Computer Engineering and Applications Vol. 13, No. 3, October 2024

Imbalanced Data NearMiss for Comparison of SVM and Naive Bayes Algorithms

Wawan Gunawan [1*], Yudo Devianto [2], Anggi Puspita Sari[3]

[1]*Department of Informatic, Faculty of Computer Science, Universitas Mercu Buana*
[2]*Department of Information System, Faculty of Computer Science, Universitas Mercu Buana*
[3]*Department of Information System, Faculty of Technology and Informatics, Universitas Bina Sarana Informatika*
**wawan.gunawan@mercubuana.ac.id*

ABSTRACT

The study aims to improve the diagnosis, management, and prevention of HIV/AIDS by using classification algorithms. The dataset used consists of 707,379 records and 89 columns. Data preprocessing includes removing irrelevant attributes, handling inconsistencies, and balancing the data using the NearMiss method, resulting in a balanced proportion of reactive and non-reactive HIV cases. Once the data is balanced, it is split into several ratios: 60:40, 70:30, 80:20, and 90:10. The classification models used in this study are Naive Bayes and SVM. The models are evaluated using the metrics Accuracy, Precision, Recall, and F1-Score. The results show that the SVM model achieves the highest accuracy of 82.6% with a 90:10 data split at a 6-fold value, and 82.2% with a 60:40 data split at a 5-fold value. On the other hand, Naive Bayes achieves the highest accuracy of 61.1% with a 60:40 data split.

Keywords: HIV, Imbalance, K-Fold, Klasifikasi, ODHA.

1. INTRODUCTION

The increase in HIV cases in Indonesia surged significantly in 2023 [1]. Transmission of cases is dominated by housewives, who account for 35 percent of all cases. This alarming statistic underscores a critical public health issue. The global community is not on track to meet the HIV prevention target by 2025, which aims for less than 370,000 new HIV infections annually. In 2022, there were 1.3 million people newly infected with HIV [2], highlighting the urgent need to accelerate progress in HIV prevention efforts. In Indonesia, there are 14,150 HIV cases among children aged 1 to 14 years. This number is increasing annually by about 700 to 1,000 children. This rise in pediatric HIV cases points to systemic gaps in both prevention and treatment strategies. One of the main causes of high HIV transmission among housewives is the lack of knowledge about prevention and the impact of the disease [3]. This gap in awareness and education poses a significant challenge to curbing the spread of the virus. Gender inequality and sexual violence further exacerbate the vulnerability of women and girls to HIV. These conditions can limit their ability to negotiate safe sex practices, thereby increasing their risk of infection. Women and girls often face structural barriers that make it difficult to access HIV prevention and treatment services. Socioeconomic factors, cultural norms, and limited healthcare resources

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

contribute to these disparities, making it imperative to address these underlying issues to effectively combat the HIV epidemic.

In several countries, discriminatory laws and social norms against certain groups, such as LGBTQ+ communities, sex workers, and drug users [4], can limit their access to healthcare services [5], including those related to HIV. Resource shortages in many countries [6], especially in developing nations, result in insufficient allocation of resources for HIV/AIDS prevention, testing, and treatment programs.

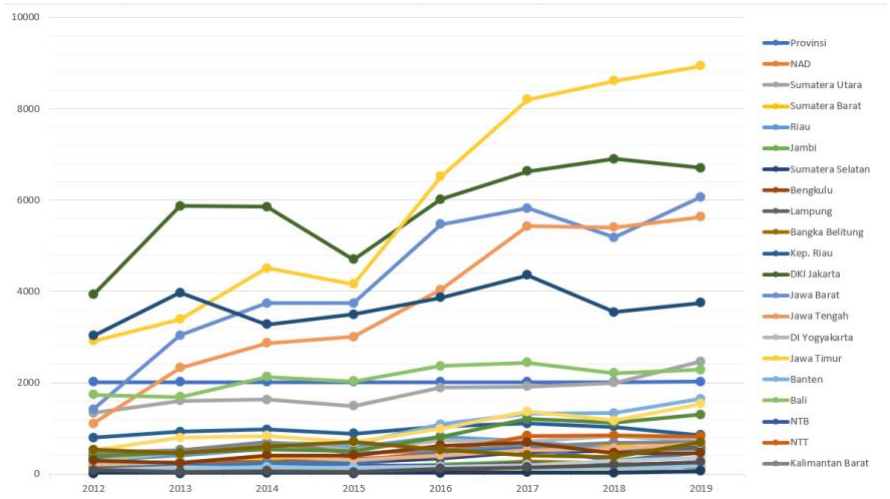


FIGURE 1. Graph of HIV/AIDS patients per year

Based on previous research, this study will use Naive Bayes, which is suitable for classification based on probabilities and is often used in medical classification. Meanwhile, Support Vector Machines (SVM) are highly effective for classification with a clear margin between classes. The Performance Evaluation of Classification Models for the HIV/AIDS Dataset shows the experimental results of the proposed methods for classifying the HIV/AIDS dataset, achieving training accuracy of up to 0.9755 and testing accuracy of up to 0.8721. This study compiles HIV prevention research studies and analyzes their effectiveness in reducing risky sexual behaviors [7] and drug-related behaviors for HIV transmission [8]. The database contains over 5000 reports, including 586 reports with outcome data from intervention studies. By identifying response patterns to treatment, the study aims to facilitate more personalized and effective care. In the journal, the results presented include the use of machine learning, specifically the XGBoost model [9], to predict the prognosis of talaromycosis in HIV/AIDS patients. The results show that the XGBoost model has the best predictive ability compared to other machine learning algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM).

Furthermore, another study mentioned that from the analysis results, it was found that the Logistic Regression model had the highest accuracy of 82.46%, compared to other data mining techniques. Meanwhile, SVM had a lower accuracy of 79.22% [10].

Another study focuses on the use of data mining techniques for data classification, specifically to estimate the potential for having breast cancer using anthropometric data and routine blood analysis parameters. The aim of this study is to apply classification algorithms to data from patients suspected of having breast cancer and

to compare the performance of Artificial Neural Networks (ANN) with Naive Bayes classification.

2. MATERIAL AND METHODS

The overall goal of these research stages is to improve the diagnosis, management, and prevention of HIV/AIDS by developing and validating robust predictive models. By systematically addressing each stage on Figure 2, the research seeks to provide reliable tools and insights that can significantly enhance the quality of HIV/AIDS healthcare and prevention strategies.

FIGURE 2. Graph of HIV/AIDS patients per year

This research process is conducted systematically and comprehensively to achieve optimal results. The study begins with Problem Identification, which involves in-depth analysis to recognize and understand the main issues related to HIV/AIDS. This step includes identifying the problems faced by patients, partners, and housewives.

Next, Data Collection is carried out from various relevant sources. The collected data includes information from patients, partners, and housewives, which is then used for further analysis. Data collection is performed meticulously to ensure the completeness and accuracy of the data obtained. After the data is collected, the next step is Data Cleaning. In this stage, the raw data is examined and cleaned of inconsistencies, duplications, and missing values [11]. This cleaning process is crucial to ensure that the data is ready for accurate analysis.

Then, the cleaned data is used for the Application of SVM and Naive Bayes Algorithms. These algorithms are chosen for their ability to classify and predict with high accuracy [12]. The use of these algorithms aims to build predictive models that can assist in the diagnosis and management of HIV/AIDS. The final stage is Testing the developed models. At this stage, the models are tested using the prepared data to evaluate their performance. The evaluation is done by measuring the accuracy, precision, recall, and F1-Score of the models used. The results of this testing provide insights into the effectiveness of the models in classification and prediction, and serve as a basis for future model improvements. With this systematic and comprehensive approach, the research is expected to make a significant contribution to improving the diagnosis, management, and prevention of HIV/AIDS.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

3. RESULT AND DISCUSSION

3.1. DATA COLLECTION

The obtained dataset consists of 707,379 rows and 89 columns, originating from the merging of several tables. This data was then inspected to ensure its readiness for processing and modeling. The inspection process included identifying and handling missing data, removing duplicates, and validating values to ensure consistency and accuracy before further analysis.

3.2. EXPLORATORY DATA

In this stage, the initial step is to select attributes that do not contribute to this research, followed by checking the remaining data in the dataset. Subsequently, each data tuple for each attribute is examined to identify whether there are any inappropriate values or data duplication. Attributes that have inconsistencies include gender, number of KIE, number of condoms, number of lubricants, number of syringes used, amount of alcohol, marital status, last education, steady partners, non-steady partners, anal sex, vaginal sex, condom use during intercourse, last condom use during intercourse, frequency of syringe use, IMS referral, TB referral results, and HEPC referral results. Additionally, missing or inconsistent data must be addressed to ensure the quality of the dataset. This process is crucial to preparing the data for accurate processing and modeling.

The steps taken to correct inconsistencies in records with these attributes are as follows:

1. Remove data records containing incomprehensible values, such as gender data in Table 1. For instance, if Code #1 is used for Male and Code #2 for Female, but values that do not match the predefined codes, such as Code #3 or unrecognized symbols, are found, those records should be deleted or corrected to ensure data consistency. This process is crucial to avoid errors in data analysis caused by inaccuracies in data entries.

TABLE 1.
Unique gender data

Code Gender	Count of ODHA
1	707.035
2	343
3	1

It is evident that there are records with a value of 3 in the gender column, which does not indicate whether the individual is male or female. Therefore, these records with a value of 3 must be deleted..

2. Replace inappropriate values with correct ones, for example, in this case related to the education level of people living with HIV/AIDS (ODHA) as shown in Figure 3. This step ensures that all data entries are accurate and reflective of the true information, thereby improving the overall quality and reliability of the dataset. Correcting these values is essential to avoid any misinterpretation or errors during data analysis.

```
hiv.klien_pendidikan_terakhir.replace('SD / SEDERAJAT', 'SD', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SMP/SLTP', 'SMP', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SLTP', 'SMP', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SMP / SEDERAJAT', 'SMP', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SMA/SMU/SMK', 'SMA', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SLTA', 'SMA', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SMK', 'SMA', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SMA / SEDERAJAT', 'SMA', inplace=True)
hiv.klien_pendidikan_terakhir.replace('SMU', 'SMA', inplace=True)
hiv.klien_pendidikan_terakhir.replace('STM', 'SMA', inplace=True)
```

FIGURE 3. Change educational data

3. Replace NAN values with 0, depending on the type of data used for the attribute. This step is crucial for maintaining the integrity of the dataset, as it ensures that missing values do not negatively impact the analysis. By filling in NAN values with 0 or another appropriate value, we can prevent potential biases or inaccuracies in the results. This process helps in creating a more complete and usable dataset, facilitating more reliable and meaningful data analysis.
4. Transform data previously in array format, such as the data related to non-steady partners in Table 2, which is currently in an array format. To make this data readable and usable in modeling, it needs to be transformed into columnar format. This transformation is necessary to ensure that the data can be properly analyzed and processed within the modeling framework. Converting array data into a column format allows for more straightforward manipulation and integration into various analytical and predictive models, thereby enhancing the accuracy and effectiveness of the data analysis process.

TABLE 2.
Array data couple of ODHA

Couple	Count of ODHA
1,2	1.959
2	1.951
1,3	242

Based on the presented data, it is evident that there are entries for partners with more than one value. Therefore, the data needs to be transformed into a columnar format to be fully processed. This transformation is crucial to ensure that all entries are appropriately handled and analyzed. By converting the data into a column format, we can streamline the processing and analysis, making it easier to apply various data manipulation techniques and analytical models. The transformation can be accomplished using the following script:

```
hiv['klien_pasangan_tidak_tetap'] = hiv['klien_pasangan_tidak_tetap'].apply(lambda
x: [int(n) for n in x.split(',')])

# Membuat kolom untuk klien_pasangan_tidak_tetap, 4 untuk TIDAK PUNYA hapus
for naza_number in range(1, 4):
    column_name = f'pasangan_tidak_tetap_{naza_number}'
    # Setiap baris diisi dengan 1 jika nomor Naza ada di baris tersebut, jika tidak
diisi dengan 0
    hiv[column_name] = hiv['klien_pasangan_tidak_tetap'].apply(lambda x: 1 if
naza_number in x else 2)
```


ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

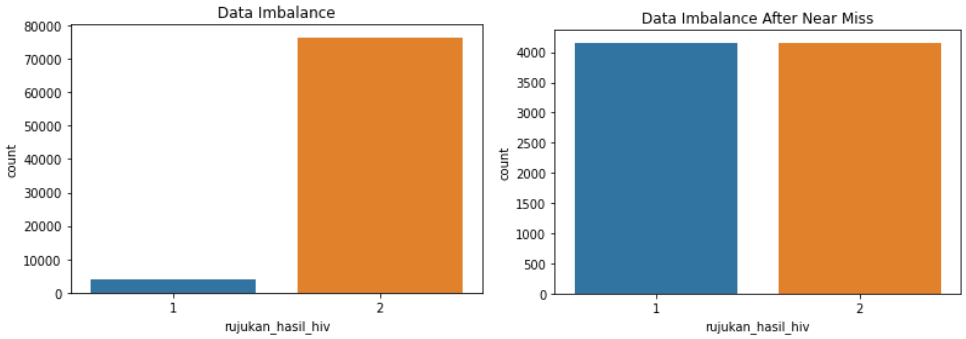
This script ensures that each entry is separated into its own row, making the dataset more manageable and suitable for detailed analysis. Converting array data into individual columns allows for more accurate and efficient data manipulation, thereby improving the overall quality of the dataset and the reliability of the results obtained from subsequent analysis.

5. Remove duplicate data entries to ensure the dataset's integrity and accuracy. Duplicates can skew analysis results and lead to incorrect conclusions, so it is essential to identify and eliminate them. This process involves scanning the dataset for identical rows and removing any repetitions. Ensuring that each data entry is unique improves the reliability and validity of the analysis. Additionally, it helps in maintaining a clean and efficient dataset, which is crucial for accurate data modeling and interpretation.

From the initial dataset, a total of 80,529 records were obtained, which are divided into two classes: reactive and non-reactive. According to the data, there are 4,517 reactive cases and 76,012 non-reactive cases. Given this class distribution, it is clear that the data is still imbalanced, necessitating data balancing. Data imbalance can lead to biased models that perform poorly on the minority class. Therefore, balancing the dataset is crucial to ensure that the classification algorithms can effectively learn and make accurate predictions across both classes. Various techniques, such as oversampling the minority class, undersampling the majority class, or using synthetic data generation methods like NearMiss, can be employed to achieve a balanced dataset. This step is vital to enhance the performance and reliability of the predictive models used in the analysis..

3.3. IMBALANCED DATA

Given the data imbalance in the HIV result reference classes, it is necessary to balance the dataset. The data balancing process is carried out using the NearMiss method, with the data before balancing shown in Figure 4(a) and after balancing shown in Figure 4(b). In Figure 4(a), the proportion of reactive HIV cases is 5.16% and non-reactive cases is 94.84%. After data balancing, as shown in Figure 4(b), the proportions become equal at 50%. Balancing the dataset is crucial to ensure that the classification algorithms do not favor the majority class and can accurately predict both reactive and non-reactive cases. The NearMiss method is particularly effective in handling imbalanced data by selecting a subset of the majority class that is most similar to the minority class, thereby reducing the imbalance. This approach helps improve the performance and reliability of the predictive models, ensuring that they can generalize well to new, unseen data. By achieving a balanced dataset, we can enhance the robustness of the analysis and obtain more accurate and meaningful results.



(a) (b)

FIGURE 4. Data balancing using NearMiss (a) before balancing, (b) after balancing

3.4. CLASSIFICATION

This process involves splitting the data into ratios of 60:40, 70:30, 80:20, and 90:10, utilizing multiple algorithms and k-fold cross-validation to measure the distance between values. The results for accuracy, precision, recall, and F1 score for the Naïve Bayes algorithm can be seen in Table 3, while the results for the SVM algorithm can be seen in Table 4.

TABLE 3.
Result using Naïve Bayes

Fold	60:40 (dalam %)				70:30 (dalam %)				80:20 (dalam %)				90:10 (dalam %)			
	Accuracy	Precision	Recall	F1Score	Accu-racy	Preci-sion	Re-call	F1-Score	Accu-racy	Preci-sion	Re-call	F1-Score	Accu-racy	Preci-sion	Re-call	F1-Score
1	56,5	76,8	56,5	46,4	58,8	75,6	58,8	50,7	59,0	76,7	59,0	50,9	59,4	76,9	59,4	51,5
2	58,5	77,4	58,5	50,0	58,8	77,4	58,8	50,3	57,4	76,0	57,4	48,2	57,0	74,4	57,0	47,7
3	55,9	75,0	55,9	45,6	59,5	77,6	59,5	51,5	58,8	77,4	58,8	50,3	59,4	76,3	59,4	51,6
4	56,7	76,8	56,7	46,8	58,8	75,6	58,8	50,7	58,2	74,0	58,2	49,9	58,0	74,3	58,0	49,6
5	61,1	78,1	61,1	54,2	56,5	75,5	56,5	46,6	59,9	77,0	59,9	52,2	58,3	76,5	58,3	49,6
6	57,7	74,7	57,7	48,9	57,6	74,1	57,6	48,8	58,2	75,5	58,2	49,6	60,0	77,8	60,0	52,4
7	58,9	77,4	58,9	50,5	57,6	77,0	57,6	48,2	57,7	76,2	57,7	48,8	58,7	75,9	58,7	50,5
8	56,3	73,8	56,3	46,4	59,6	77,7	59,6	51,8	59,1	76,0	59,1	51,2	58,0	76,4	58,0	49,2
9	59,8	75,0	59,8	52,7	58,9	77,5	58,9	50,6	58,4	76,4	58,4	49,8	59,2	76,2	59,2	51,4
10	59,4	76,6	59,4	51,6	57,8	74,2	57,8	49,2	58,4	76,4	58,4	49,8	58,2	77,2	58,2	49,3

TABLE 4.
Result using SVM

Fold	60:40 (dalam %)				70:30 (dalam %)				80:20 (dalam %)				90:10 (dalam %)			
	Accuracy	Precision	Recall	F1Score	Accu-racy	Preci-sion	Re-call	F1-Score	Accu-racy	Preci-sion	Re-call	F1-Score	Accu-racy	Preci-sion	Re-call	F1-Score
1	73,0	73,2	73,0	72,9	81,4	81,6	81,4	81,4	78,7	78,8	78,7	78,7	79,2	79,5	79,2	79,1
2	79,4	79,9	79,4	79,3	77,0	77,1	77,0	76,9	76,5	76,8	76,5	76,5	79,2	79,7	79,2	79,1
3	78,2	78,8	78,2	78,0	78,4	78,4	78,4	78,3	77,7	77,9	77,7	77,7	77,9	78,4	77,9	77,9
4	79,4	80,0	79,4	79,3	79,4	79,6	79,4	79,4	79,0	79,3	79,0	78,9	78,9	79,4	78,9	78,8
5	82,2	82,4	82,2	82,1	78,7	79,4	78,7	78,6	78,1	78,2	78,1	78,0	77,8	78,2	77,8	77,7
6	78,0	78,0	78,0	77,9	75,4	75,9	75,4	75,3	80,0	80,5	80,0	79,9	82,6	83,0	82,6	82,6
7	77,2	77,3	77,2	77,1	75,3	76,1	75,3	75,1	80,3	80,7	80,3	80,2	77,4	77,9	77,4	77,3
8	76,6	77,0	76,6	76,5	81,3	81,5	81,3	81,2	79,7	80,0	79,7	79,7	76,6	77,1	76,6	76,5
9	78,3	78,5	78,3	78,3	78,4	78,7	78,4	78,3	76,7	77,2	76,7	76,6	77,3	77,7	77,3	77,2
10	79,7	79,9	79,7	79,7	79,0	79,1	79,0	79,0	77,1	77,7	77,1	77,0	77,7	78,2	77,7	77,6

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Based on the displayed data, the results of the k-fold values can be visualized as shown in Figure 5.

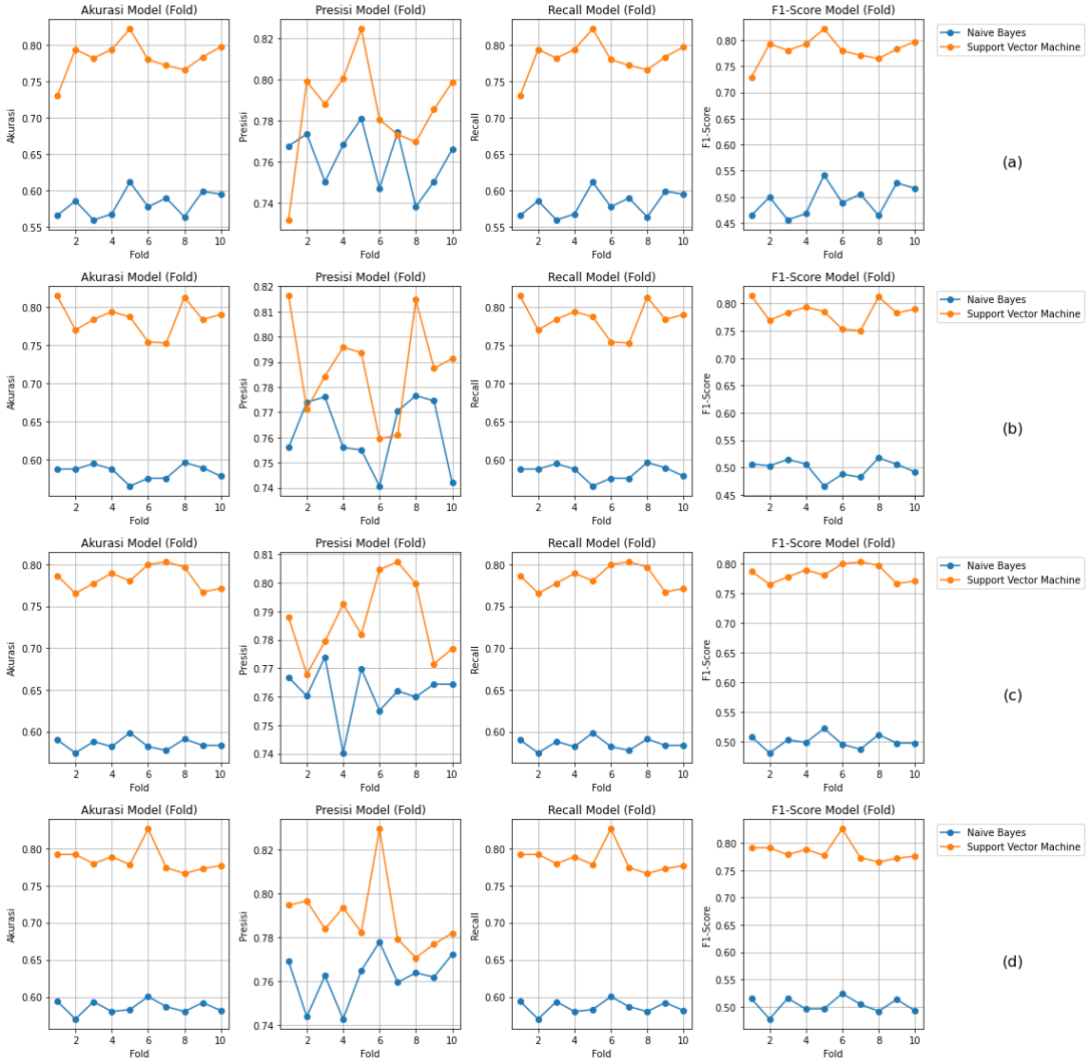


FIGURE 5. Visualization Fold (a) 60:40, (b) 70:30, (c) 80:20, (d) 90:10

From the visualizations provided, it is clear that the SVM algorithm outperforms the Naïve Bayes algorithm in terms of accuracy, precision, recall, and F1-Score. The superiority of SVM in these metrics indicates its effectiveness in handling the dataset and making accurate predictions. This suggests that SVM is a more robust and reliable model for this particular classification task, highlighting its capability to better capture the underlying patterns and relationships within the data compared to Naïve Bayes.

4. CONCLUSION

This study employs the Naïve Bayes and SVM algorithms for classification to develop predictive models with high accuracy. The data, with selected features, is classified using these algorithms, resulting in the highest accuracy when using the SVM model. The SVM model achieved its highest accuracy of 82.6% with a 90:10 data split and a 6-fold value, and an accuracy of 82.2% with a 60:40 data split and a

5-fold value. In contrast, the Naïve Bayes algorithm attained its highest accuracy of 61.1% with 60:40 data splits. These results indicate that the SVM model is significantly more effective for this classification task compared to the Naïve Bayes, demonstrating its superior capability in capturing and predicting the patterns within the dataset.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Universitas Mercu Buana through LPPM for funding the research we conducted.

REFERENCES

[1] I. Expat, "Bali Enters Top 10 Most HIV Cases In Indonesia," *https://indonesiaexpat.id/news/bali-enters-top-10-most-hiv-cases-in-indonesia/*, 2022. .

[2] UNAIDS, "HIV data and statistics," *https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics*, 2023. .

[3] T. G. Hailu, "Comparing Data Mining Techniques in HIV Testing Prediction," *Intell. Inf. Manag.*, vol. 07, no. 03, pp. 153–180, 2015, doi: 10.4236/iim.2015.73014.

[4] A. Tohari, N. Chamidah, and Fatmawati, "Modelling of hiv and aids cases in indonesia using bi-response negative binomial regression approach based on local linear estimator," *Ann. Biol.*, vol. 36, no. 2, pp. 215–219, 2020.

[5] N. H. Sweilam, S. M. AL-Mekhlafi, Z. N. Mohammed, and D. Baleanu, "Optimal control for variable order fractional HIV/AIDS and malaria mathematical models with multi-time delay," *Alexandria Eng. J.*, vol. 59, no. 5, pp. 3149–3162, 2020, doi: 10.1016/j.aej.2020.07.021.

[6] A. Ramachandran *et al.*, "Predictive Analytics for Retention in Care in an Urban HIV Clinic," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-62729-x.

[7] S. S. Mukerji *et al.*, "Machine Learning Approaches to Understand Cognitive Phenotypes in People With HIV," *J. Infect. Dis.*, vol. 227, no. Suppl 1, pp. S48–S57, 2023, doi: 10.1093/infdis/jiac293.

[8] S. Saravanakumar, A. Eswari, L. Rajendran, and M. Abukhaled, "A Mathematical Model of Risk Factors in HIV/AIDS Transmission Dynamics: Observational Study of Female Sexual Network in India," *Appl. Math. Inf. Sci.*, vol. 14, no. 6, pp. 967–976, 2020, doi: 10.18576/amis/140603.

[9] C. K. Mutai, P. E. McSharry, I. Ngaruye, and E. Musabanganji, "Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa," *BMC Med. Res. Methodol.*, vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12874-021-01346-2.

[10] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Meas. Sensors*, vol. 25, no. November 2022, p. 100605, 2023, doi:

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

10.1016/j.measen.2022.100605.

[11] S. Sandiwarno, Z. Niu, and A. S. Nyamawe, "SES-Net: A Novel Multi-Task Deep Neural Network Model for Analyzing E-learning Users' Satisfaction via Sentiment, Emotion, and Semantic," *Int. J. Hum. Comput. Interact.*, vol. 0, no. 0, pp. 1–24, 2024, doi: 10.1080/10447318.2024.2356356.

[12] W. Gunawan, R. A. Wiradiputra, P. Sari, D. Prayama, and R. Nainggolan, "Prediction of Cross-Platform and Native Apps Technology Opportunities for Beginner Developers Using C 4.5 and Naïve Bayes Algorithms," vol. 7, no. December, 2023, [Online]. Available: www.joiv.org/index.php/joiv.