Implementation of the ordinal logistic regression method for air quality classification based on the air pollution standard index

Dwi Nur Fitrianto [1], **Suwanto Sanjaya** [2*], **Fadhilah Syafria** [3], **Elin Haerani** [4], **Siska Kurnia Gusti** [5]

[1-5] Faculty of Science and Technology, Sultan Syarif Kasim Riau State Islamic University, Pekanbaru, Indonesia

* Corresponding Author: **Suwanto Sanjaya**

Article Info

ISSN (online): 2582-7138
Volume: 05
Issue: 03
May-June 2024
Received: 08-04-2024
Accepted: 14-05-2024
Page No: 938-949

Abstract

Air pollution is a serious problem in many cities around the world, caused by human and natural activities. Jakarta, as an Indonesian economic and transportation hub, faces serious air pollution challenges. This study uses the Ordinal Logistics Regression Method to develop an optimal classification model for identifying air quality based on the air pollution index. The aim is to contribute to dealing with air contamination and improve understanding of the use of such methods in the classification of air quality. Data used from 2012 to 2021 covered parameters such as PM10, SO2, CO, O3, and NO2. Oversampling is done using SMOTE to address imbalances in the datasets used. The data used is divided into two parts: training data and validation test data, with an 80:20 ratio. The model training and testing process is carried out with a variety of scenarios, including parameter significance tests using probability tests and Wald tests, cross-validation with fold numbers 5, 10, and 15, as well as evaluation using a confusion matrix. Models are used for Mord libraries such as Ordinalridge, LogisticAT, logisticIT, and LogisticS. There were a total of 72 model testing experiments to find the best model of the six data model outputs. The test results showed that the optimal model was obtained with a data deletion scenario of null values, data oversampling using the LogisticIT model, and K-fold = 5, with a training accuracy of 0.8628 and validation data test accuracy of 0.8599, as well as each precision value of 0.86, recall value of 0.86, and F1-score of 0.86. The model's performance was satisfactory in handling different data variations, according to the evaluation. These results show that the model is able to generalize data well and is reliable in predicting air quality accurately.

DOI: https://doi.org/10.54660/.IJMRGE.2024.5.3.938-949

Keywords: Air pollution, ordinal logistics regression, classification, cross validation, likelihood, wald

1. Introduction

Air pollution is a serious problem facing many cities and regions around the world (Handayani *et al.*, 2020) [6]. Factors that can degrade air quality come from nature and human activities, such as transportation, industry, land and forest fires (Valentino Jayadi *et al.*, 2023) [20]. High population growth, rapid urbanization, and increased industrial activity makes air quality a major concern for human health and the environment (Sang *et al.*, 2021) [18]. The impact of air pollution is enormous on human health, the environment, and the global climate (Henri, 2021) [8]. Air pollution not only affects the air we breathe every day, but also has a long-term impact on respiratory health and urban ecosystems. The rapid and dense population growth in the major cities of Indonesia is the main cause of air pollution (Astriyani *et al.*, 2023) [4]. The Jakarta DKI region, as an economic and transportation hub, faces serious air pollution problems due to rapid population growth and significant industrial activity (Agista *et al.*, 2020) [2]. Jakarta, as an economic and transportation hub, faces complex air quality challenges that affect public health. Data from the Ministry of Environment and Forestry shows that the transport sector accounts for 44% of air pollution in Jakarta

(Hasiman, 2023) [7]. According to the Air Quality Index, Jakarta will be the fourth worst air quality city in the world by August 2023.

The air quality index in Jakarta has reached 157, which falls into the unhealthy air quality category The World Health Organization reports that there are at least 7 million premature deaths every year worldwide due to exposure to air pollution (WHO, 2023) [21].

According to the Government of the Republic of Indonesia Regulation No. 41 of 1999 on the control of air pollution, air contamination refers to a mixture of substances, energy, and/or other components that enter the air through human activity or natural processes, causing a decrease in the quality of the air to a certain degree so that the air no longer fulfills its function optimally (Presiden Republik Indonesia, 1999) [12]. The Indonesian government has taken steps to address air pollution, one of which is the publication of Decision No. P.14/MENLHK/SETJEN/KUM.1/7/2020 regulating the Air Pollution Standard Index (ISPU). ISPU is a report that presents information to the public about the level of air pollution in a region over a certain period of time to the public, issued by the Ministry of Environment and Forestry. Based on the decision, ISPU were divided into five categories: good, moderate, unhealthy, highly unhealthier, and dangerous. ISPU classification is based on the content of such parameters as SO2 (Sulfur Dioxide), CO (Carbon Monoxide), NO2 (Nitrogen Dioxide), O3 (Ozone), PM10 (Particulate Matter 10), PM25 (Particulate Matter 25), and HC (Hydrocarbons) (MenLHK, 2020) [10].

Air quality monitoring should be measured daily using the standard index officially published by the government, ISPU. Based on the ISPU parameter index, it requires a system of rapid and accurate data classification through data mining techniques. The results help the city government make decisions and control air pollution to maintain good air quality for the community. Data mining aims to dig up information efficiently. One of the techniques used to predict is classification (Etriyanti *et al*., 2020) [5]. Classification is a process in machine learning or statistics in which data is divided into certain classes or categories based on their attributes. The goal is to create a model using labeled training data to predict a new data class. If there are only two classes, it's called a binary classification; if there are more than two classes, it's called a multiclass classification (Raharjo, 2021) [16].

The Ordinal Logistic Regression Method is a classification method that can be used. This method is an analytical technique applied to understand the relationship between predictor variables (X) that are categorical or numerical and response variables (Y) that are categorical, in which response variables have more than two categories that follow the ordinal scale or level (Addini *et al*., 2022) [1]. The ordinal logistic regression method is used to analyze the answer variable that has an ordinal scale with three categories or more. This variable has a level that can be ordered, with a sequence determined by the value of the variable response in increasing order. The first category is considered the one with the lowest score (Hosmer & Lemeshow, 2000) [9]. This method is a variation of binary logistic regression that is generally applied to estimate binary or categorical variables with two classes. However, in ordinal logistics regression, the response variable has three or more classes that follow a certain order, such as "low," "sized," and "high," which can be ordered in order. This allows for a more detailed analysis of the relationship between predictor variables and responses in scenarios in which responses can not only be divided into two categories. Therefore, this method is often used in various classification applications where responses are not only binary but have several categories that the ordinal order.

Several studies on the classification of air quality have been carried out using the K-Nearest Neighbor algorithm by producing the K=7 model with the best performance of 96%, precision of 92%, recall of 95%, and f-measure of 93% (Amalia *et al*., 2022) [3]. Similar research has been done using the Artificial Neural Network (ANN) Backpropagation algorithm method by obtaining the most optimal classification results obtained from 5 layers of input, 4 hidden layers, and 2 output layers, as well as 5000 epochs and a 0.001 learning rate, obtaining an accuracy of 94%, precision of 90%, and recall of 100% (Putri & Suwanda, 2023) [15]. In another study using the Support Vector Machine method with Hyperparameter Optimization, GridSearch CV for air quality prediction produced an accuracy that has been optimized at 94.8% using a polynomial kernel with 2 degrees that gives an improvement in accuracy of 21.5% (Toha *et al*., 2022) [19]. The research on air quality was also conducted using the Fuzzy Tsukamoto method, which yielded accuracy values of 97% of the classification system carried out (A. E. Putra *et al*., 2023) [13]. Other research using the method of ordinal logistic regression in the classification was conducted by Gusti Ngurah Sentana Putra, dkk, entitled "Classification of Family Welfare Level in the Sidemen District Using Bootstrap Aggregating (Bagging) Regression Logistic Ordinal," which resulted in a classification accuracy of 79.40%, and the method of bagging ordinal logistic regression with 50,000 iterations resulted in an accuracy of 82.78%. So the process of bootstrapping and aggregating increased accuracy by 3.21% (I. G. N. S. Putra *et al*., 2023) [14].

Based on the explanation above, the study focuses on finding the best model of the many scenarios that are undertaken to produce a reliable and accurate classification model. The study uses the Ordinal Logistic Regression method to classify air quality based on the standard air pollution index using five parameters, namely PM10, NO2, SO2, CO, and O3. Uses datasets from 2012 to 2021. Perform some data cleaning techniques, such as deleting null values and entering null values with an average, so that you have some data model scenarios to use in research. Execute oversampling on data that is imbalanced using the SMOTE technique. Conduct the parameter significance test with the likelihood ratio test and the Wald test. Model training using k-fold cross-validation with K values = 5,10 and 15. Then perform several test scenarios to find the best model using the models found in the Mord library. Then perform tests using validation test data to prevent overfitting of the model. Model evaluation using a confusion matrix as well as measuring accuracy, precision, recall, and F1-score values. The research is aimed at producing accurate and efficient classification models that are expected to help in decision-making related to air pollution management by governments and related agencies. It is hoped that this research can contribute to improving air quality, protecting public health, and preserving the living environment. In addition, it is also expected to expand the understanding of the use of ordinary logistics regression in the context of air quality classification.

2. Material and Method

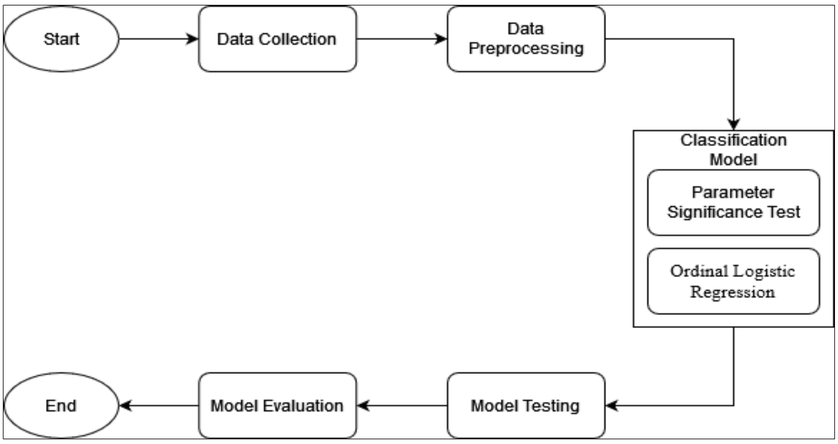

Fig 1: Research Methodology

2.1. Data collection

At this stage, the required data, such as parameter values that determine air quality according to the Air Pollution Standard Index (ISPU), which includes PM10 (particulates), SO2 (sulfur dioxide), CO (carbon monoxide), O3 (ozone), and NO2 (nitrogen dioxide), as well as air quality categories, are collected. This research data is obtained as secondary data from the Jakarta Open Data website, which is available through the link https://satudata.jakarta.go.id/. This data set is downloaded from the website in.csv file format.

2.2. Data Processing

Data sets are not currently ready for use in classification models, so data processing efforts are required, including data cleaning to form the transformation process. At this stage, it is done as follows:

a. Combine air quality data from 2012 to 2021 into a single data file.
b. Clean the data by identifying incomplete data, duplicates, and empty rows or columns. How to overcome this by removing or filling in average values.
c. Doing attribute selection and deleting irrelevant attributes. It aims to reduce the complexity of the attributes processed by the algorithm.
d. Perform a descriptive analysis to understand the image of the data obtained.
e. Perform oversampling to deal with data imbalances using the SMOTE (Synthetic Minority Oversampling Technique) technique, so that it can help improve the performance of models in minority classes with fewer observations. Oversampling is a technique that involves randomly adding data from a minority class to the training data. This adding process is repeated until the amount of data from the minority classes equals the number of majority classes (Zhafirah, 2023) [24].

2.3. Modeling

Once the data processing phase is complete, the next step is to apply the data to the Ordinal Logistics Regression method to classify air quality. The procedure for the analysis of the classification model in this study is as follows can be seen in Fig 2.

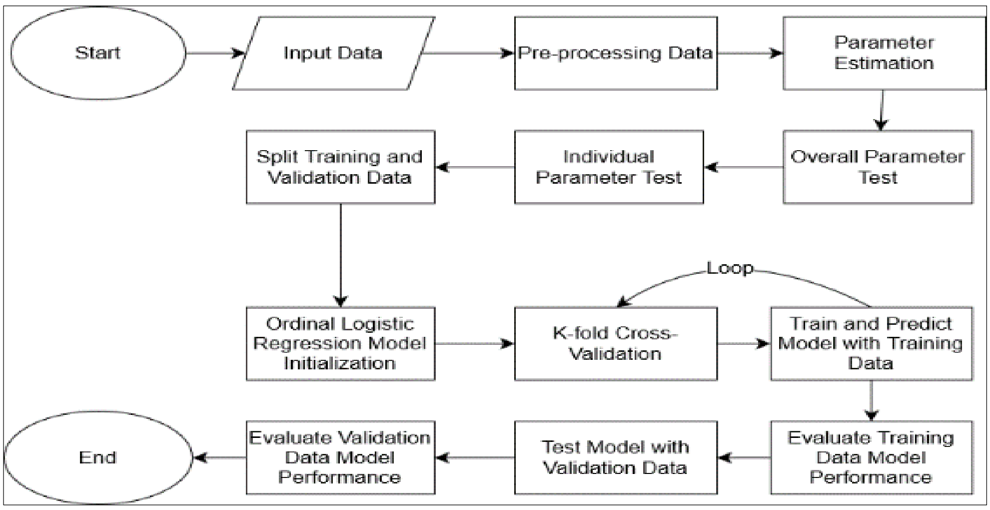

Fig 2: Ordinal Logistics Regression Method Flow Diagram

a. Estimate Parameters

Parameters are estimated using the Maximum Likelihood Estimation Method. The parameter estimation stage is a step in finding optimal parameter values to minimize model errors and match the observed data. This parameter represents the relationship between the predictor and the responsive variable (Nurmita, 2022) [11].

b. Parameter Significance Test

The significance of the parameter is examined as a whole using the probability ratio test and individually using the Wald test to identify the influential variable (Hosmer & Lemeshow, 2000) [9].

1. Test Overall

The overall test is used to evaluate the impact of predictor variables together on the response variable. The following hypothesis is used:

H_0: The predictor variable (X) does not affect the model

$$\beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_1: One or more predictor variables (X) affect the model. $\beta_r \neq 0$ with r = 1, 2,..., p; p = number of predictors

$$X^2_{\text{Calculate}} = -2\, log \left(\frac{\text{likelihood without independent variable}}{\text{likelihood with independent variables}}\right)$$

Test criteria: Reject H_0 when X^2 calculated > $X^2(\alpha,p)$ or p-value (sig) < α.

2. Individual Test

This test is used to test whether each parameter in the model contributes significantly to the response variable. The following hypothesis is used:

H_0: $\beta_r = 0$; Predictor variable (X) has no strong relationship with response variable (Y).

H_1: $\beta_r \neq 0$, with r = 1,2,…,p; p = number of predictors; predictor variable (X) has a strong relationship with response variables (Y).

$$W = \left[\frac{\beta_j}{SE\,(\beta_j)}\right]^2$$

W: Wald test statistics.

βj: Parameter estimate (coefficient) for predictor variable j.

SE (βj): Standard error of parameter estimation βj $\sqrt{var\beta_j}$

Test criteria: Reject H_0 when $W^2 > X^2(\alpha,1)$ or p-value (sig) < α.

c. Ordinal Logistic Regression Model

The logistical model used for ordinal response data is often referred to as the cumulative logit model. A cumulative logit is used to evaluate the relationship between dependent variables, which are staged ordinal responses, and a set of independent variables formed by considering significant variables (Yuleoni, 2022) [23]. In a cumulative logit model, the response is layered data represented by numbers 1, 2, 3,..., r, where r is the number of categories of responses that exist. The ordinal nature of the Y response is reflected through cumulative probability. A comparison is made between a cumulative probability that is less than or equal to the r response category on the predictor variable p, represented by the vector X, $P(Y \leq r \mid X)$, and a greater probability of the r-response category, $P(Y \geq r \mid X)$. Ordinal Logistic Regression Opportunities can be expressed as follows:

$$logit\,(P(Y \leq r|x_i)) = \beta_{0r} + \sum_{j=1}^{p} \beta_j X_j$$

This model will be processed using a Python library called 'Mord' which includes four types of models:

1. OrdinalRidge is similar to standard logistic regression with additional L2 regularization so that the model is more stable and generalization better. Suitable for datasets with lots of features and risks of overfitting.
2. LogisticAT (All-Threshold) is that it takes into account all the thresholds between the responsive categories, providing more accurate and detailed predictions. Suitable for high precision prediction needs.
3. Logistic IT (Immediate-Threshold) is only considering the threshold between two consecutive categories. Suitable for simple and fast models with sequential categories.
4. LogisticSE (Squared Error) is using the square error function for model evaluation. Suitable for model evaluation using easy-to-understand square errors.

d. Split Data

The data is divided into two parts, namely training data for cross-validation and validation test data, with an 80:20 ratio.

e. Cross-Validation

80% of training data is used for training models using cross-validation with equal divisions of 5, 10, and 15 folds. Cross-validation is a statistical technique for evaluating and comparing the performance of learning algorithms. It entails dividing the dataset into two parts, one for model training and the other for model testing. The aim is to provide a more objective estimate of the model's ability to generalize on data that has never been seen before (Witten & Frank, 2005) [Error! Reference source not found.].

f. Air Quality Classification

The classification is done using a model that has been developed using cumulative logit functions.

2.4. Model testing

Models that have been trained using k-fold cross-validation with 80% of the training data subsequently tested using 20% of the previously separate validation test data. These data were never seen by the model during the training to evaluate its ability to predict new data and prevent overfitting.

2.5. Model Evaluation

Model evaluation is done using a confusion matrix to see overall performance and detail in classifying each class (Salam, 2023) [17]. The confusion matrix divides the predictions of models into four categories: good, ongoing, unhealthy, and very bad. Classification evaluation metric summaries, such as accuracy, precision, recall, and F1-score, are provided to provide information about the model performance for each class in the data set. Evaluation metrics help in evaluating how well the model can classify each class and identify areas where the model needs to be improved. The result of this confusion matrix will be used to calculate the evaluation metrics. The following table of multi-class matrix confusion can be seen in Table 1.

Table 1: Confusion Matrix Multiclass

		Predicted Values			
		Good	Medium	Unhealthy	Very Unhealthy
Actual Values	Baik	TP	FP	FP	FP
	Medium	FN	TP	FP	FP
	Unhealty	FN	FN	TP	FP
	Very Unhealthy	FN	FN	FN	TP

While accuracy provides a general overview of model performance, there are situations where it alone is insufficient to provide a comprehensive understanding of the model's performance, especially when there is a class imbalance in the dataset. Therefore, precision and recall testing are needed to provide a more comprehensive and contextual evaluation of the performance of the classification model. Here are the formulas used in calculating the model performance evaluation matrix:

a. Accuracy

Accuracy is a classification model's performance metric that shows how well a model can predict classes accurately. Here's the formula:

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

b. Precision

Precision measures the relevance of a model's positive prediction. Precision provides an overview of how accurate a model is in predicting a positive class. Here's the formula:

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

c. Recall

Recall is a model evaluation for identifying all instances that should be in a class. Recall describes how many positive instances can be predicted as positive. Here's the formula:

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

d. F1-Score

The F1-score combines precision and recall into one value, finding a balance between prediction precision and the ability to find as many positive examples as possible. (recall). Here's the formula:

$$\text{F1-Score} = 2\,\frac{Presisi\ x\ Recall}{Presisi + Recall}$$

3. Results

The study focused specifically on the ability of the ordinal logistic regression method to classify air quality levels, as well as providing valuable insights related to the reliability and accuracy of the model. The study conducted several test scenarios, ranging from the pre-processing stage to the model testing stage, to find the best model that can be used as a model of air quality classification.

3.1. Data collection results

The data set used contains DKI Jakarta air quality data from 2012 to 2021. Each year is presented in a separate file with a total of 18,265 pieces of data, covering 9 attributes (X) and 1 class (Y). Measurements were carried out at 5 stations spread across Jakarta: DKI1 (Bundaran HI), DKI2 (Kelapa Gading), DKA3 (Jagakarsa), DKE4 (Lubang Krokaya), and DKI5. (Kebon Jeruk). Here's the data set collected can be seen in Table 2 below:

Table 2: Air Pollution Dataset Collected

File Name	Amount
(ISPU) Year 2012	1830
(ISPU) Year 2013	1825
(ISPU) Year 2014	1825
(ISPU) Year 2015	1825
(ISPU) Year 2016	1830
(ISPU) Year 2017	1825
(ISPU) Year 2018	1825
(ISPU) Year 2019	1825
(ISPU) Year 2020	1830
(ISPU) Year 2021	1825
Total	18265

The data used in this study are on an ordinal or categorical scale of the response variable (Y). The response variables reflect the values of the air quality status category based on ISPU, divided into four categories: good, current, unhealthy, and very unhealthy.

3.2. Data Processing Results

Datasets that were previously divided into several files have now been merged into one for ease of data processing. In this dataset, there are empty values, and some irrelevant attributes have been deleted, such as date, station, max, and critical. This is because the definition of air quality is based on the measured values of parameters such as PM10 (particulates), SO2 (sulfur dioxide), CO (carbon monoxide), O3 (ozone), and NO2 (nitrogen dioxide). To deal with empty values, two approaches are used: delete them or replace them with the mean values of the associated columns. The table can be seen in Table 3.

Table 3: Dataset after Cleaning

Data Cleaning Scenario	Initial Data Amount	Total Final Data
Drop Null Value	18265	15357
Fill Null Values (Mean)	18265	17595

In the above table, the section containing null values with averages reduces the amount of data because some rows have a null value for all the columns. After processing the null value, the non-relevant attribute columns are deleted, leaving 6 columns consisting of 5 parameter attributes and 1 category column. Parameters are PM10 (particulates 10), SO2 (sulfur dioxide), CO (carbon monoxide), O3 (ozone), and NO2 (nitrogen dioxide). The table can be seen in Table 4.

Table 4: Format of Dataset Used

PM10	SO2	CO	O3	NO2	Category
33	30	14	80	5	Medium
56	30	18	37	5	Medium
35	13	18	46		Good
22	15	24	102	7	Unhealthy
44	17	21	63	7	Medium
58	12	25	215	15	Very Unhealthy
….	….	….	….	….	….

After pre-processing the data, a descriptive analysis is performed to understand the characteristics of the data. Here is a bar diagram showing the results of the analysis.

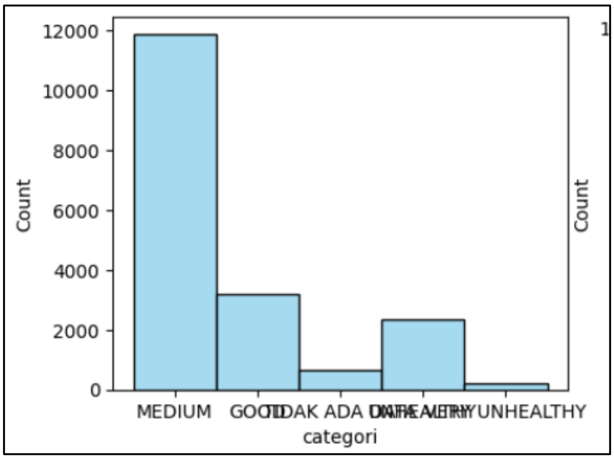

Fig 3: Before Oversampling

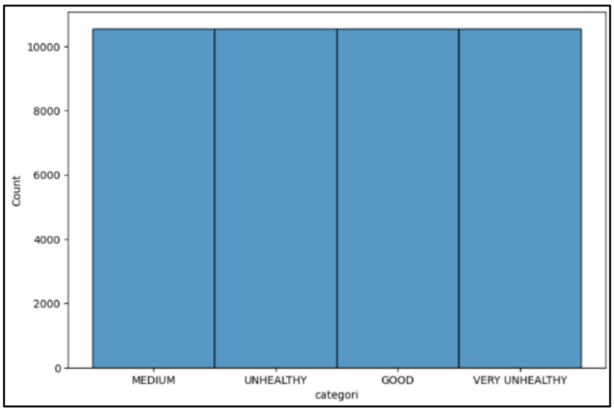

Fig 4: After Oversampling

From the bar diagram in Fig 3 above, there is a data imbalance between categories that can cause problems in modeling because models trained with imbalanced data tend to predict the majority class accurately but ignore the minority class. The solution is oversampling, i.e., adding data from the minority class to balance it with the dominant data. The SMOTE technique is used as an oversampling technique. This technique creates additional samples from the minority class by paying attention to the relationship between the original sample and its neighbors, so that the number of samples from a minority group can be multiplied. The oversampling results can be seen on the diagram in Fig 4.

After oversampling, the amount of data in both scenarios increases as the oversampler increases the number of data in the minority category. Subsequently, duplicate data from the oversampling result is deleted to prevent overfitting. This is critical to ensuring the balance of datasets and preventing the model from overfitting. From these data processing steps, a total of 6 data models from 2 scenarios are obtained at the pre-processing stage. All of these data models will be used from the modeling stage to the model testing and evaluation stage. Here are six data models derived from the data processing process. The table can be seen in Table 5.

Table 5: Model of Dataset Used

Data Cleaning Scenario	Data Model	Total Data
Drop Null Value	Original	15357
	Oversampling	42232
	Drop Duplicates	33948
Fill Null Values (Mean)	Original	17595
	Oversampling	47540
	Drop Duplicates	38212

Datasets that use ordinal logistic regression require the transformation of a categorical response variable into a numerical format. For example, change "Good" to 0, "Medium" to 1, "Unhealthy" to 2, and "Very Unhealthy" to 3. This process replaces a qualitative label with a numeric value in accordance with its ordinal order, which can be done manually or using an encoding label scheme.

3.3. Modeling Results

Before modeling, the initial step is to evaluate the estimates and significance of the parameters. Parameter estimates involve calculating model coefficients from observation data. Parameter estimation in ordinal logistic regression entails determining the coefficient values used to connect independent variables with dependent variables that have sequential levels. The significance test determines whether the coefficient is statistically significant. Here are the results of the parameter significance test, the table can be seen in Table 6 and Table 7:

Table 6: Overall Test Results

Data Model	Without Independent Variables	With Independent Variables	Likelihood Results	Critical Value	P-Value	Alpha Value	Criteria
Null Original	-15.494	445.024	921.037	3,8415	0,0	0,05	Parameters affect
Null Oversampling	-57.895	1.336.584	2.788.958	3,8415	0,0	0,05	Parameters affect
Null Drop Duplicates	-46.805	1.077.512	2.248.634	3,8415	0,0	0,05	Parameters affect
Mean Original	-18.293	524.310	1.085.206	3,8415	0,0	0,05	Parameters affect
Mean Oversampling	-66.238	1.447.706	3.027.889	3,8415	0,0	0,05	Parameters affect
Mean Drop Duplicates	-53.481	1.144.333	2.395.629	3,8415	0,0	0,05	Parameters affect

The overall parameter test results show that all parameters in the six data models reject the zero hypothesis, indicating that the parameters have statistical significance. This indicates that these parameters have a significant influence on the response variable.

Table 7: Individual Test Results

Data Model	Parameter	Coefficient Estimation	Standard Error	Wald Result	Critical Value	Alpha Value	P-Value	Criteria
Null Original	PM10	0,0171	0,0003	3.163	3,8415	0,05	0,0	Parameters affect
	SO2	0,0058	0,0004	167	3,8415	0,05	0,0	Parameters affect
	O2	0,0018	0,0004	16	3,8415	0,05	6,47E-05	Parameters affect
	NO2	0,0092	0,0001	3.685	3,8415	0,05	0,0	Parameters affect
	O3	0,0142	0,0005	586	3,8415	0,05	0,0	Parameters affect
Null Oversampling	PM10	0,0286	0,0002	11.424	3,8415	0,05	0,0	Parameters affect
	SO2	0,0030	0,0004	47	3,8415	0,05	7,12E-12	Parameters affect
	O2	0,0026	0,0004	34	3,8415	0,05	5,84E-09	Parameters affect
	NO2	-0,0015	6,68E+09	564	3,8415	0,05	0,0	Parameters affect
	O3	0,0118	0,0005	451	3,8415	0,05	0,0	Parameters affect
Null Drop Duplicates	PM10	0,0276	0,0002	9.145	3,8415	0,05	0,0	Parameters affect
	SO2	0,0045	0,0004	89	3,8415	0,05	0,0	Parameters affect
	O2	0,0023	0,0004	23	3,8415	0,05	2,08E-06	Parameters affect
	NO2	0,0005	8,43E+10	42	3,8415	0,05	1,00E-10	Parameters affect
	O3	0,0127	0,0005	464	3,8415	0,05	0,0	Parameters affect
Mean Original	PM10	0,0177	0,0002	3.606	3,8415	0,05	0,0	Parameters affect
	SO2	0,0070	0,0004	250	3,8415	0,05	0,0	Parameters affect
	O2	0,0015	0,0004	12	3,8415	0,05	4,08E-04	Parameters affect
	NO2	0,0099	0,0001	4.390	3,8415	0,05	0,0	Parameters affect
	O3	0,0141	0,0005	630	3,8415	0,05	0,0	Parameters affect
Mean Oversampling	PM10	0,0277	0,0002	11.721	3,8415	0,05	0,0	Parameters affect
	SO2	0,0051	0,0004	142	3,8415	0,05	0,0	Parameters affect
	O2	3,51E+10	0,0004	0,0067	3,8415	0,05	9,34E-01	Parameters do not affect
	NO2	-0,0012	6,39E+10	359	3,8415	0,05	0,0	Parameters affect
	O3	0,0049	0,0004	109	3,8415	0,05	0,0	Parameters affect
Mean Drop Duplicates	PM10	0,0270	0,00027	9.559	3,8415	0,05	0,0	Parameters affect
	SO2	0,0065	0,0004	199	3,8415	0,05	0,0	Parameters affect
	O2	-5,78E+11	0,0004	0,0160	3,8415	0,05	8,99E-01	Parameters do not affect
	NO2	0,0010	8,13E+10	169	3,8415	0,05	0,0	Parameters affect
	O3	0,0071	0,0005	187	3,8415	0,05	0,0	Parameters affect

From the individual test results, it was seen that of the 6 data models tested, 2 of them had one non-significant parameter, namely O2, in the data models Mean_Oversampling and Mean_HapusDuplicat. Though not statistically significant, these parameters remain important because they provide valuable insights into the factors that affect air quality. By using these parameters, the model can become more holistic and comprehensive in taking into account these factors, although they are statistically significant.

Of the previous six data models, each is divided by an 80:20 ratio, where 80% will be used as training data and 20% as validation test data. Here's a table of the data splitting done can be seen in Table 8.

Table 8: Number of Datasets After Data Split

Scenario Data	Data Model	Volume of Data	Data splitting	
			Training (80%)	Validate (20%)
Drop Null Value	Original	15357	12285	3072
	Oversampling	42232	33785	8447
	Drop Duplicates	33948	27158	6790
Fill Null Values (Mean)	Original	17595	14076	3519
	Oversampling	47540	38032	9508
	Drop Duplicates	38212	30569	7643

The next step is to build a model and perform k-fold cross-validation for training and prediction. The model used comes from the Python library 'Mord'. Next, k-fold cross-validation with k values = 5, 10, and 15 is used to evaluate the model. 80% of training datasets are used for training and testing, where each iteration divides the data into test and training data. This process is repeated as many times as has been specified. The table can be seen in Table 9.

Table 9: Scenario of Testing Model

Model	K-fold	Model	K-fold	Model	K-fold	Model	K-fold
Ordinal Ridge	5	Logistic AT	5	Logistic IT	5	Logistic SE	5
	10		10		10		10
	15		15		15		15

3.4. Model Testing Results

At the test stage, trained models will be tested to measure their ability to classify air quality. The 20% validation test data that was previously separated and never seen by the model was used to test the model, with the aim of avoiding overfitting and finding the best model. This test has a total of 72 experiments from 2 data scenarios and 6 data models that have been made before. A model test results table with validation data is presented in Table 10 below.

Table 10: Testing Results

No	Scenario Data	Data Model	Model	K-Fold	Training Accuracy	Validation Test Accuracy
1	Drop Null Value	Original	Ordinal Ridge	5	0,8557	0,8516
2				10	0,8559	0,8551
3				15	0,8559	0,8564
4			Logistic AT	5	0,8568	0,8587
5				10	0,8572	0,8580
6				15	0,8574	0,8590
7			Logistic IT	5	0,8569	0,8590
8				10	0,8573	0,8590
9				15	0,8573	0,8593
10			Logistic SE	5	0,8567	0,8580
11				10	0,8569	0,8570
12				15	0,8576	0,8580
13		Oversampling	Ordinal Ridge	5	0,8228	0,8252
14				10	0,8224	0,8245
15				15	0,8226	0,8251
16			Logistic AT	5	0,8617	0,8587
17				10	0,8615	0,8584
18				15	0,8615	0,8584
19			Logistic IT	5	0,8628	0,8599
20				10	0,8626	0,8599
21				15	0,8623	0,8598
22			Logistic SE	5	0,8596	0,8577
23				10	0,8598	0,8577
24				15	0,8597	0,8575
25		Drop_Duplicates	Ordinal Ridge	5	0,8120	0,8073
26				10	0,8120	0,8073
27				15	0,8120	0,8085
28			Logistic AT	5	0,8392	0,8325
29				10	0,8392	0,8326
30				15	0,8393	0,8325
31			Logistic IT	5	0,8407	0,8335
32				10	0,8403	0,8337
33				15	0,8405	0,8338
34			Logistic SE	5	0,8371	0,8315
35				10	0,8374	0,8309
36				15	0,8373	0,8312

37	Fill Null Values (Mean)	Original	ordinal Ridge	5	0,8432	0,8448
38				10	0,8427	0,8459
39				15	0,8429	0,8442
40			Logistic AT	5	0,8445	0,8451
41				10	0,8448	0,8471
42				15	0,8445	0,8465
43			Logistic IT	5	0,8449	0,8456
44				10	0,8448	0,8479
45				15	0,8447	0,8473
46			Logistic SE	5	0,8447	0,8445
47				10	0,8447	0,8456
48				15	0,8446	0,8454
49		Oversampling	Ordinal Ridge	5	0,8009	0,8119
50				10	0,8009	0,8115
51				15	0,8008	0,8119
52			Logistic AT	5	0,8471	0,8545
53				10	0,8470	0,8544
54				15	0,8467	0,8548
55			Logistic IT	5	0,8485	0,8568
56				10	0,8486	0,8568
57				15	0,8484	0,8564
58			Logistic SE	5	0,8447	0,8525
59				10	0,8445	0,8527
60				15	0,8446	0,8528
61		Drop_Duplicates	Ordinal Ridge	5	0,7934	0,7880
62				10	0,7934	0,7876
63				15	0,7933	0,7873
64			Logistic AT	5	0,8248	0,8241
65				10	0,8246	0,8244
66				15	0,8248	0,8244
67			Logistic IT	5	0,8262	0,8262
68				10	0,8262	0,8261
69				15	0,8263	0,8257
70			Logistic SE	5	0,8234	0,8219
71				10	0,8233	0,8219
72				15	0,8234	0,8219

From the table, it is concluded that the best model obtained is a model from a data scenario that removes null values using data oversampling, a LogisticIT model, and k-fold cross-validation with K = 5. This model has a training accuracy of 0.8628, or 86.28%, and a validation test of 0.8599, or 85.99%.

3.5. Model Evaluation Results

The model was evaluated using a confusion matrix, a multiclass matrix. The results were used to calculate evaluation metrics such as accuracy, precision, recall, and F1-Score. In addition, the training graphs and validation tests of the best models were also displayed to give a more complete picture. Here's the best model evaluation can be seen in Fig 5 and Fig 6.

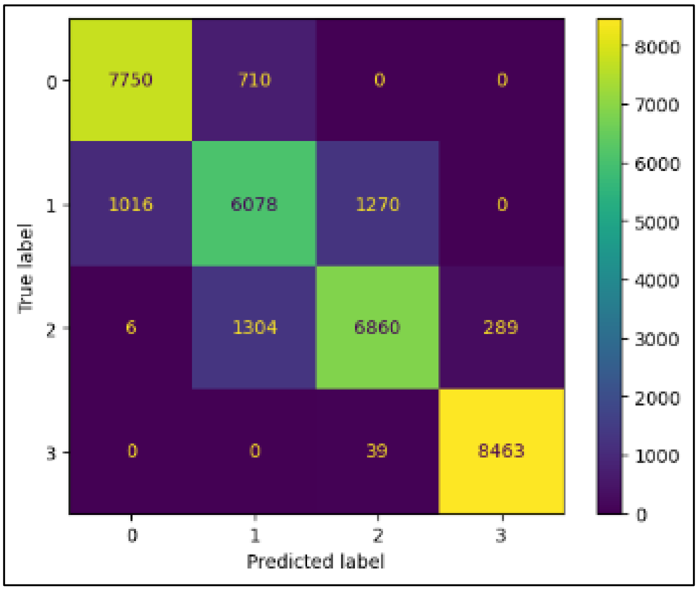

Fig 5: Confusion Matrix Training

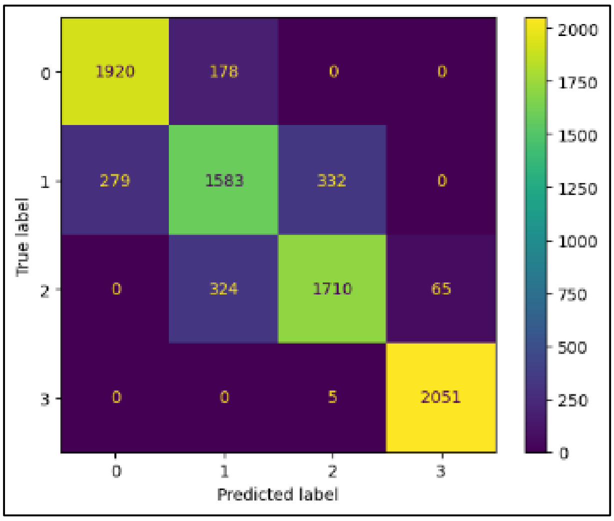

Fig 6: Confusion Matrix Validation Test

After the confusion matrix is formed, the next step is to create an evaluation matrix to provide additional information about the best model performance. The result of evaluation metrics can be seen in Table 11 and Table 12.

Table 11: Training Evaluation Matrix

Class	Precision	Recall	F1-Score
Good	0.88	0.92	0.90
Currently	*0.75*	*0.73*	0.74
Unhealthy	0.84	0.81	0.83
Very Unhealthy	0.97	1.00	0.98
Average	***0.86***	***0.86***	**0.86**
Training Accuracy			**0.8628**

Table 12: Validation Test Evaluation Matrix

Class	Precision	Recall	F1-Score
Good	0.87	0.92	0.89
Currently	0.76	0.72	0.74
Unhealthy	0.84	0.81	0.82
Very Unhealthy	0.97	1.00	0.98
Average	0.86	0.86	0.86
Validation Test Accuracy			**0.8599**

After that, the model will be compared with the accuracy of the training with a validation test to see its performance. The bar graph below displays the average accuracy of training (blue) and validation (orange). This graph helps in comparing the relative performance between training and validations. The graph can be seen in Fig 7.

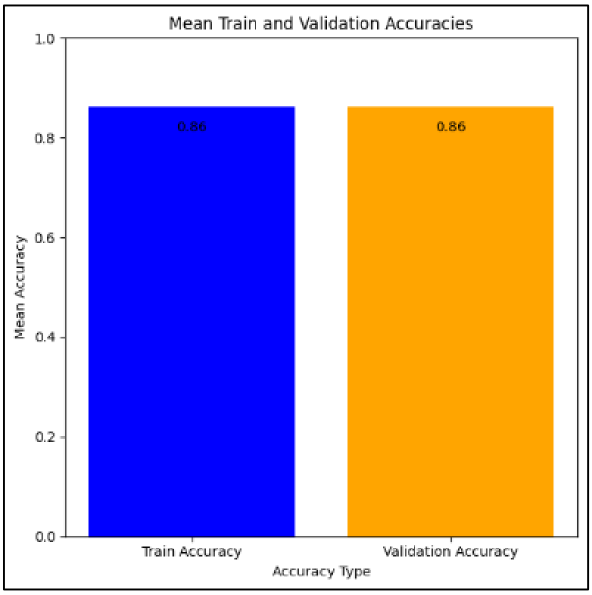

Fig 7: Accuracy Comparison Graph

If the training accuracy is much higher than the validation accuracy, this may indicate overfitting, in which the model is too good at studying patterns from the training data and cannot generalize well to new data. On the contrary, if the verification accuracy is higher or comparable to the training precision, it indicates that the model can generalize properly. If you look at the model's accuracy result of 0.8628 for cross-validation and 0.8599 for validation data, the small difference between the accuracy of cross-validation and the validation test data indicates that the model has good performance in classifying data, indicating its ability to make consistent predictions.

4. Discussion

Based on 72 test experiments, it was found that some of the best models came from pre-processing results with a null-value deletion scenario and data being over-sampled. It can then be concluded that the use of data cleaning techniques with the elimination of null values is more appropriate and correctly applied to the research data. Later, data processing techniques such as oversampling using SMOTE are also suitable for use in this research because they can deal with class imbalance problems in a data set. Thus, oversampling becomes a more appropriate strategy for addressing class imbalances and improving the quality of predictive models.

The LogisticIT model is the best for this research data because of its superior ability to handle sequential categories, making it perfect for models that require simplicity and speed. Of the many tests carried out, LogisticIT yielded consistent and satisfactory results in terms of accuracy, precision, recall, and f1-score compared to other models such as OrdinalRidge, logisticAT, and LogisticSE. The advantages of logisticIT lie in its ability to accurately calculate the threshold between two consecutive categories, which is important in classifying ordinal data. Furthermore, the model offers good computational efficiency and ease in interpreting results, making it the best choice for ordinal analysis in this research context. But bear in mind that each model has its own advantages that can be more suitable for different situations or datasets, so the selection of the model should be done taking into account the context and the goal of the research as a whole. This research uses the entire model as a result of finding the best matching model used in the research data.

This research divides the data into 80% for cross-validation training and 20% for validation tests to ensure that the resulting model is not only optimal based on training data but also has good generalization capabilities against data that has never been seen before. Cross-validation aids in the effective selection and setting of hyperparameter models, while the final validation test with 20% data provides an objective assessment of model performance in real-world scenarios, ensuring that models do not overfit and have reliable performance. The results of this study showed the best model accuracy of 0.8628 for cross-validation and 0.8599 for validation test data, with a small difference between the two, suggesting that the model has the ability to generalize well on new data that has never been seen before. This suggests that a model is not only optimal based on training data but also capable of providing consistent and reliable predictions based on the new data. Therefore, the resulting model can be considered a stable and reliable model for use in real-world applications.

The implementation of the Ordinal Logistics Regression method for air quality classification has shown successful results. Based on the results obtained, this study has not surpassed previous research performance in terms of accuracy, precision, recall, and F1-score. However, the ordinary logistics regression classification method used shows significant potential for analyzing air quality. With a training accuracy of 0.8628 and a validation data test accuracy of 0.8599, as well as a precision, recall, and F1 score of 0.86, this method has demonstrated good ability in generalizing data and coping with data variation. In addition, this method offers advantages in terms of model interpretability and a deeper understanding of the relationship between predictor variables and air quality categories. Therefore, this research is expected to make a significant contribution to the development of a reliable and transparent classification model for air quality analysis. With further parameter perfection and optimization, this model has the potential to improve performance and produce more competitive results in the future.

5. Conclusion

The research focuses on finding the optimal model from the many test scenarios carried out to find classification models that have optimal performance in conducting air quality classifications. This study compared the accuracy of training using cross-validation with the accuracy of validation data testing using an 80:20 data division ratio. Data cleaning results obtained two data scenarios, i.e., delete null values and contain null with average values. Oversampling is done to address data imbalances using the SMOTE (Synthetic Minority Over-sampling Technique) technique to prevent overfitting of the model. The research data sets used after processing the data are divided into six data models, each of which will be used entirely in modeling and testing. Based on the results of the tests, the optimal model was obtained in conducting the classification using five significant parameters: PM10 (particulates 10), SO2 (sulfur dioxide), CO (carbon monoxide), O3 (ozone), and NO2 (nitrogen dioxide) that had previously passed the probability ratio test and the Wald test. Using 4 models from the Mord library, such as OrdinalRidge, LogisticAT, logisticIT, and logisticSE, as well as using cross-validation for model training with fold values 5, 10, and 15, There were 72 test trials to find the optimal model, and the model was obtained that had the best and highest accuracy value. The optimal model was achieved by using data scenarios deleting null values, data models that had been oversampled, the LogisticIT model, and a K-fold value of 5. The optimum model had a cross-validation training accuracy value of 0.8628, or 86.28%, and a validation test data accuracy score of 0.8599, or 85.99%. As well as the respective precision values of 0.86, recall 0.86, and F1-Score 0.86. Both the cross-validation training and the validation data test of the developed ordinal logistic regression model show that it works well enough when dealing with different types of data. The very small difference in values between the two suggests that the model is able to generalize data well, thus giving confidence in its reliability in predicting air quality.

6. References

1. Addini FF, Haryanto D, Maolani RA. Klasifikasi Tingkat Risiko Kerugian Kecelakaan berdasarkan Karakteristik Pengemudi dengan Analisis Regresi Logistik Ordinal. Jurnal Matematika Integratif.

2022;18(2):167-177.
DOI: 10.24198/jmi.v18.n2.41317.167-177.
2. Agista PI, Gysdini N, Maharani MDD. Analisis Kualitas Udara Dengan Indeks Standar (ISPU) And The Distribution Of Pollutant Levels In DKI. Jurnal Penelitian dan Pengembangan Sains dan Humaniora. 2020;2:39-57.
3. Amalia A, Zaidiah A, Isnainiyah IN. Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika). 2022;7(2):496-507.
DOI: 10.33387/jiko.v4i2.2871.
4. Astriyani M, Laela IN, Lestari DP, Anggraeni L, Astuti T. Analisis Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma C.45. JuSiTik: Jurnal Sistem dan Teknologi Informasi Komunikasi. 2023;6(1):36-41. doi: 10.32524/jusitik.v6i1.790.
5. Etriyanti E, Syamsuar D, Kunang N. Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa. Telematika. 2020;13(1):56-67.
DOI: 10.35671/telematika.v13i1.881.
6. Handayani AS, Soim S, Agusdi TE, Rumiasih, Nurdin A. Klasifikasi Kualitas Udara Dengan Metode Support Vector Machine. JIRE (Jurnal Informatika & Rekayasa Elektronika). 2020;3(2):187-199.
7. Hasiman F. Mencermati Polusi Udara Jakarta. Kompas.Id; c2023. Available from: https://www.kompas.id/baca/opini/2023/09/20/mencermati-polusi-udara-jakarta.
8. Henri J. Klasifikasi Kualitas Udara Menggunakan Algoritma Learning Vector Quantization Di Kota Pekanbaru. Jurnal Penelitian dan Pengembangan Sains dan Humaniora. 2021;2:39-57.
9. Hosmer DW, Lemeshow S. Applied Logistic Regression. In Applied Logistic Regression; c2000.
DOI: 10.1002/0471722146.
10. Menteri Lingkungan Hidup Dan Kehutanan Republik Indonesia. Peraturan Menteri Lingkungan Hidup Dan Kehutanan Republik Indonesia Nomor P.14/Menlhk/Setjen/Kum.1/7/2020 Tentang Indeks Standar Pencemar Udara. In: Permen LHK Nomor 14 Tahun 2020 Tentang Indeks Standar Pencemar Udara (ISPU); c2020. p. 1-16.
11. Nurmita A. Pemodelan Persepsi Mahasiswa Uin Suska Riau Terhadap Vaksinasi COVID-19 Menggunakan Ordinal Logistic Regression. In Repository UIN Suska; c2022.
12. Presiden Republik Indonesia. Peraturan Pemerintah Republik Indonesia Nomor 41 Tahun 1999 Tentang Pengendalian Pencemaran Udara. Demographic Research; c1999. p. 4-7.
13. Putra AE, Rismawan T, Rekayasa J, Komputer S. Klasifikasi Kualitas Udara Berdasarkan Indeks Standar Pencemaran Udara (ISPU) Menggunakan Metode Fuzzy Tsukamoto. Jurnal Komputer dan Aplikasi. 2023;11(2):190-196.
14. Putra IGNS, Susilawati M, Gautama IPW. Klasifikasi Tingkat Kesejahteraan Keluarga Di Kecamatan Sidemen Menggunakan Bootstrap Aggregating (Bagging) Regresi Logistik Ordinal. Jurnal Penelitian dan Pengembangan Sains dan Humaniora. 2023;12(2):121-131.
15. Putri LA, Suwanda. Implementasi Metode Artificial Neural Network (ANN) Algoritma Backpropagation untuk Klasifikasi Kualitas Udara di Provinsi DKI Jakarta Tahun 2021. Bandung Conference Series: Statistics. 2023;3(2):184-191. DOI: 10.29313/bcss.v3i2.7826.
16. Raharjo B. Pembelajaran Mesin. In Yayasan Prima Agus Teknik; c2021. p. 1-315.
17. Salam A. Perbandingan Klasifikasi Citra CT-Scan Kanker Paru-Paru Menggunakan Contrast Stretching Pada CNN dengan EfficientNet-B0. KLIK: Kajian Ilmiah Informatika dan Komputer. 2023;4(3):1341-1351. doi: 10.30865/klik.v4i3.1448.
18. Sang AI, Sutoyo E, Darmawan I. Analisis Data Mining untuk Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma Decision Tree dan Support Vector Machine. E-Proceeding of Engineering. 2021;8(5):8954-8963.
19. Toha A, Purwono P, Gata W. Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV. Buletin Ilmiah Sarjana Teknik Elektro. 2022;4(1):12-21.
DOI: 10.12928/biste.v4i1.6079.
20. Jayadi BV, Handhayani T, Lauro MD. Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta. Jurnal Ilmu Komputer dan Sistem Informasi. 2023, 11(2). DOI: 10.24912/jiksi.v11i2.26006.
21. Fleischer NL, Merialdi M, van Donkelaar A, Vadillo-Ortega F, Martin RV, Betran AP, Souza JP, Marie S. O´ Neill. Outdoor air pollution, preterm birth, and low birth weight: analysis of the world health organization global survey on maternal and perinatal health. Environmental health perspectives. 2014;122(4):425-430.
22. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Morgan Kaufmann; c2005. Available from: http://books.google.com/books?hl=en&lr=&id=QTnOcZJzlUoC&oi=fnd&pg=PR17&dq=Data+Mining+Practical+Machine+Learning+Tools+and+Techniques&ots=3gpDdrWiOc&sig=TZS7G8l1eXSa2SpAvfD6aBoJ2lw.
23. Yuleoni T. Pemodelan Akreditasi Sekolah Menengah Atas Menggunakan Ordinal Logistic Regression (Studi Kasus: Kota Pekanbaru). In Repository UIN Suska; c2022.
24. Zhafirah D. Penanganan Imbalance Data Dengan Random Oversampling (ROS) Pada Klasifikasi Penderita Diabetes Menggunakan Support Vector Machine (SVM); c2023. Available from: http://digilib.unila.ac.id/.