BAREKENG: Journal of Mathematics and Its Applications
September 2024 Volume 18 Issue 3 Page 1829–1838
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss3pp1829-1838

APPLICATION OF ADASYN OVERSAMPLING TECHNIQUE ON K-NEAREST NEIGHBOR ALGORITHM

Herina Marlisa [1], Neva Satyahadewi [2*], Nurfitri Imro'ah [3], Naomi Nessyana Debataraja [4]

[1,2,3,4]Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak, 78124, Indonesia

*Corresponding author's e-mail: * **neva.satya@math.untan.ac.id***

ABSTRACT

Article History:

Received: 15th February 2024
Revised: 7th April 2024
Accepted: 6th July 2024
Published: 1st September 2024

Keywords:

Adaptive Synthetic;
Classification;
Imbalanced Data;
Diabetes Mellitus.

The K-Nearest Neighbor Algorithm is a commonly used data mining algorithm for classification due to its effectiveness with large datasets and noise. However, class imbalance may impact classification results, where data with unbalanced classes may classify new data based on the majority class and ignore minority class data. The research analyzed whether applying the Adaptive Synthetic (ADASYN) oversampling technique in the K-Nearest Neighbor Algorithm can handle data imbalance problems. The study looks at the resulting accuracy, specificity, and sensitivity values. ADASYN oversamples the minority class data based on the model's difficulty level of data learning using distribution weights. This research uses the Pima Indian Diabetes Dataset from the Kaggle website. The dependent variable was diabetes mellitus status, while the independent variables were number of pregnancies, glucose levels, diastolic blood pressure, insulin levels, Body Mass Index (BMI), and age. The study found that the accuracy, specificity, and sensitivity values were 72.88%, 73.42%, and 71.79%, respectively. Based on the results of the analysis, it can be concluded that using ADASYN in the K-Nearest Neighbor Algorithm to classify diabetes mellitus in Pima Indian women is good enough to address imbalanced data. It is shown that the ADASYN oversampling technique can help the K-Nearest Neighbor Algorithm to classify new data without ignoring the data of the minority class.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

H. Marlisa, N. Satyahadewi, N. Imro'ah and N. N. Debataraja., "APPLICATION OF ADASYN OVERSAMPLING TECHNIQUE ON K-NEAREST NEIGHBOR ALGORITHM," *BAREKENG: J. Math. & App.*, vol. 18, iss. 3, pp. 1829-1838, September, 2024.

Copyright © 2024 Author(s)
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

Data mining uses statistical techniques to identify patterns and extract valuable insights from large datasets [1]. One of the most used techniques in data mining is classification. Classification is often used to make decisions using an algorithm based on new information derived from previous data processing [2]. Good classification results depend on the algorithm used in the classification process. There are several algorithms for the classification process, including the K-Nearest Neighbor Algorithm, Support Vector Machine, Decision Tree, Naïve Bayes, etc. The K-Nearest Neighbor algorithm is often used in the classification process because it is effective for extensive and noisy data [3]. The K-Nearest Neighbor algorithm is also quite simple because no assumptions about data distribution must be met, and it is easy to implement [4].

In addition to the classification algorithm used, class imbalance can also affect classification results. Data with unbalanced classes tend to classify new data based on the majority class and ignore minority class data so that classification performance is less than optimal [5]. These problems can be overcome through data balancing before classification, namely by generating synthetic data on minority classes using the Adaptive Synthetic (ADASYN) oversampling technique. ADASYN is a technique to overcome the problem of unbalanced data by oversampling minority class data based on the model's difficulty level of data learning using distribution weights [6]. The application of ADASYN in the classification process can be utilized in various fields of life, including the health sector. There are many records in the health sector regarding patient examination results from various diseases that can be utilized to find helpful information, including diabetes mellitus disease data.

Diabetes mellitus is a chronic disease characterized by high blood glucose levels due to the body's inability to produce or use insulin optimally [7]. Diabetes mellitus is often caused by age, excess weight, heredity, and lack of physical activity [8]. Many people are late to realize that they have diabetes mellitus because the initial symptoms of this disease are often not visible, so the disease is detected after complications occur. According to the International Diabetes Federation (IDF), the number of people with diabetes mellitus in every country continues to increase and has caused the death of 6.7 million adults in 2021 [9].

The high mortality rate due to diabetes mellitus can be triggered by late diagnosis in patients due to the limited number of medical personnel, especially in small cities [10]. In addition, several clinical trial processes that need to be carried out to determine whether someone has diabetes mellitus also take a long time [11]. Therefore, patient data collected previously can be utilized by processing it using the classification function in data mining to obtain information to support decisions in identifying diabetes mellitus. A large amount of medical record data tends to be prone to class imbalance, so the ADASYN oversampling technique is needed to improve the classification algorithm's accuracy.

In 2021, Ramadhan conducted research by comparing the SVM, SMOTE-SVM, and ADASYN-SVM methods in classifying diabetes mellitus type 2 [12]. The data used is secondary data obtained from the Karya Medika Clinic, Indonesia. In this case, the ADASYN-SVM oversampling technique is superior and can solve the problem of class imbalance with an accuracy of 87.3%. Durahim also conducted research by comparing several oversampling and undersampling techniques in 2016 [13]. The study concluded that the ADASYN oversampling technique is the best choice among several existing sampling techniques, considering execution time, increased data size, and classification performance. In the 2019 research of Hasmita, Nhita, Saepudin, and Aditsania, ADASYN was applied to predict chili prices in the Bandung Regency with the K-Nearest Neighbor Algorithm [14]. The accuracy value obtained before oversampling was 93% with an F1-Score value of 0%, while after applying ADASYN, the accuracy value and F1-Score reached 100%. The results showed that using ADASYN on unbalanced data proved helpful and played an essential role in the classification process.

Therefore, research was conducted to overcome the class imbalance problem in the case study of determining diabetes mellitus status in Pima Indian women with the ADASYN oversampling technique and the K-Nearest Neighbor classification algorithm. The research is expected to overcome the problem of unbalanced data so that the classification results become more accurate and can become a decision support system in identifying diabetes mellitus, especially in the health sector.

2. RESEARCH METHODS

Data mining uses statistical techniques to identify patterns and extract valuable insights from large datasets. The six data mining functions are description, estimation, prediction, classification, clustering, and association [15]. Classification involves determining the class of an object based on previous data processing using an algorithm [2]. However, class imbalance can impact classification results, where one class has more cases than others. To solve this problem, it is necessary to balance the data by generating synthetic data on the minority classes using the Adaptive Synthetic (ADASYN) oversampling technique before classifying.

The data used in this study is secondary data, namely the Pima Indian Diabetes Dataset, which was obtained from Kaggle with the web address https://www.kaggle.com/datasets/mathchi/diabetes-data-set. The research data are health examination data of 768 Pima Indian women from the city of Phoenix, Arizona, United States.

2.1 Adaptive Synthetic (ADASYN) Oversampling Technique

Adaptive Synthetic (ADASYN) is a technique to overcome the problem of unbalanced data by oversampling the minority class data based on the difficulty level of data learning by the model using distribution weights [6]. The main idea of the ADASYN technique is to use distribution weights for minority class data based on the difficulty level of understanding. ADASYN can reduce bias in unbalanced data and is relatively easy to implement [16]. The stages of synthetic data generation with ADASYN are [17]:

a. Determine the ADASYN parameter values, namely d_{th} and β. The d_{th} value is the maximum tolerance limit of class imbalance with $d_{th} \in [0,1]$. The β is the expected level of class balance after synthetic data generation with $\beta \in [0\,,1]$. $\beta = 1$ describes that the data is fully balanced after the generation process.

b. Calculate the degree of class balance using Equation (1):

$$d = \frac{m_s}{m_l} \tag{1}$$

where m_s is the number of minority class data and m_l is the number of majority class data. If the condition $d < d_{th}$, then the class imbalance in the research data cannot be tolerated and needs to be balanced so that the ADASYN process can continue to the next stage. If $d > d_{th}$, then the class imbalance in the research data can still be tolerated or is already balanced, so there is no need to balance the data with ADASYN.

c. Calculate the amount of minority class synthetic data that needs to be generated with Equation (2):

$$G = (m_l - m_s)(\beta) \tag{2}$$

d. Find the K-Nearest Neighbor of each minority class data based on Euclidean Distance and calculate the ratio r_i using Equation (3):

$$r_i = \frac{\Delta_i}{K}, i = 1,2, \dots, m_s \tag{3}$$

where K is the neighboring value and Δ_i is the amount of data in the K-Nearest Neighbor that does not belong to the minority class.

e. Normalize r_i, so that the value of $\hat{r}_i$ is a density distribution using Equation (4):

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{4}$$

f. Calculate the number of synthetic data that needs to be generated from each minority class data using Equation (5):

$$g_i = (\hat{r}_i)(G) \tag{5}$$

where G is the total synthetic data that needs to be generated for the minority class. G and g_i are non-negative integers.

g. Generate synthetic data samples with Equation (6):

$$\boldsymbol{s_i} = \boldsymbol{x_i} + (\boldsymbol{x_{zi}} - \boldsymbol{x_i})(\lambda) \tag{6}$$

where $\boldsymbol{s_i}$ is the synthetic data calculated by ADASYN, $\boldsymbol{x_i}$ is the minority class data that needs to be generated, $\boldsymbol{x_{zi}}$ is the minority class data in the immediate neighbor or the same neighborhood, dan λ is a random number between 0 and 1.

2.2 K-Nearest Neighbor Algorithm

K-Nearest Neighbor Algorithm is used to classify an object based on the training data class with the closest distance [18]. The K-Nearest Neighbor Algorithm is relatively simple because no assumptions about data distribution must be met, and it is easy to implement [4].The steps in the K-Nearest Neighbor Algorithm are [19]:

a. Preparing training data and testing data.

b. Calculate the distance between training and testing data based on Euclidean Distance with Equation (7):

$$D(x_i, y_i) = \sqrt{\sum_{j=1}^{n} \left(x_{i,j} - y_{i,j}\right)^2} \tag{7}$$

where $x_{i,j}$ is the i-th training data of the j-th attribute, $y_{i,j}$ is the i-th testing data of the j-th attribute, and n is the number of independent attributes.

c. Sort the distances that have been obtained from most minor to most significant.

d. Determine the value of the parameter K (number of nearest neighbors).

e. Check the class of the K nearest training data and classify the testing data based on the majority class of the K nearest neighbors.

2.3 Classification Performance Evaluation

Classification performance can be evaluated using a confusion matrix. The confusion matrix is a table that displays how many correct and incorrect classification algorithms exist [20]. Table 1 presents a confusion matrix with two classes on the dependent attribute.

Table 1. Confusion Matrix

Confusion Matrix		Actual	
		Positive	**Negative**
Prediction	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Based on the TP, FP, TN, and FN values, the accuracy, specificity, and sensitivity values can be obtained, which are used to determine the classification performance with the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{10}$$

Accuracy is a value used to determine how accurate an algorithm is in performing classification. Specificity is the value used to determine how successful an algorithm is in classifying antagonistic classes.

Sensitivity is a value used to evaluate how successful an algorithm is in correctly classifying positive classes. The values obtained from the confusion matrix calculation results can be classified into five categories shown in **Table 2 [21]**.

Table 2. Classification Performance Goodness Level

Value (%)	Category
90.01 - 100.00	Very good
80.01 - 90.00	Good
70.01 - 80.00	Good Enough
60.01 - 70.00	Bad
≤ 60.00	Failed

3. RESULTS AND DISCUSSION

There are two processes to go through in the preprocessing phase, namely the attribute selection and data cleaning phases. The attribute selection process is based on the factors that influence an individual to be diagnosed with diabetes mellitus, according to the American Diabetes Association.

Table 3. Research Attributes

Attributes	Description
Status	Diagnosis of the patient after examination, Diabetic or Non-Diabetic
Pregnancies	The number of pregnancies the patient has had
Glucose	Patient's blood sugar level two hours postprandial or two hours after a meal
Diastolic	Blood pressure when the heart relaxes or receives blood back from the body
Insulin	Insulin levels in the body, a hormone that plays a role in controlling proper glucose levels in the blood
BMI	Body Mass Index
Age	Patient's age at the time of examination

Table 3 describes all data attributes used with diabetes mellitus status as the dependent attribute. At the same time, the rest are the number of pregnancies, glucose levels, diastolic blood pressure, insulin levels, BMI, and age as independent attributes. Furthermore, data cleaning was carried out by eliminating data that had missing values, so that the research dataset which initially amounted to 768 data was reduced to 392 data.

3.1 Descriptive Statistics

Descriptive statistics aimed to obtain an overview of the health examination data from 392 Pima Indian women.

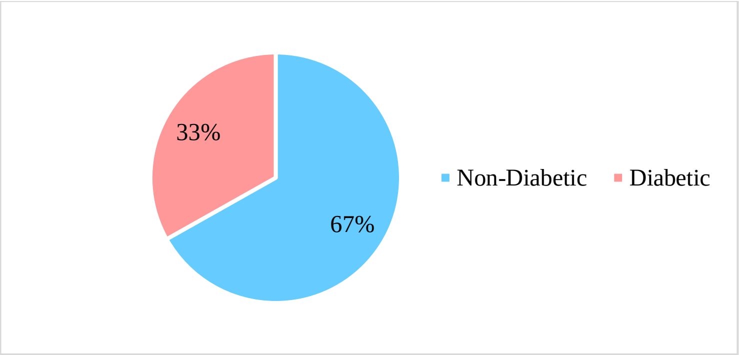

Figure 1. Pie Chart of Diabetes Mellitus Status

Figure 1 shows the percentage of diabetes mellitus status of the 392 Pima Indian women who underwent health screening. Based on the figure, it is known that 33% or 130 people of the total Pima Indian women were identified as having diabetes mellitus. Meanwhile, 67% or 262 people are categorized as not having diabetes mellitus. It indicates a class imbalance in the research data.

Table 4. Descriptive Statistics of Independent Attributes

Attributes	Average
Pregnancies	3.00
Glucose (mg/dL)	122.63
Diastolic (mmHg)	70.66
Insulin (µU/mL)	156.06
BMI (kg/m^2)	33.09
Age (Years)	30.86

Table 4 displays the results of a health check-up conducted on Pima Indian women. The average age of the patients was 30.86 years, with an age range of 21 to 81 years. Most patients were overweight, as indicated by an average BMI of 33.09 kg/m^2, greater than 25 kg/m^2. The average insulin level of patients in two hours after eating was 156.0 **µ**6 U/mL, which is within the normal range as it does not exceed 200 **µ**U/mL.

3.2 Splitting Data

The division of training and testing data is done randomly with the help of R-Studio software. The division of data training and testing is presented in Table 5.

Table 5. Comparison of Training and Testing Data

Description	Training	Testing	Total
Diabetic Class	91	39	130
Non-diabetic Class	183	79	262
Total	274	118	392
Percentage	70%	30%	100%

Table 5 shows that this training data is not balanced. Therefore, it is necessary to balance the data first to maximize the classification results.

3.3 Application of Adaptive Synthetic (ADASYN)

This research uses the ADASYN parameter value $d_{th} = 0.75$ [6], which means the maximum tolerance limit of class imbalance is 0.75, and $\beta = 1$ explains that the new training data formed is expected to be fully balanced. The ADASYN calculation for the training data starts by calculating the degree of class balance with Equation (1):

$$d = \frac{m_s}{m_l} = \frac{91}{183} = 0.50$$

The calculation results show the condition $d < d_{th}$, which means that the class imbalance in the research data cannot be tolerated so that the ADASYN process can proceed to the next stage. Then, calculate the amount of synthetic data for the minority class to be formed using Equation (2) with $\beta = 1$.

$$G = (m_l - m_s)(\beta) = (183 - 91)(1) = 92$$

Based on these calculations, there are around 92 synthetic data with minority classes to be formed. Next, calculate the ratio based on K-Nearest Neighbor using Euclidean Distance for each data in the minority class. The nearest neighbor evaluation on the second data (x_2) with $K = 5$ is presented in **Table 6**.

Table 6. Five Nearest Neighbors for the Second Data

Data	Class	Distance
x_{121}	Diabetic	19.16
x_{127}	Non-Diabetic	23.36
x_{144}	Diabetic	24.86
x_{79}	Diabetic	27.93
x_{183}	Non-Diabetic	28.43

Table 6 shows that there are two data with majority classes, so $\Delta_2 = 2$. Therefore, the ratio of data x_2 can be calculated with **Equation (3)**.

$$r_2 = \frac{\Delta_2}{K} = \frac{2}{5} = 0.40$$

In the same way, the ratio value for each minority data is obtained, as shown in **Table 7**.

Table 7. The Ratio Value for Each Minority Class Data

No.	Data	Nearest Data	Δ_i	r_i
1	x_2	$x_{121}, x_{127}, x_{144}, x_{79}, x_{183}$	2	0.40
2	x_3	$x_{93}, x_{141}, x_{239}, x_{32}, x_{72}$	1	0.20
⋮	⋮	⋮	⋮	⋮
90	x_{272}	$x_{224}, x_{149}, x_{260}, x_{78}, x_{217}$	0	0.00
91	x_{274}	$x_{101}, x_{91}, x_{265}, x_{169}, x_{140}$	4	0.80

Next, normalize r_i using **Equation (4)**. The following is the calculation of $\hat{r}_i$ for the second data.

$$\hat{r}_2 = \frac{r_2}{\sum_{i=1}^{91} r_i} = \frac{0.40}{0.40 + 0.20 + \cdots + 0 + 0.80} = \frac{0.40}{46.60} = 0.009$$

The next step is to calculate the amount of synthetic data that needs to be generated from each minority class data with **Equation (5)**. The following is the calculation of g_i for the second data.

$$g_2 = (\hat{r}_2)(G) = (0.009)(92) = 0.828 \approx 1$$

The calculation results show that one synthetic data with the minority class of having diabetes mellitus was generated from the second data. **Table 8** presents the values of $\hat{r}_i$ and g_i for each minority class data obtained by the same calculation.

Table 8. Number of Synthetic Data Duplications

Data	$\hat{r}_i$	g_i
x_2	0.009	1
x_3	0.004	0
⋮	⋮	⋮
x_{272}	0.000	0
x_{274}	0.017	2
Total		**94**

Next, generate synthetic data samples using **Equation (6)**. The following is an example of the calculation of the second data with $\lambda = 0.80$ and $\boldsymbol{x_{zi}} = \boldsymbol{x_{121}}$, which is the closest minority class data to the second data.

$$\boldsymbol{s_1} = \boldsymbol{x_2} + (\boldsymbol{x_{121}} - \boldsymbol{x_2})(\lambda)$$
$$= \begin{bmatrix} 0 \\ 137 \\ 40 \\ 168 \\ 43.1 \\ 33 \end{bmatrix} + \left(\begin{bmatrix} 1 \\ 125 \\ 50 \\ 167 \\ 33.3 \\ 28 \end{bmatrix} - \begin{bmatrix} 0 \\ 137 \\ 40 \\ 168 \\ 43.1 \\ 33 \end{bmatrix} \right)(0.80)$$
$$= \begin{bmatrix} 1 \\ 127 \\ 48 \\ 167 \\ 35.3 \\ 29 \end{bmatrix}$$

The calculation results show that the synthetic data generated is class data with diabetes mellitus with $s_{1,1} = 1$, $s_{1,2} = 127$, $s_{1,3} = 48$, $s_{1,4} = 167$, $s_{1,5} = 35.3$, and $s_{1,6} = 29$. Other synthetic data are generated in the same manner as the number of g_i obtained and added to the initial training data to generate new ADASYN training data. The number of diabetic and non-diabetic classes in the balanced training data is ready to be used in the classification process.

3.4 Application of K-Nearest Neighbor Algorithm

This research uses 368 training data derived from the application of ADASYN and 118 testing data. The first stage calculates the distance between training and testing data based on Euclidean Distance with **Equation (7)**. The following is an example of calculating the distance between the first training data and testing data.

$$D(x_1, y_1) = \sqrt{\sum_{j=1}^{6} \left(x_{1,j} - y_{1,j}\right)^2}$$
$$= \sqrt{\begin{matrix}(1-3)^2 + (89-78)^2 + (66-50)^2 + (94-88)^2 \\ +(28.1-31)^2 + (21-26)^2\end{matrix}}$$
$$= 21.22$$

The distance calculation results are sorted from the value with the smallest (closest) distance to the largest distance. The next stage is to determine the value of K nearest neighbors used. Then, an examination is carried out in each class based on the K nearest neighbors. In this study, the K value used are $K = 1, 3, 5, 7, 9$, so the 10 closest data to the first test data are presented in **Table 9.**

Table 9. Descriptive Statistics of Independent Attributes

Data	Euclidean Distance	Ranking	Status
184	14.37	1	Non-Diabetic
26	18.15	2	Non-Diabetic
159	19.24	3	Non-Diabetic
44	19.27	4	Non-Diabetic
128	20.21	5	Non-Diabetic
1	21.22	6	Non-Diabetic
186	22.23	7	Non-Diabetic

Data	Euclidean Distance	Ranking	Status
124	22.61	8	Non-Diabetic
52	23.41	9	Non-Diabetic
67	23.56	10	Non-Diabetic

Based on Table 9, the class that mostly appears when $K = 1$, $K = 3$, $K = 5$, $K = 7$, or $K = 9$ is the non-diabetic class, so the classification result for the first test data is classified as non-diabetic. The same method is applied to all test data to obtain classification results.

This study uses 118 test data with parameters $K = 1, 3, 5, 7, 9$. Classification performance was evaluated with a confusion matrix. The K value that produces the highest accuracy, specificity, and sensitivity is set as the neighboring value used in the classification process. The confusion matrix for $K = 3$ with the K-Nearest Neighbor Algorithm is presented in Table 10.

Table 10. Confusion Matrix of Classification Results

Confusion Matrix		Actual	
		Diabetic	Non-Diabetic
Prediction	Diabetic	28	21
	Non-Diabetic	11	58

Based on these results, the accuracy value of diabetes mellitus disease classification with K = 3 is 72.88%, specificity is 73.42%, and sensitivity is 71.79%. The same calculation is applied to all classification results for each K parameter value so that a comparison of the accuracy, specificity, and sensitivity results of the K-Nearest Neighbor Algorithm based on the $K = 1, 3, 5, 7, 9$ parameter values are shown in Table 11.

Table 11. Accuracy, Specificity, and Sensitivity of K-NN Parameters

Parameter K	Accuracy (%)	Specificity (%)	Sensitivity (%)
$K = 1$	72.03	**77.22**	61.54
$K = 3$	**72.88**	73.42	**71.79**
$K = 5$	66.10	63.29	**71.79**
$K = 7$	64.41	62.03	69.23
$K = 9$	65.25	63.29	69.23

Based on Table 11, the value of the parameter $K = 3$ produces the best accuracy and sensitivity, which are 72.88% and 71.79%. The resulting specificity value is also strong, ranking second only to the specificity value of $K = 1$, which is 73.42%. Therefore, we conclude that the optimal neighboring value for the K-Nearest Neighbor Algorithm in the diabetes mellitus classification case study of Pima Indian women is $K = 3$. The results of the calculation are described in Table 12 below.

Table 12. Interpretation

	Value	Description	Interpretation
Accuracy	72.88%	Good Enough	Application of ADASYN in the K-Nearest Neighbor Algorithm achieved a 72.88% success rate in correctly classifying all tested subjects.
Specificity	73.42%	Good Enough	Application of ADASYN in the K-Nearest Neighbor Algorithm can predict the class of not having diabetes mellitus in Pima Indian women who do not have diabetes mellitus by 73.42%.
Sensitivity	71.79%	Good Enough	Application of ADASYN in the K-Nearest Neighbor Algorithm can predict the class of diabetes mellitus in Pima Indian women who suffer from diabetes mellitus by 71.79%.

4. CONCLUSIONS

The research results suggest that using ADASYN in the K-Nearest Neighbor Algorithm is good enough to overcome imbalanced data in classifying diabetes mellitus in Pima Indian women. The accuracy, specificity, and sensitivity values are 72.88%, 73.42%, and 71.79%, respectively. This indicates that the ADASYN oversampling technique assists the K-Nearest Neighbor Algorithm in classifying new data without disregarding minority class data.

REFERENCES

[1] M. A. Muslim *et al.*, *Data Mining Algoritma C4.5*. Semarang: UNNES Repository, 2019. Accessed: May 15, 2023. [Online]. Available: http://lib.unnes.ac.id/id/eprint/33080

[2] Indrayanti, D. Sugianti, and M. A. Al Karomi, “Optimasi Parameter K pada Algoritma K-Nearest Neighbor untuk Klasifikasi Penyakit Diabetes Melitus,” *Prosiding SNATIF*, pp. 823–829, 2017, Accessed: Nov. 02, 2023. [Online]. Available: https://jurnal.umk.ac.id/index.php/SNA/article/view/1456

[3] N. M. Putry and B. N. Sari, “Komparasi Algoritma KNN dan Naïve Bayes untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus,” *EVOLUSI : Jurnal Sains dan Manajemen*, vol. 10, no. 1, Sep. 2022, doi: 10.31294/evolusi.v10i1.12514.

[4] F. Yunita, “Sistem Klasifikasi Penyakit Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-NN),” *Selodang Mayang: Jurnal Ilmiah Badan Perencanaan Pembangunan Daerah Kabupaten Indragiri Hilir*, vol. 2, no. 1, pp. 223–230, Apr. 2016, doi: https://doi.org/10.47521/selodangmayang.v2i1.10.

[5] T. Tanti, P. Sirait, and A. Andri, “Optimalisasi Kinerja Klasifikasi Melalui Seleksi Fitur dan AdaBoost dalam Penanganan Ketidakseimbangan Kelas,” *Jurnal Media Informatika Budidarma*, vol. 5, no. 4, p. 1377, Oct. 2021, doi: 10.30865/mib.v5i4.3280.

[6] D. V. Ramadhanti, R. Santoso, and T. Widiharih, “Perbandingan SMOTE dan ADASYN pada Data Imbalance untuk Klasifikasi Rumah Tangga Miskin di Kabupaten Temanggung dengan Algoritma K-Nearest Neighbor,” *Jurnal Gaussian*, vol. 11, no. 4, pp. 499–505, Feb. 2022, doi: 10.14710/j.gauss.11.4.499-505.

[7] World Health Organization, *Classification of Diabetes Mellitus*. 2019. Accessed: Nov. 01, 2023. [Online]. Available: https://www.who.int/publications/i/item/classification-of-diabetes-mellitus

[8] American Diabetes Association, “Standards of Medical Care in Diabetes—2015,” *Diabetes Care*, vol. 38, no. Supplement_1, pp. S4–S4, Jan. 2015, doi: 10.2337/dc15-S003.

[9] International Diabetes Federation, *IDF Diabetes Atlas, 10th edn. Brussels*, 10th ed. Belgium: International Diabetes Federation, 2021. [Online]. Available: www.diabetesatlas.org

[10] U. I. Lestari, A. Y. Nadhiroh, and C. Novia, “Penerapan Metode K-Nearest Neighbor Untuk Sistem Pendukung Keputusan Identifikasi Penyakit Diabetes Melitus,” *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 8, no. 4, pp. 2071–2082, Dec. 2021, doi: 10.35957/jatisi.v8i4.1235.

[11] R. R. Santoso, R. Megasari, and Y. A. Hambali, “Implementasi Metode Machine Learning Menggunakan Algoritma Evolving Artificial Neural Network Pada Kasus Prediksi Diagnosis Diabetes,” *JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer*, vol. 3, no. 2, pp. 85–97, Sep. 2020, Accessed: Nov. 07, 2023. [Online]. Available: https://ejournal.upi.edu/index.php/JATIKOM/article/view/27885

[12] N. G. Ramadhan, “Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus,” *Scientific Journal of Informatics*, vol. 8, no. 2, pp. 276–282, Nov. 2021, doi: 10.15294/sji.v8i2.32484.

[13] A. O. Durahim, “Comparison of Sampling Techniques for Imbalanced Learning,” *Yönetim Bilişim Sistemleri Dergisi*, vol. 2, no. 2, pp. 181–191, 2016, [Online]. Available: http://dergipark.ulakbim.gov.tr/ybs/

[14] S. Hasmita, F. Nhita, D. Saepudin, and A. Aditsania, “Chili Commodity Price Forecasting in Bandung Regency using the Adaptive Synthetic Sampling (ADASYN) and K-Nearest Neighbor (KNN) Algorithms,” in *2019 International Conference on Information and Communications Technology (ICOIACT)*, IEEE, Jul. 2019, pp. 434–438. doi: 10.1109/ICOIACT46704.2019.8938525.

[15] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Second. Canada: John Wiley & Sons, Inc., 2014. doi: 10.1002/9781118874059.

[16] W. Hidayat, M. Ardiansyah, and A. Setyanto, “Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 1, pp. 11–20, Jun. 2021, doi: 10.29408/edumatic.v5i1.3125.

[17] A. M. Halim, M. Dwifebri, and F. Nhita, “Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets,” *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3647.

[18] M. M. Baharuddin, H. Azis, and T. Hasanuddin, “Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca,” *ILKOM Jurnal Ilmiah*, vol. 11, no. 3, pp. 269–274, Dec. 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[19] A. Sugesti, Moch. A. Mukid, and Tarno, “Perbandingan Kinerja Mutual K-Nearest Neighbor (MKNN) dan K-Nearest Neighbor (KNN) dalam Analisis Klasifikasi Kelayakan Kredit,” *Jurnal Gaussian*, vol. 8, no. 3, pp. 366–376, 2019, doi: https://doi.org/10.14710/j.gauss.8.3.366-376.

[20] D. Normawati and S. A. Prayogi, “Implementasi Naïve Bayes Classifier dan Confusion Matrix pada Analisis Sentimen Berbasis Teks Pada Twitter,” *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 5, no. 2, pp. 697–711, 2021, doi: http://dx.doi.org/10.30645/j-sakti.v5i2.369.

[21] Y. Crismayella, N. Satyahadewi, and H. Perdana, “Algoritma Adaboost pada Metode Decision Tree untuk Klasifikasi Kelulusan Mahasiswa,” *Jambura Journal of Mathematics*, vol. 5, no. 2, pp. 278–288, Aug. 2023, doi: 10.34312/jjom.v5i2.18790.