e-ISSN: 2614-8404
p-ISSN: 2776-3234

Air Quality Classification Using Naive Bayes Algorithm With SMOTE Technique Based on ISPU Data

Alya Febriyanti Fadhilah[1*)], Ayu Ratna Juwita[2], Yusuf Eka Wicaksana[3], Tohirin Al Mudzakir[4]
[1234] Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang
Email: [1]if21.alyafadhilah@mhs.ubpkarawang.ac.id, [2]ayurj@ubpkarawang.ac.id, [3]yusuf.eka@ubpkarawang.ac.id, [4]tohirin@ubpkarawang.ac.id

Abstract *– Air pollution in DKI Jakarta is an important issue and has a negative impact on public health. This study applies the naive Bayes algorithm to classify air quality, Utilizing the SMOTE technique effectively addresses the issue of data imbalance. The data analyzed came from air pollution index data from 2022 to 2024, taken from five air monitoring stations in Jakarta. The analysis process was carried out following the CRISP-DM stages, starting from understanding the problem to evaluating the model. The results showed that SMOTE succeeded in increasing prediction accuracy in fewer classes. Without SMOTE, the model accuracy reached 90% but appeared biased towards fewer classes, with a recall value of only 0.75 and a precision of 0.62. While SMOTE, the model accuracy became 88%, with a precision value of 0.86, recall 0.87, and f1-score 0.87, which showed more balanced results across classes.*

Keywords – classification, naive bayes, SMOTE, ISPU, data mining

I. INTRODUCTION

Air pollution is a major issue that has a negative impact on human health, ecosystems, and infrastructure, especially in urban areas. Transportation, industrial, and other activities are the main causes of air pollution [1]. Air pollution is a global challenge that is concerning across various countries, one of which is Indonesia. Jakarta as the largest metropolitan city is known to have a fairly high level of air pollution. In early October 2024, air pollution in DKI Jakarta entered the "unhealthy" category, based on the Air Quality Index at 160 and a PM2.5 concentration of 68.7 µg/m³. This figure is 13.7 times higher than the WHO safe limit [2]. The standard recommended by the World Health Organization is 5 µg/m³ [3]. Greenpeace recorded around 7,390 premature deaths and 2,000 low birth weight babies due to exposure to air pollution. Children, the elderly, and people with comorbidities are the most vulnerable to the impacts [4].

The problem of air pollution in DKI Jakarta can be stated as the toughest and most complex challenge faced by the government [5]. The Indonesian government based on decree on Environmental Impact Management Agency (Bapedal) Number KEP-107/Kabapedal/11/1997 has prepared steps to overcome environmental impacts, namely deciding that the Air Pollution Standard Index is a tool for assessing air quality in an area and its effects on human health, animals, plants, as well as aesthetic value. The Air Pollution Standard Index is a value without a unit that indicates the state of ambient air quality at a certain location and time, in order to assess its effects on human health, aesthetics, and other living things [6].

Determining the ISPU level can be done by applying classification methods in data mining as an efficient solution [7]. Data mining is a method used to obtain new information from a group of data by identifying certain patterns and rules obtained from large data sets [8]. The application of data mining is one solution to analyze air quality in DKI Jakarta using a classification technique approach.

Based on research conducted by Lasulika [9]. About Comparison of Naive Bayes, Support Vector Machine and K-Nearest Neighbor to Find the Highest Level of Accuracy in Smoothness of Cable TV Payment. The results of the study stated that naive bayes produced the highest level of accuracy of 96% and AUC of 0.99, while K-NN obtained an accuracy of 92% at K=3 and SVM provided an accuracy of 66% and an AUC value of 0.786. Then by Arnap [10] conducted a study on Improving SQL Injection Attack Detection Using Naive Bayes and SMOTE Methods on Imbalanced Datasets. Producing an accuracy rate of 99.50%, f1-score of 99.50%, precision of 99.50% and recall of 99.50%, this shows that the SMOTE technique is effective in balancing classes in the dataset and improving SQL injection attack detection.

Another research by Nurhariza [11] namely the Implementation of the Niave Bayes Algorithm for Classification to Determine Student Achievement Based on Average Values, produced an accuracy of 98%, 100% precision and 98% recall. Next, Kurniadi [12]conducted a study on the Classification of Village Fund Direct Cash Assistance Recipients Using Naïve Bayes and SMOTE. His study conducted a comparison that produced an accuracy value of 97.07% for the naïve bayes model without SMOTE, while with SMOTE it increased to 97.80%, precision of 96.67%, and recall of 99.02%.

Based on the previous explanation, the purpose of this study is to analyze the performance of the naïve Bayes algorithm using the SMOTE technique in classifying air pollution in DKI Jakarta. It is expected that this study can produce a more accurate classification to support the decision-making process in managing and overcoming air pollution efficiently

e-ISSN: 2614-8404
p-ISSN: 2776-3234

II. METHODOLOGY

A. Reseacrch Stages

This study utilizes the research stages of a data mining series known as CRISP-DM. In research [13] according to Laroske, CRISP-DM is a standard data mining process used to solve general problems in research or business. CRISP-DM is an approach that uses a data development stage model that is often applied by experts in solving a problem [14]. The CRISP-DM method consists of six stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

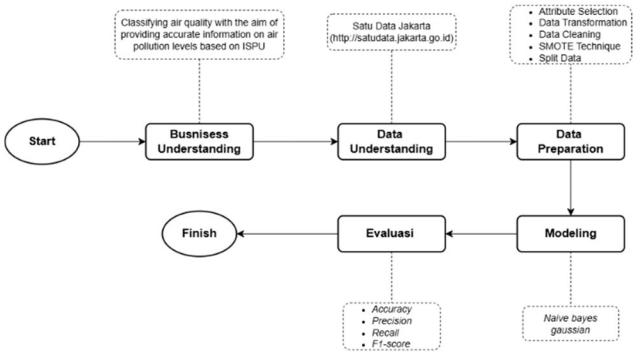

Fig 1. CRISP-DM methodology

1. Business Understanding

The first stage is to understand or analyze the problems that need to be solved in business. In this process, understanding is carried out on the monitoring of DKI Jakarta's air quality as a whole, so that accurate information is obtained.

2. Data Understanding

The second process includes data collection, data analysis and explanation, and identification of problems related to the data. The data collection process was obtained from the website http://satudata.jakarta.go.id about the DKI Jakarta Air Pollution Standard Index data in excel format. The data used covers the years 2022 to 2024 Each row of data describes the results of daily ISPU monitoring with the main attributes being the monitoring date, monitoring station, pollutant parameters such as PM10, PM2.5, CO, NO2, O3, SO2, and air quality category. Which were taken from collected five air quality measurement stations in DKI Jakarta, namely DKI1 Bundaran HI, DKI2 Kelapa Gading, DKI3 Jagakarsa, DKI4 Lubang Buaya, and DKI5 Kebon Jeruk. The reason for choosing these three years is because the time difference is not too far from the implementation of this research and can provide a more accurate picture of the current situation.

3. Data Preparation

In the third step, it includes the process of compiling the final dataset that will be used as input in data mining modeling. The data preparation process is carried out as follows:

1) Attribute Selection, namely the process of selecting the most relevant attributes such as CO, O3, SO2, NO2, PM10, PM2 and categories.
2) Transformation data, which is the process of changing the format or structure of data according to modeling such as recategorization and coding of categorical features.
3) Data Cleaning, which is the process of cleaning data that has the potential to interfere with modeling analysis, such as handling missing values and outliers.
4) Balancing Data is the process of equalizing the distribution of classes in a dataset so that the quantity of samples in every class becomes more balanced, such as the SMOTE technique which is applied to balance the distribution of datasets in minority classes by creating minority datasets so that the number is the same as the dataset in the majority class [15]. The SMOTE equation can be seen in equation (1).

$$Xnew = Xi + (Xi\hat{} - Xi) \times \delta \quad (1)$$

5) Data Division, which is the stage of separating the dataset into training data and test data with a certain ratio such as 80:20.

4. Modeling

In the modeling stage, the selected algorithm model is applied to the dataset. The data mining modeling process applied is naive bayes gaussian. The naive Bayes algorithm has superior performance in performing classification, especially in terms of the accuracy of the classification results produced [16]. The function of naive bayes gaussian is to calculate data attributes that have continuous properties directly [17]. The naive bayes gaussian equation can be seen in equation (2).

$$(Xi = xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu\ )^2}{2\sigma^2 ij}} \quad (2)$$

One method to build a simple model is to assume that the data follows a Gaussian distribution with no covariance between dimensions. Thus, the model can be created simply by calculating the average and standard deviation of the data for each label.

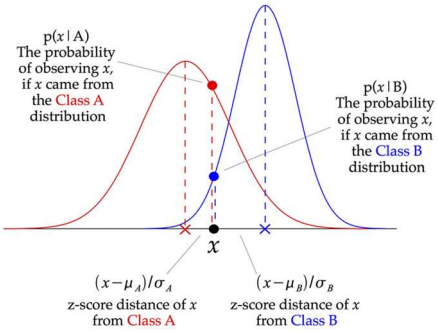

Fig 2. How the naive bayes gaussian classification works (www.researchgate.net)

Figure 2 illustrates how the naive bayes gaussian works, where each data point is calculated its z-score value against the class mean, which is the difference between the data

e-ISSN: 2614-8404
p-ISSN: 2776-3234

value and the class mean divided by the class standard deviation. Therefore, the naive bayes gaussian uses a slightly different approach and is quite effective in handling continuous data. Furthermore, the classification process is carried out for each attribute by referring to Bayes' Theorem, where new information is obtained through the calculation of the statistical probability of each attribute.

5. Evaluation

At the evaluation stage, an analysis is carried out on the accuracy of the results of the data processing process that has been applied. The evaluation process uses a confusion matrix that specifically describes the performance of the model. In this matrix, each row denotes the actual data class, while each column signifies the predicted data class [18].

Table 1. Confusion matrix

Actual Values	**Predicted Values**	
	Positive	***Negative***
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

1) Accuracy
The total count of frequencies accurately classified by the model. The mathematical equation for accuracy is given in formula 3.

$$\frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

2) Precision
When a model makes a positive prediction, the accuracy of that prediction indicates how often the result is correct. The mathematical equation for precision is given in formula 4.

$$\frac{TP}{TP+FP} \quad (4)$$

3) Recall
When the actual class has a positive value, this metric measures the extent to which the model correctly predicted positive. The mathematical equation for recall is in equation 5.

$$\frac{TP}{TP+} \quad (5)$$

4) F1-Score
This harmonic mean value is obtained by precision and recall. The mathematical equation of f1-score is found in equation 6.

$$\frac{2(Recall*Precision)}{(Recall+Pre\quad)} \quad (6)$$

B. Indeks Standar Pencemaran Udara

Indeks Standar Pencemaran Udara (ISPU) is a parameter applied to measure air quality by indicating the level of pollution due to the influence of chemical substances and particles in the atmosphere [19]. The calculation of ISPU is based on the upper and lower threshold values for ambient, as well as ambient concentrations obtained from measurement results using the following mathematical formula:

$$I = \frac{Ia-Ib}{xa-xb}(Xx - Xb) + Ib \quad (7)$$

The upper bound value of ISPU (Ia) is set at 100, while the lower bound value of ISPU (Ib) is set at 50. The ambient concentration for the upper limit (Xa) and lower limit (Xb) is obtained from the ISPU parameter concentration conversion table, with different Xa and Xb values depending on each parameter. Meanwhile, the actual ambient concentration value (Xx) is calculated based on the average ambient concentration during 24 hours of measurement.[20]. The air pollution index value has various categories which are used to measure the level of air pollution detected in Figure 3.

Kategori	Status Warna	Angka Rentang
Baik		1 – 50
Sedang		51 – 100
Tidak Sehat		101 – 200
Sangat Tidak Sehat		201 - 300
Berbahaya		≥ 301

Fig 3. ISPU Category (www.climate4life.info)

Figure 3 illustrates the air quality classification system based on the ISPU standard issued by the Ministry of Environment and Forestry of the Republic of Indonesia. Air quality is classified into five levels based on the ISPU value: Good (1–50), Moderate (51–100), Unhealthy (101–200), Very Unhealthy (201–300), and Hazardous (≥301). Each category is marked with a certain color as a visual indicator, namely green for Good, blue for Moderate, yellow for Unhealthy, red for Very Unhealthy, and black for Hazardous.

III. RESULTS AND DISCUSSION

1. Business Understanding

DKI Jakarta air quality monitoring is carried out to obtain data that supports the air quality classification process. The data obtained provides accurate information on the level of air pollution based on the air pollution standard index and its impact on health due to exposure to pollution, so that it can help the community and government in making decisions based on current conditions.

2. Data Understanding

The total data successfully collected was 3,506 Air Pollution Standard Index data for 2022-2024, with details of 321 data in the good category, 269 data in the moderate category, 482 data in the unhealthy category and 4 data in the very unhealthy category.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

tanggal	pm_10	pm_duakomalima	so2	co	o3	no2	max	critical	categori	stasiun
2022-12-31	54	73	56	24	23	24	73	PM2,5	SEDANG	DKI4
2022-12-30	40	64	57	21	17	24	64	PM2,5	SEDANG	DKI4
2022-07-06	75	129	45	25	71	26	129	PM2,5	TIDAK SEHAT	DKI4
2022-07-05	66	110	47	16	61	23	110	PM2,5	TIDAK SEHAT	DKI4
2022-07-04	56	78	49	11	60	13	78	PM2,5	SEDANG	DKI4
2022-07-03	78	126	48	17	126	25	126	PM2,5	TIDAK SEHAT	DKI2
2022-07-02	81	137	47	18	106	29	137	PM2,5	TIDAK SEHAT	DKI4
2022-08-01	68	109	50	12	91	19	109	PM2,5	TIDAK SEHAT	DKI4
2022-08-31	85	148	51	25	74	37	148	PM2,5	TIDAK SEHAT	DKI4
2022-08-30	59	84	50	15	56	29	84	PM2,5	SEDANG	DKI4
2022-08-29	58	82	49	16	55	29	82	PM2,5	SEDANG	DKI4
2022-08-28	64	85	49	17	56	30	85	PM2,5	SEDANG	DKI5
2022-08-27	80	130	50	25	57	37	130	PM2,5	TIDAK SEHAT	DKI4
2022-08-26	76	119	53	19	53	31	119	PM2,5	TIDAK SEHAT	DKI4
2022-08-25	64	98	50	12	66	19	98	PM2,5	SEDANG	DKI4
2022-08-24	71	119	50	15	82	20	119	PM2,5	TIDAK SEHAT	DKI4
2022-08-23	76	116	49	17	81	27	116	PM2,5	TIDAK SEHAT	DKI4
2022-08-22	57	78	49	13	79	22	79	O3	SEDANG	DKI2
2022-08-21	59	101	49	13	65	17	101	PM2,5	TIDAK SEHAT	DKI4
2022-08-20	76	117	50	14	69	18	117	PM2,5	TIDAK SEHAT	DKI4
2022-08-19	70	105	49	14	66	26	105	PM2,5	TIDAK SEHAT	DKI4
2022-08-18	64	89	49	25	84	28	89	PM2,5	SEDANG	DKI4
2022-08-17	55	78	48	17	108	25	108	O3	TIDAK SEHAT	DKI2
2022-08-16	61	89	51	15	181	33	181	O3	TIDAK SEHAT	DKI2
2022-08-15	64	84	50	17	63	28	84	PM2,5	SEDANG	DKI4

Fig 4. Air Quality Dataset

3. Data Preparation

Data Preprocessing includes attribute selection, Data Transformation, Data Cleaning, Data Balancing, and Data Splitting.

1) Attribute Selection

Based on the dataset from 2022 to 2024, there are 11 attributes shown in Figure 4. However, only 7 attributes are used in the classification, namely the categories PM10, PM2.5, SO2, CO, O3, NO2 and the category.

	pm_10	pm_duakomalima	so2	co	o3	no2	categori
0	54	73	56	24	23	24	SEDANG
1	40	64	57	21	17	24	SEDANG
2	75	129	45	25	71	26	TIDAK SEHAT
3	66	110	47	16	61	23	TIDAK SEHAT
4	56	78	49	11	60	13	SEDANG
...	...	...	...	...	...	...	...
3501	51	68	60	10	26	51	SEDANG
3502	65	87	59	11	35	22	SEDANG
3503	53	70	60	8	32	47	SEDANG
3504	51	74	61	8	33	44	SEDANG
3505	59	84	61	9	28	44	SEDANG

Fig 5. Dataset after attribute selection

2) Transformasi Data

Data transformation is carried out through two processes, namely recategorization, combining several categories into one, and categorical feature coding, converting categorical data into a numeric format.

a) Recategorization

Recategorization is done by combining the "very unhealthy" category into the "unhealthy" category because it only has 4 data that have little influence on the model. This recategorization is done to overcome data limitations, so that it can maintain model performance and increase the effectiveness of the SMOTE technique.

categori	
SEDANG	2699
TIDAK SEHAT	482
BAIK	321
SANGAT TIDAK SEHAT	4

Fig 6. Before re-categorization

categori	
SEDANG	2699
TIDAK SEHAT	486
BAIK	321

Fig 7. Re-categorization results

b) Encoding Categorical Feature

Categorical feature encoding is done by changing the category parameter numbers that were originally strings into numeric types. In this case, the categorical labels "good", "moderate", and "unhealthy" are changed to 2, 1, and 0.

pm_10	pm_duakomalima	so2	co	o3	no2	categori
54	73	56	24	23	24	1
40	64	57	21	17	24	1
75	129	45	25	71	26	0
66	110	47	16	61	23	0
56	78	49	11	60	13	1
...	...	...	...	...	...	...
51	68	60	10	26	51	1
65	87	59	11	35	22	1
53	70	60	8	32	47	1
51	74	61	8	33	44	1
59	84	61	9	28	44	1

Fig 8. Results of encoding categorical features

3) Data Cleaning

Data cleaning is carried out to clean data that has the potential to interfere with analysis and modeling. One of the main focuses in this process is handling data with "-" values and outliers so that the data is cleaner and ready for analysis.

	pm_10	pm_duakomalima	so2	co	o3	no2	categori
647	-	35	52	2	8	8	1
648	32	64	52	7	8	12	1
649	49	73	-	-	-	-	1
650	52	-	18	17	27	5	1
651	48	-	20	18	21	5	2

Fig 9. Dirty air quality data

In row 647, pm_10 column, there is data with the value "-", likewise in row 649 there is data with the value "-" in the so2, co, o3 and no2 columns. Data with the value "-" is converted to NaN to facilitate the calculation of missing data. The missing numbers are then overcome by filling in using the median of each parameter.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

	pm_10	pm_duakomalima	so2	co	o3	no2	categori
647	55	35	52	2	8	8	1
648	32	64	52	7	8	12	1
649	49	73	42	13	26	20	1
650	52	78	18	17	27	5	1
651	48	78	20	18	21	5	2

Fig 10. Missing Value filled with median

In Figure 10, there is no "-" data because it has been replaced using the median of each parameter. The median value used for each parameter is pm_10 has a median value of 55, pm_duakomalima has a median value of 78, so2 is 42, co is 13, o3 is 26 and no2 is 20.

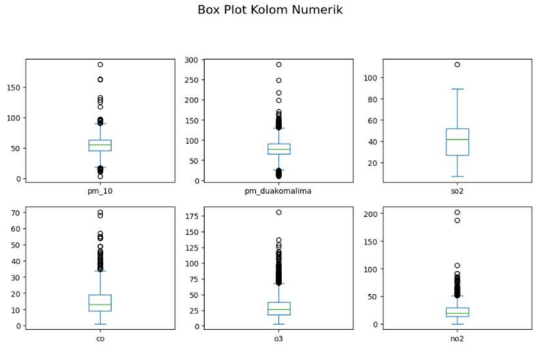

Fig 11. Outlier checking

In Figure 11, it can be seen that all parameters have outliers, parameters such as pm_10, pm_duakomalima, co, o3 and no2 have significant outliers, while the so2 parameter is not very significant. Parameters that have extreme outliers are handled with the Interquartile Range (IQR). Outliers will be replaced with lower or upper limits so as not to eliminate data.

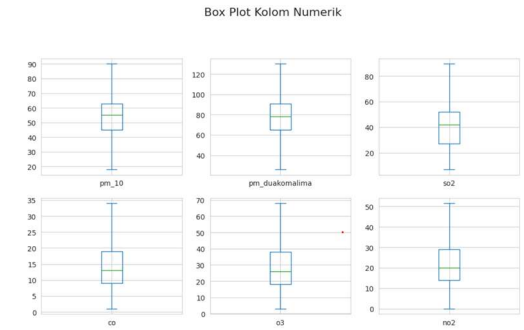

Fig 12. After performing the Interquartile Range

4) Balancing Data
Data balancing is done using SMOTE which is a technique in data imbalance.

Table 2. SMOTE Technique Results

Class	Original	SMOTE
0	486	2699
1	2699	2699
2	321	2699
Total	**3506**	**8097**

Category 0 is the minority class, while category 1 is the majority class, the SMOTE technique balances the data with the number of majority classes reaching 2699 data.

5) Data Sharing
The data obtained from the balancing process amounted to 2699 per category. As much as 80% was allocated as training data while 20% was test data. The results of the initial data division were 2804 for training data and 702 testing data. Meanwhile, the outcomes of the division in the SMOTE method were 6477 training data and 1620 test data.

4. Modeling

After the data preparation is complete, the next step is to perform the naive bayes gaussian algorithm modeling. In the modeling stage, two scenarios are created as in Figure 12.

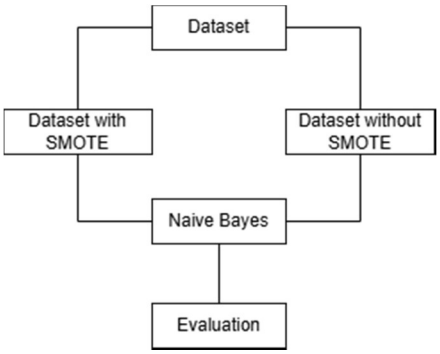

Fig 13. Modeling scenario

This study scenario intends to compare results of the naïve bayes gaussian classification algorithm on datasets that go through the SMOTE sampling technique process with datasets that do not use the SMOTE technique. Figures 14 and 15 are the classification results.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95       540
           1       0.85      0.77      0.81       540
           2       0.86      0.87      0.87       540

    accuracy                           0.88      1620
   macro avg       0.88      0.88      0.88      1620
weighted avg       0.88      0.88      0.88      1620
```

Fig 14. Naïve Bayes Gaussian + SMOTE results

Figure 14 shows the train and test process the performance of the naïve Bayes model on the dataset that applies the SMOTE method. This stage produces an accuracy percentage of 0.88, with an f1-score of 0.88, precision of 0.88, and recall of 0.88. Although these results show a slight decrease in model accuracy, the distribution of precision, recall, and f1-score values between classes becomes more balanced. In minority class 2, there is a significant increase in performance, with a precision of 0.86, a recall of 0.87, and an f1-score of 0.87. With this, it can be shown that the application of SMOTE has succeeded in increasing the ability of the naïve Bayes model to

e-ISSN: 2614-8404
p-ISSN: 2776-3234

recognize data in the minority class without significantly reducing model performance.

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.86	0.88	97
1	0.94	0.93	0.94	541
2	0.62	0.75	0.68	64
accuracy			0.90	702
macro avg	0.82	0.85	0.83	702
weighted avg	0.91	0.90	0.91	702

Fig 15. Naïve Bayes Gaussian results without SMOTE

Figure 15 shows the results of naïve bayes on a dataset that does not apply the SMOTE method. The model produces an accuracy percentage of 0.90, with a precision 0.90, recall 0.90, and f1-score 0.90. However, in minority class 2, the values obtained tend to be lower than the other classes, namely precision of 0.62, recall of 0.75, and f1-score of 0.68. This shows that the classification model is biased towards the majority class and is less than optimal in classifying the minority class.

Table 3. Performance Comparison

Metode	***Accuracy***	***Precision***	***Recall***	***F1 - Score***
Naïve bayes + SMOTE	88%	88%	88%	88%
Naïve bayes	90%	90%	90%	90%

5. Evaluation

The evaluation and comparison process is carried out to evaluate the performance of the classification model with naive bayes gaussian algorithm using a confusion matrix.

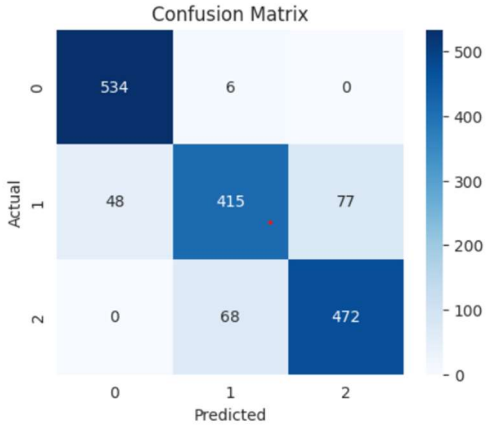

Fig16. Confusion Matrix Naive Bayes + SMOTE

Figure 16 shows the confusion matrix generated naive Bayes model test results that have applied the SMOTE method. The outcomes show that the model is able to predict classes 0 and 2 better and more balanced.

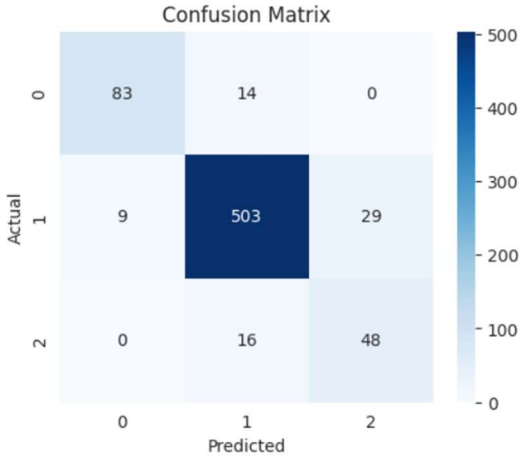

Fig 17. Naive Bayes Confusion Matrix

Figure 17 shows confusion matrix from the results of the naive Bayes test alone, which results in the model tending to be biased towards class 1, while the predictions for classes 0 and 2 are still low.

.

IV. CONCLUSION

Based on the results of the research that has been conducted, the following are the conclusions that can be obtained:

1. The application SMOTE successfully overcomes the imbalance of ISPU data by increasing the amount of data in the minority class. This makes the classification model, especially the naïve bayes gaussian algorithm, fairer in recognizing various classes. After SMOTE was applied, the prediction performance for the minority class increased, as indicated by the more balanced precision, recall, as well as f1-score values in every category compared to when not using SMOTE.
2. Based on the evaluation results, the naïve bayes model devoid of using SMOTE has an accuracy of 90%, but its performance is biased towards the majority class, which can be seen in the low recall value of 0.75 and precision of 0.62 for minority class. After the SMOTE technique was applied, the accuracy did decrease slightly, namely 88%, but at a recall value of 0.87, precision 0.86 and f1-score 0.87 in the minority class increased. This shows that the model is more balanced in classifying each air quality class.

REFERENCES

[1] A. Handayani, Silva, S. Soim, T. Agusdi, Enim, Rumsiasih, and A. Nurdin, "KLASIFIKASI KUALITAS UDARA DENGAN METODE SUPPORT VECTOR MACHINE," *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 3, no. 2, pp. 187–199, 2020.

[2] Tempo, "Hari Pertama Oktober 2024, IQAir Catat Kualitas Udara Jakarta Memburuk," TEMPO. Accessed: Nov. 28, 2024. [Online]. Available:

e-ISSN: 2614-8404
p-ISSN: 2776-3234

https://www.tempo.co/lingkungan/hari-pertama-oktober-2024-iqair-catat-kualitas-udara-jakarta-memburuk--4050

[3] IQAir Staff Writers, "STANDAR WHO," IQAir. Accessed: Nov. 24, 2024. [Online]. Available: https://www.iqair.com/id/newsroom/new-epa-annual-pm25-air-quality-standards-expected-save-lives

[4] O. Health, "Polusi Jakarta Peringkat 1 di Dunia, Bagaimana Dampaknya pada Kesehatan?," ONE HEALTH CENTER OF EXCELLENCE UNIVERSITAS GADJAH MADA. [Online]. Available: https://ohce.wg.ugm.ac.id/polusi-jakarta-peringkat-1-di-dunia-bagaimana-dampaknya-pada-kesehatan/

[5] Bernadet, S. Listyarini, and L. Warlina, "Pengaruh Kebijakan Pencemaran Udara Sektor Transportasi Terhadap Nilai Indeks Kualitas Udara (Iku) Di Dki Jakarta," *Jurnal Ilmiah Pendidikan Lingkungan dan Pembangunan*, vol. 24, no. 01, pp. 1–13, 2023, doi: 10.21009/plpb.v24i01.30798.

[6] L. Qosim, Ahmad, "PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) DAN NAIVE BAYES CLASSIFIER (NBC) UNTUK MENENTUKAN KUALITAS UDARA," p. 6, 2021.

[7] R. Firdaus *et al.*, "Implementasi Algoritma Random Forest Untuk Klasifikasi Pencemaran Udara di Wilayah Jakarta Berdasarkan Jakarta Open Data," vol. 14, no. 2, pp. 520–525, 2021.

[8] A. T. Ramadhan *et al.*, "Penerapan Algoritma Decision Tree Dalam Melakukan Analisis Klasifikasi Harga Handphone," *Jurnal Sistem Informasi dan Ilmu Komputer*, vol. 1, no. 4, pp. 195–206, 2023, [Online]. Available: https://doi.org/10.59581/jusiik-widyakarya.v1i4.1861

[9] M. E. Lasulika, "Komparasi Naïve Bayes, Support Vector Machine Dan K-Nearest Neighbor Untuk Mengetahui Akurasi Tertinggi Pada Prediksi Kelancaran Pembayaran Tv Kabel," *ILKOM Jurnal Ilmiah*, vol. 11, no. 1, pp. 11–16, 2019, doi: 10.33096/ilkom.v11i1.408.11-16.

[10] A. Arnap, "Enhancing SQL Injection Attack Detection Using Naïve Bayes and SMOTE Method on Imbalanced Datasets," vol. 4, no. 1, 2024.

[11] M. Nurhariza, A. Ratna Juwita, and D. Sulistya Kusumaningrum, "Implementasi Algoritma Naive Bayes Untuk Klasifikasi Menentukan Prestasi Siswa Berdasarkan Nilai Rata-Rata," vol. V, no. 1, pp. 65–77, 2024.

[12] D. Kurniadi, F. Nuraeni, and M. Firmansyah, "Klasifikasi Masyarakat Penerima Bantuan Langsung Tunai Dana Desa Menggunakan Naïve Bayes dan SMOTE," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, pp. 309–320, 2023, doi: 10.25126/jtiik.20231026453.

[13] N. Ajijah and A. Kurniawan, "Klasifikasi Teks Mining Terhadap Analisa Isu Kegiatan Tenaga Lapangan Menggunakan Algoritma K-Nearest Neighbor (KNN)," *J-SAKTI (Jurnal Sains Komputer & Informatika)*, vol. 7, no. 1, pp. 254–262, 2023.

[14] K. Suhada, A. Elanda, and A. Aziz, "Klasifikasi Predikat Tingkat Kelulusan Mahasiswa Program Studi Teknik Informatika dengan Menggunakan Algoritma C4.5 (Studi Kasus: STMIK Rosma Karawang)," *Dirgamaya: Jurnal Manajemen dan Sistem Informasi*, vol. 1, no. 2, pp. 14–27, 2021, doi: 10.35969/dirgamaya.v1i2.182.

[15] G. Gumelar, Q. Ain, R. Marsuciati, S. Agustanti Bambang, A. Sunyoto, and M. Syukri Mustafa, "Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance," *SISFOTEK : Sistem Informasi dan Teknologi*, pp. 250–255, 2021.

[16] A. A. Mahran, R. K. Hapsari, and H. Nugroho, "Penerapan Naive Bayes Gaussian Pada Klasifikasi Jenis Jamur Berdasarkan Ciri Statistik Orde Pertama," *Network Engineering Research Operation*, vol. 5, no. 2, p. 91, 2020, doi: 10.21107/nero.v5i2.165.

[17] A. A. H. Kirono, I. Asror, and Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma Naïve Bayes," *e-Proceeding of Engineering*, vol. 9, no. 3, p. 1962, 2022.

[18] I. W. Saputro and B. W. Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa," *Creative Information Technology Journal*, vol. 6, no. 1, p. 1, 2020, doi: 10.24076/citec.2019v6i1.178.

[19] D. Perdana and A. Muklason, "Machine Learning untuk Peramalan Kualitas Indeks Standar Pencemar Udara DKI Jakarta dengan Metode Hibrid ARIMAX-LSTM," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 5, no. 3, pp. 209–222, 2023, doi: 10.28926/ilkomnika.v5i3.588.

[20] I. S. R. Dasrul Chaniago, Annisa Zahara, "Kementerian Lingkungan Hidup dan Kehutanan," 24-09. [Online]. Available: https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia