIGARAKSA	Tiga raksa	Tangerang	1.0	0	1
7	SDN 1 HAURGAJRUG	Cipanas	Lebak	0.0	0	0
8	SD NEGERI TANGERANG 15	Tangerang	Tangerang	0.0	0	0
9	SD NEGERI PONDOK RANJI 02	East Ciputat	South Tangerang	0.0	0	0

Table 5 shows that both models produce the exact predictions but differ in some cases. For example, at SD NEGERI JOMBANG 02 in South Tangerang City, the actual label of potential risk is 0 (not at risk), but the Naive Bayes model predicts 1 (risky). In contrast, SVM predicts 0, according to the actual label. On the other hand, in SKH KEKINATAN in Lebak Regency, the actual label and SVM prediction are both 1 (risky), while Naive Bayes predicts 0.

4.6. Comparative Analysis of Model Evaluation

At the model evaluation stage, the performance of Naive Bayes and Support Vector Machine (SVM) was compared using several scenarios, namely without normalization and SMOTE, with normalization without SMOTE, with SMOTE without normalization, and a combination of normalization and SMOTE. Each model was tested using accuracy, precision, recall, and F1-score metrics to assess its effectiveness in detecting educational units at risk of fraud in the distribution of PIP funds.

Table 6. The results of the comparison test before and after normalization with 80% training data, 20% test data

Type	CA		F-1 Score		Accuracy		Recall	
	Before	Scaling	Before	Scaling	Before	Scaling	Before	Scaling
SVM	0.70	0.99	0.58	0.99	0.79	0.99	0.70	0.99
NB	0.70	0.81	0.58	0.78	0.79	0.83	0.70	0.81

The test results in Table 6 show that in the data without normalization and SMOTE, Naive Bayes and SVM could only achieve an accuracy of about 70%. The recall for the fraud class in this condition is very low, at 0% for both models, indicating that the model fails to recognize fraud cases when the data is not processed correctly.

Table 7. The results of the comparison trial before and after SMOTE with 80% training data, 20% test data

Type	CA		F-1 Score		Accuracy		Recall	
	Before	SMOTE	Before	SMOTE	Before	SMOTE	Before	SMOTE
SVM	0.70	0.62	0.58	0.63	0.79	0.65	0.70	0.62
NB	0.70	0.36	0.58	0.30	0.79	0.64	0.70	0.36

The application of SMOTE without normalization produced suboptimal results in both models, as shown in Table 7. Naive Bayes experienced a decrease in accuracy of up to 36%, with recall fraud indeed high (91%). Still, the precision and F1-score were very low. SVM also does not perform well in this scenario, with an accuracy of only 62% and an F1-score below 0.65.

Table 8. Comparison test results before and after Combination Normalization and SMOTE with 80% training data, 20% test data

Type	CA		F-1 Score		Accuracy		Recall	
	Before	Combination	Before	Combination	Before	Combination	Before	Combination
SVM	0.70	0.99	0.58	0.99	0.79	0.99	0.70	0.99
NB	0.70	0.81	0.58	0.79	0.79	0.83	0.70	0.81

In Table 8, the combination of normalization and SMOTE gives the best results for both models, especially on SVM. Naive Bayes achieved 81% accuracy, 89% fraud accuracy, 42% recall fraud, and an F1 score of 0.57. However, SVM consistently excels with 99% accuracy, 96% accuracy for fraud, 99% recall, and an F1-score of 0.98.

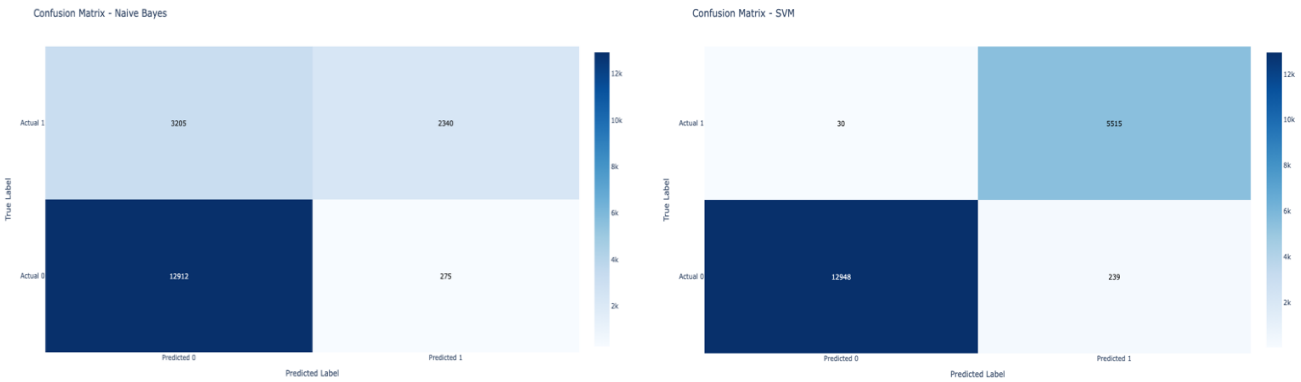


Fig 4. Comparison of Naive Bayes and SVM confusion matrix results

Figure 4 shows that The Naive Bayes model confusion matrix shows 2,340 True Positives (TP), 12,912 True Negatives (TN), 275 False Positives (FP), and 3,205 False Negatives (FN), indicating that while the model correctly identifies a large number of normal cases (TN), it misses many fraud cases (high FN) and detects fewer actual frauds (lower TP). In contrast, the SVM model achieves 5,515 TP, 12,948 TN, 239 FP, and only 30 FN, demonstrating a much higher ability to correctly detect fraud (higher TP and much lower FN), slightly better at identifying normal cases (higher TN), and making fewer mistakes in labeling normal cases as fraud (lower FP). Overall, the SVM model outperforms Naive Bayes in both fraud detection and minimizing missed fraud cases, as reflected in the significantly higher TP and lower FN values.

Table 9. Comparison test results before and after Combination Normalization and SMOTE with 80% training data, 20% test data

Type	CA		F-1 Score		Accuracy		Recall	
	Data Test	New Dataset	Data Test	New Dataset	Data Test	New Dataset	Data Test	New Dataset
SVM	0.99	0.845	0.99	0.641	0.99	0.757	0.99	0.556
NB	0.81	0.639	0.79	0.419	0.83	0.349	0.81	0.524

Table 9 shows that performance was significantly decreased when both models were tested on the new dataset, especially in the Naive Bayes model. The SVM accuracy dropped from 0.99 on the test data to 0.845 on the new dataset, the F1-score dropped to 0.641, the precision to 0.757, and the recall to 0.556. The Naive Bayes model experienced a sharper decline in performance on the new dataset. The accuracy dropped to 0.639, the F1-score to 0.419, the precision to 0.349, and the recall to 0.524.

5. Conclusion

The study demonstrates that Support Vector Machines (SVMs) with RBF kernels and optimized parameters consistently outperform other models in identifying educational units at risk of fraud, achieving up to 99% accuracy and F1-scores in the fraud class. In contrast, although simple and computationally efficient, the Naive Bayes model shows limitations in detecting complex fraud patterns, especially under imbalanced data conditions, with performance dropping to 81% accuracy and only 42% recall for fraud detection. When applied to a new dataset from Banten Province, both models experienced performance declines, highlighting challenges in generalizing across datasets with differing characteristics. SVM remained more robust with 84.5% accuracy and a 64.1% F1-score, whereas Naive Bayes fell to 63.9% accuracy and 41.9% F1-score. These results underscore the importance of regular model updates and external validation to ensure adaptability in real-world settings. Key factors influencing fraud detection include recipient discrepancies, budget checks, stakeholder proposals, risk indicators, and anomalous savings patterns—features that consistently emerged as significant in the analysis. Overall, the developed predictive model holds promise as a practical tool for monitoring and early detection, enabling more targeted and efficient oversight of PIP fund distribution.

References

[1] I. Kusumawati *et al.*, *Pengantar Pendidikan*. CV Rey Media Grafika, 2023.
[2] N. Sufni, "Analisis Keberhasilan Program Kartu Indonesia Pintar (KIP) dalam Meningkatkan Akses Pendidikan di Indonesia," *Benefit J. Bussiness, Econ. Financ.*, vol. 2, no. 2, pp. 38–45, 2024.
[3] B. G. Dimmera and P. D. Purnasari, "Permasalahan Dan Solusi Program Indonesia Pintar Dalam Mewujudkan Pemerataan Pendidikan Di Kabupaten Bengkayang," *Sebatik*, vol. 24, no. 2, pp. 307–314, 2020.
[4] J. Antoni, D. Candira, M. Istan, and others, "Implementasi Fraud Control Plan Dalam Pengelolaan Dana Bantuan Operasional Sekolah," *El-Idare J. Manaj. Pendidik. Islam*, vol. 10, no. 2, pp. 126–135, 2024.
[5] Gumay, R. I., & Anggai, S. (2023). Analisis dan Deteksi Risiko Fraud Pada Data Program Indonesia Pintar (PIP) Menggunakan Algoritma Machine Learning (Studi Kasus Penyaluran Dana PIP di Kab. Cianjur). Jurnal Teknologi Informasi Dan Ilmu Komputer, 10(2), 285–292.
[6] Nur, A. M. (2022). Penerapan Metode Naïve Bayes Untuk Penentuan Penerima Beasiswa Program Indonesia Pintar (PIP) di SMAN 1 Sukamulia. Jurnal Informatika, 9(1), 1–8.
[7] H. Manossoh, "Faktor-faktor penyebab terjadinya fraud pada pemerintah di Provinsi Sulawesi Utara," *J. EMBA J. Ris. Ekon. Manajemen, Bisnis dan Akunt.*, vol. 4, no. 1, 2016.
[8] N. Aini, W. Handoko, and R. Nurhaliza, "Prediksi Penerimaan Bantuan Pip Pada Smks Al-Furqon Batubara Dengan Metode Naïve Bayes," *JUTSI J. Teknol. Dan Sist. Inf.*, vol. 4, no. 1, pp. 11–20, 2024.

[9] I. Priyanto, E. M. Dewanti, T. Tundo, M. Nurdin, and R. Kasiono, “Penerapan Algoritma Metode Naïve Bayes Untuk Penentuan Penerimaan Bantuan Program Indonesia Pintar (PIP),” *J. Manajamen Inform. Jayakarta*, vol. 4, no. 2, pp. 162–172, 2024.
[10] D. Kristianti and M. A. Hariyadi, “Support Vector Machine (SVM) dan Algoritma Na{\"\i}ve Bayes (NB) untuk mengklasifikasi keterlambatan pembayaran sumbangan pendidikan di Madrasah Ibtidaiyah,” *J. Pendidik. Tambusai*, vol. 6, no. 2, pp. 13468–13477, 2022.
[11] I. A. Rahma and L. H. Suadaa, “Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1329–1340, 2023.
[12] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022.
[13] E. Edrial, R. Putrama, and A. Sujastiawan, “Evaluasi Kebijakan Program Indonesia Pintar (PIP) di SMA Negeri 1 Utan Tahun 2019-2020,” *J. Kapita Sel. Adm. Publik*, vol. 3, no. 1, pp. 109–116, 2022.
[14] F. Uriyalita, “Evaluasi Program Indonesia Pintar (PIP) Telaah tentang Aksesibilitas, Pencegahan dan Penanggulangan Anak Putus Sekolah di Wilayah Urban Fringe Harjamukti, Cirebon,” S-2 Manajemen Pendidikan Islam, 2020.
[15] E. Navira, “Pendeteksian Fraud Pada Pemerintahan Daerah Di Indonesia,” Universitas Lampung, 2023.
[16] R. S. Y. Zebua *et al.*, *Fenomena Artificial Intelligence (AI)*. PT. Sonpedia Publishing Indonesia, 2023.
[17] C. Carudin *et al.*, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia, 2024.
[18] R. M. Sari, *Klasifikasi Data Mining*. Serasi Media Teknologi, 2024.
[19] A. Pebdika, R. Herdiana, and D. Solihudin, “Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima Pip,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 452–458, 2023.
[20] M. F. Al Fikri, A. Asrianda, and Z. Fitri, "Implementation Of The Adaboost Method on Linear Kernel SVM for Classifying Pip Assistance Recipients at SMP Negeri 2 Kejuruan Muda," in *Proceedings of International Conference on Multidisciplinary Engineering (ICOMDEN)*, 2024, p. 25.
[21] B. S. Larsen, "Synthetic minority over-sampling technique (SMOTE)," *GitHub (https://github. com/dkbsl/matlab_smote/releases/tag/1.0)*, 2022.
[22] C. Anam and H. B. Santoso, “Perbandingan Kinerja Algoritma C4. 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa,” *ENERGY J. Ilm. Ilmu-Ilmu Tek.*, vol. 8, no. 1, pp. 13–19, 2018.