Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction

by Bambang Suprihatin

symmetry

Article

Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction

Winoto Chandra [1,2], Bambang Suprihatin [3] and Yulia Resti [3,*]

[1] Doctoral Study Program, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Padang Selasa Bukit Besar, Palembang 30139, Sumatera Selatan, Indonesia
[2] Department of Information System, Faculty of Computer Science, Universitas Bina Darma, Jl. Jenderal A. Yani No. 3, Palembang 30111, Sumatera Selatan, Indonesia
[3] Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih, Km.32, Inderalaya 30062, Sumatera Selatan, Indonesia
* Correspondence: yulia_resti@mipa.unsri.ac.id

Abstract: The Air Quality Index (AQI) dataset contains information on measurements of pollutants and ambient air quality conditions at certain location that can be used to predict air quality. Unfortunately, this dataset often has many missing observations and imbalanced classes. Both of these problems can affect the performance of the prediction model. In particular, predictions for the minority class are very important because inaccurate predictions can be fatal or cause big losses. Moreover, the missing data may lead to biased results. This paper proposes the single imputation of the median and the multiple imputations of the *k*-Nearest Neighbor (KNN) regressor to handle missing values of less than or equal to 10% and more than 10%, respectively. At the same time, the SMOTE-Tomek Links address the imbalanced class. These proposed approaches to handle both issues are then used to assess the air quality prediction of the India AQI dataset using Naive Bayes (NB), KNN, and C4.5. The five treatments show that the proposed method of the Median-KNN regressor-SMOTE-Tomek Links is able to improve the performance of the India air quality prediction model. In other words, the proposed method succeeds in overcoming the problems of missing values and class imbalance.

Keywords: air quality; missing values; imbalanced data; median; KNN; SMOTE-Tomek Links

Citation: Chandra, W.; Suprihatin, B.; Resti, Y. Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction. *Symmetry* **2023**, *15*, 887. https://doi.org/10.3390/sym15040887

Academic Editors: Jianqiang Wang and Florentin Smarandache

Received: 28 February 2023
Revised: 30 March 2023
Accepted: 3 April 2023
Published: 9 April 2023

1. Introduction

The problems of missing data and class imbalance often occur in datasets from various fields, for example, arrhythmia [1], financial fraud [2], air pressure systems [3], diabetes [4], and so on. Missing and imbalanced data significantly affect the performance of the classification method, so it is essential to overcome these problems. Missing data is a condition where an observation has no value [5]. Imbalanced data is a condition where the number of instances in a class is very small compared to other classes [6]. Missing data can cause the loss of information in the datasets, and imbalanced data can cause the information in the majority class to become easy to obtain. On the contrary, obtaining information about the minority class becomes challenging. Missing data that are not handled can lead to incorrect analysis results and conclusions. The loss of datasets can affect the accuracy of the classification and may give biased results [7]. In addition, some classification algorithms do not allow missing values in the dataset. Handling missing data requires techniques that obtain a representative value when filling in lost data. Imputation is a technique used to fill in or replace missing data. Imputation techniques to handle missing data are classified into two groups: single imputation and multiple imputations [8]. Single imputation is an imputation technique that provides a specific value that can replace the missing data directly. Single imputation gives a specified value in place of the missing data instantly. At the same time, multiple imputations select a calculated value from several possible responses based on analysis of variance or confidence intervals.

A single imputation is suitable for small amounts of missing data. When the sum of the missing data is slight, this imputation provides a reasonably effective technique. Both single-imputation methods, the mean or the median, can be used to impute missing data with an amount of at most 10% [9,10]. The median imputation involves fewer calculations and provides a specific set of data instead of the missing data [1]. Median imputation has the advantage of dealing with outliers in the observed data and a skewed data distribution, where the data distribution is not symmetrical or is more inclined to one side. However, using the median imputation on large amounts of data can produce biased or unrealistic results, thus providing misleading information [11]. Another weakness is that it reduces variability so that it reduces the estimated error compared with the deletion approach and ignores covariance and correlation with other variables [12]. Calculating the median of many inputs before performing the statistical analysis removes most of the recovered randomness and gives results close to those of simple imputation, containing minor random errors [13].

The multiple imputations are used to generate numerous values and perform statistical analysis. A simple hint adds a random value to restore the randomness lost in the imputation process. The K-Nearest Neighbor (KNN) regressor is one of the multiple-imputation methods [14,15]. The KNN regressor is the same as the classification KNN, which uses the Euclidean distance metric to take as many as k nearest neighbors. The difference is that the KNN classification takes the similarity of the label or class of the k closest neighbors using majority voting. At the same time, the KNN regressor calculates the mean or average value of the k neighbors as the value that fills in the missing data. KNN utilizes the additional information provided by each predictor to maintain the data's original structure. KNN is a non-parametric method and does not require an explanation of the relationship between variables, so it is not readily susceptible to model specification errors [3]. KNN is easy to implement and capable of handling all types of power loss, whether continuous, discrete, ordinal, or nominal. Several studies have shown that the application of a KNN regressor can improve the performance of classification methods [3,14,16].

Apart from missing data, another problem that needs to be addressed is class imbalance. Unaddressed class imbalances can lead to inaccurate predictions for the minority class. An accurate prediction for the minority class is very important because inaccuracy can be fatal or result in very expensive costs [2]. The class imbalance issue can be handled by using data-level and algorithm-level approaches [17], such as oversampling and undersampling [18]. Oversampling is a technique that balances the minority class by duplicating the minority class so that the number of minority classes increases and can avoid overfitting. The weakness of oversampling is that it can cause overfitting [19]. Undersampling is a technique that balances the minority class by eliminating the majority class until the distribution of these classes is balanced. The disadvantage of undersampling is that it can lose valuable data [18]. The Synthetic Minority Oversampling Technique (SMOTE) is a popular method to overcome the weaknesses that exist in oversampling. This technique synthesizes new samples from the minority class by identifying the vector between the sample from the minority class and the sample from the selected neighbor [20]. Several studies show that applying the SMOTE to imbalanced data can improve classification performance [6,21,22], but several studies did not obtain good results [19,20,23,24]. As an alternative, many oversampling methods based on the SMOTE have been developed in recent years [23,24]. One of them is SMOTE-Tomek Links, which combines the SMOTE as an oversampling method and the Tomek Links as undersampling method. The advantages of the SMOTE-Tomek Links method are that it can improve data imbalances more effectively than SMOTE and can improve the accuracy of the minority class [19,23]. Several previous studies have shown that the SMOTE-Tomek Links method can work well for classification [19,25]. In research of DNA methylation classification [25], using the SMOTE-Tomek Links method obtained a metric performance of recall, precision, and $F1$-score above 90%. Likewise, in research [19] detecting the error of an electrical rotating machine, satisfactory performance metrics of above 97% were achieved.

Air Quality Index (AQI) is a dataset that provides information on the results of measurements of pollutants and ambient air quality conditions at a certain location. The dataset can be used to predict air quality. Unfortunately, the dataset often has observations with missing data and uneven class distribution; for example, the India AQI datasets. This dataset has a number of variables with missing observations of less than 10% and others of more than 10%. There are two minority classes in this dataset, namely the Good category and the Severe category. Single imputation using the median is inappropriate to be applied to the India AQI dataset, since the data are split over a certain interval. At the same time, predictions for the minority classes are very important because inaccurate predictions can be fatal in relation to flight schedules, outdoor activities, motorized vehicle control, and so on. The problems of missing data and class imbalance in AQI datasets, especially the India AQI datasets, require a unique approach.

This study integrates the handling of missing observations and imbalanced class in the India AQI dataset. The missing values are handled using the median and KNN regressor, while for class imbalance, the SMOTE-Tomek Links method is used. The median and KNN regressors handle the missing data of less than 10% and more than 10%, respectively. For data with certain intervals in each class, the median approach needs to be adjusted according to the intervals in each class of the related variables. At the same time, KNN is performed by determining the proximity of the distance to each observation. These proposed approaches to handle both issues are then used to predict the air quality of the India AQI dataset using Naive Bayes (NB), KNN, and C4.5. Furthermore, we used five treatments to show the effect of Median-KNN regressor imputation and SMOTE-Tomek Links on the air quality prediction performance of the India AQI dataset. The five treatments are a combination of removing missing data and imputing missing data with SMOTE resampling and SMOTE-Tomek Links, respectively. Our proposed methods are expected to improve the performance of air quality prediction with the missing observations and imbalanced classes of the India AQI dataset.

2. Materials and Methods

The entire workflow of the method proposed in this study is shown in Figure 1. The stages are as follows:

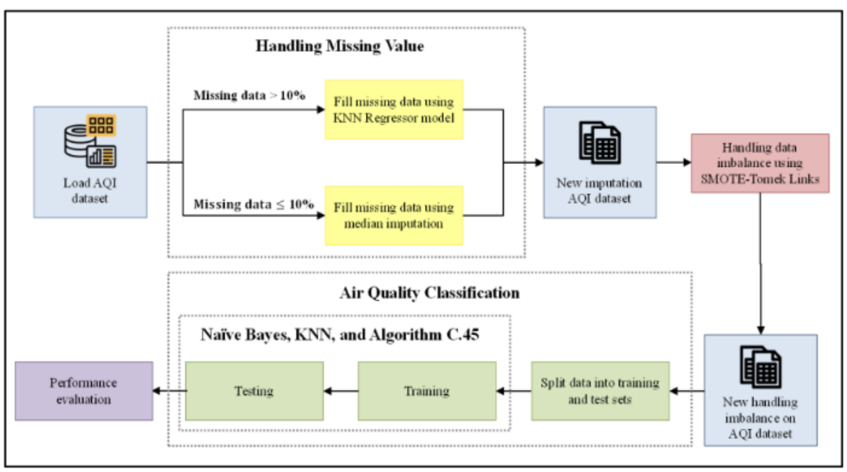


Figure 1. Median-KNN-SMOTE-Tomek Links workflow diagram for dealing with missing and imbalanced data in AQI.

2.1. Air Quality Index Dataset

The dataset used in this study is the India Air Quality Index (AQI) from 2015 to 2020, which was obtained free of charge via the https://www.kaggle.com/datasets/rohanrao/

air-quality-data-in-india, accessed on 1 September 2022. The size of the dataset is 24,850, with eight variables. Seven variables are AQI calculation parameters, and the air quality category is a label calculated based on the AQI value (known as AQI Bucket). The seven variables are PM_{10} (particulate matter 10-micrometer), $PM_{2.5}$ (particulate matter 2.5-micrometer), SO_2 (sulfur dioxide), NO_x (nitric x-oxide), CO (carbon monoxide), O_3 (ozone or trioxygen), and NH_3 (ammonia). These seven variables have ranges of values and a number of missing observations with different percentages (Table 1).

Table 1. Value intervals and percentages of the missing values.

Predictor Variable	Value Intervals	Percentage of Missing Value
$PM_{2.5}$ (μg/m^3)	0.04–914.94	2.73
PM_{10} (μg/m^3)	0.03–917.08	28.52
NO_x (μg/m^3)	0.00–378.24	7.47
NH_3 (μg/m^3)	0.01–352.89	26.30
CO (mg/m^3)	0.00–175.81	1.79
SO_2 (μg/m^3)	0.01–186.08	2.43
O_3 (μg/m^3)	0.01–257.73	3.25

The AQI Bucket variable consists of 6 classes: Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. The distribution of observations in each class is presented in Figure 2.

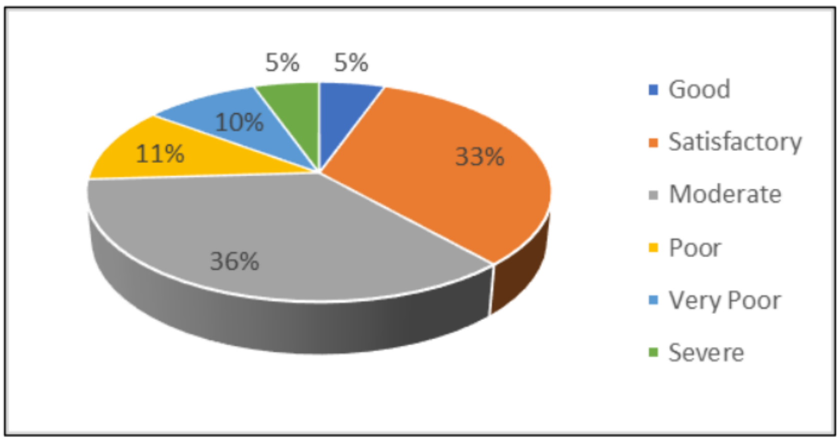


Figure 2. Class distribution of the dataset.

The number of observations in the majority class reached 36% (Moderate category) while in the minority class it was 5% (Good category and Severe category). The two minority classes in air quality prediction are often more important than the other classes because they can affect various aspects of human life. For example, if weather predictions are good, airlines can fly, but if the opposite happens, it will certainly endanger passengers. In this study, training and testing on the India AQI dataset cover data for 2015–2018 and 2019–2020, respectively.

2.2. Handling Missing Values

At this stage, missing values are handled using the imputation technique of filling in or replacing the missing value with the predicted value. Lost data handling consists of median imputation and KNN regressor imputation. Median imputation is used for variables with missing data less than or equal to 10% ($PM_{2.5}$, NO_x, O_3, CO, and SO_2). The KNN regressor imputes missing data for variables with more than 10% missing data (PM_{10} and NH_3).

2.2.1. Median Imputation

For the imputation of the missing observations, the median calculation is not carried out directly on the actual data, but is calculated based on the value range of every variable in each class (AQI category). The median imputation for missing data based on the pollutant concentration intervals obtained in each category is given in Table 2 [26].

Table 2. India Air Quality Index.

AQI Category (Range)	PM_{10} (µg/m³)	$PM_{2.5}$ (µg/m³)	NO_x (µg/m³)	O_3 (µg/m³)	CO (µg/m³)	SO_2 (µg/m³)	NH_3 (µg/m³)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400
Moderate (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800
Poor (201–300)	251–350	91–120	181–280	169–208	10.1–17	381–800	801–1200
Very Poor (301–400)	351–430	121–250	281–400	209–748	17.1–34	801–1600	1201–1800
Severe (401–500)	>430	>250	>400	>748	>34	>1600	>1800

2.2.2. KNN Regressor Imputation

For two variables, PM_{10} and NH_3, estimating the mean is ineffective because the amount of missing data needs to be more significant. So, these two variables are imputed using the KNN regressor method. The KNN regressor uses the similarity of predictor variables between samples to predict the value of missing observations [11]. The algorithm works based on the weighted average of the k-nearest neighbor [27].

The first step in the KNN regressor is calculating the Euclidean distance between the observations containing missing data (x_a) and complete observations (x_b) on the m-th variable [28]. For the i-th observation, where $i = 1, 2, \cdots, p$ and p is the number of missing data,

$$d_i(x_a, x_b) = \sqrt{\sum_{m=1}^{M} \left(x_{a_i} - x_{b_i m}\right)^2} \tag{1}$$

Next, we determine k—the number of closest observations or nearest neighbor used [28]. Furthermore, we impute missing data for the i-th observation using the weighted average of all predictor variables (excluding variables with missing observations) [11].

$$x_i = \frac{\sum_{i=1}^{k} w_i x_{a_i}}{\sum_{i=1}^{k} w_i} \tag{2}$$

where w_i is defined in (3).

$$w_i = \left(\frac{1}{d_i(x_a, x_b)}\right)^2 \tag{3}$$

Finally, the imputation process using the KNN regressor is carried out for every variable that contains missing data by prioritizing those with the least missing data. The imputation of other variables is carried out the same way after imputing all values in a variable is completed.

2.3. Handling Missing Value Imbalance Data Using Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links

The imbalanced data for each class can cause a classification bias towards the majority class while undersampling the minority class [29]. SMOTE is a method to overcome the problem of data imbalance, introduced by Chawla et al. [6], where to synthesize a new sample, random interpolation is carried out between the sample feature space for each target class and its nearest neighbor [30]. This can increase the number of minority classes and help the classifier to increase its generalization capacity [29,30]. Many oversampling methods have been developed using the SMOTE as the basis [24], including SMOTE-Tomek Links [23,24]. This method combines the SMOTE oversampling and Tomek Links undersampling techniques [23]. The SMOTE gives rise to synthetic data for the minority class, and at the same time, Tomek Links removes the data that are identified as Tomek Links from the majority class.

The SMOTE step begins by determining the number of nearest neighbors (k), then calculating the shortest distance between the random data selected from the minority class $\left(x_{c_i}\right)$ and the data of the k-nearest neighbors (x_{k_i}) using the Euclidean distance in Equation (1) [29,31]. Furthermore, based on the closest distance, the synthetic sample data (x_{s_i}) are generated for the minority class using (4) [30]:

$$x_{s_i} = x_{c_i} + r\left(x_{c_i} - x_{k_i}\right) \tag{4}$$

The process is stopped when the data for each class have been balanced [29,31].

The first step in Tomek Links is choosing a pair of samples with the minimum Euclidean distance from the k-nearest neighbors, where each sample comes from a different class $\left(x_g, x_h\right)$. The sample pair $\left(x_g, x_h\right)$ is a Tomek Link, if there is no sample x_k that satisfies the following conditions related to the Euclidean distance,

$$d\left(x_g, x_k\right) < d\left(x_g, x_h\right) \quad \text{or} \quad d\left(x_h, x_k\right) < d\left(x_g, x_h\right) \tag{5}$$

The sample of Tomek Link from the majority class is then removed from the dataset. The process ends if the balance of classes has been reached [23].

2.4. Performance Measure

The performance of the AQI quality prediction model for which missing data and class imbalances are handled using Median-KNN regressor imputation and SMOTE-Tomek Links is measured using accuracy, precision, recall, and $F1$-score, as is written in (6)–(9) [31].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

where TP is correct positive predictions (true positive), FP is wrong positive predictions, TN is correct negative predictions (true negative), and FN is wrong negative predictions (false negative).

Accuracy is used to calculate the accuracy of a classification model [31]. However, the accuracy value is not appropriate for measuring the performance of the classification model on imbalanced data. In this study, other performance measures were used to overcome this problem, such as precision, recall, and $F1$-score [31]. Precision measures the ratio of positive, correctly predicted AQI classes to the total number of positively predicted classes [31]. The recall is used to measure the true positive ratio [31]. Finally, the $F1$-score is the average harmonic of precision and recall [31].

3. Results and Discussion

3.1. Handling Missing Data

The variables with the lowest number of missing values below 10%, based on Table 1, namely $PM_{2.5}$, NO_x, O_3, CO, and SO_2, were imputed using the median value technique from the predetermined AQI category. For example, suppose there is a missing value in the $PM_{2.5}$ attribute, and the AQI Bucket category is Good. Based on Table 2, a range of values is obtained between 0 and 30, so the imputation value using the median technique is $(0 + 30)/2 = 15$. For other missing values, the same method is used, using the median value based on the category of each variable. Furthermore, the missing values above 10%, namely PM_{10} and NH_3, were imputed using the KNN regressor method. NH_3 has fewer missing values than PM_{10}, so NH_3 is imputed using the KNN regressor first by ignoring the PM_{10}. To carry out the training process on the NH_3, the k value is determined first by looking at the most optimal root mean square error (RMSE). The RMSE for the k value measured ranges from 1 to 25 for the NH_3, as can be seen in Figure 3.

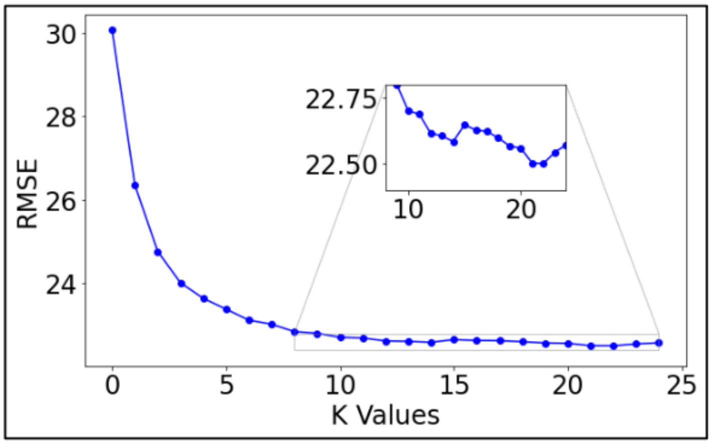


Figure 3. Optimal k values on the NH_3.

At $k = 1$, the highest RMSE value is almost around 30% and continues to decrease towards a value of 22%. The lowest RMSE value was obtained at $k = 23$, so the k value was chosen to be trained on the NH_3 attribute using the KNN regressor. After the NH_3 is filled, the PM_{10} is imputed using the KNN regressor. In the same way, the k value is determined by the PM_{10}. The RMSE results obtained for the k value in the PM_{10} can be seen as shown in Figure 4. For $k = 1$, the highest RMSE value is almost around 42% and continues to decrease towards a value of 36%. The lowest RMSE value was obtained at $k = 9$, so the k value was chosen to be trained on the PM_{10} using the KNN regressor. The results of the imputation process using the KNN regressor are then compared between the predicted value and the actual value, which can be seen as shown in Figure 5.

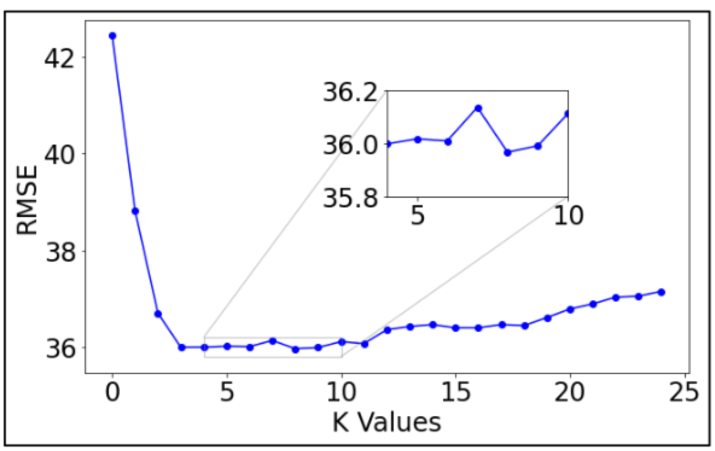


Figure 4. Optimal k values on the PM_{10}.

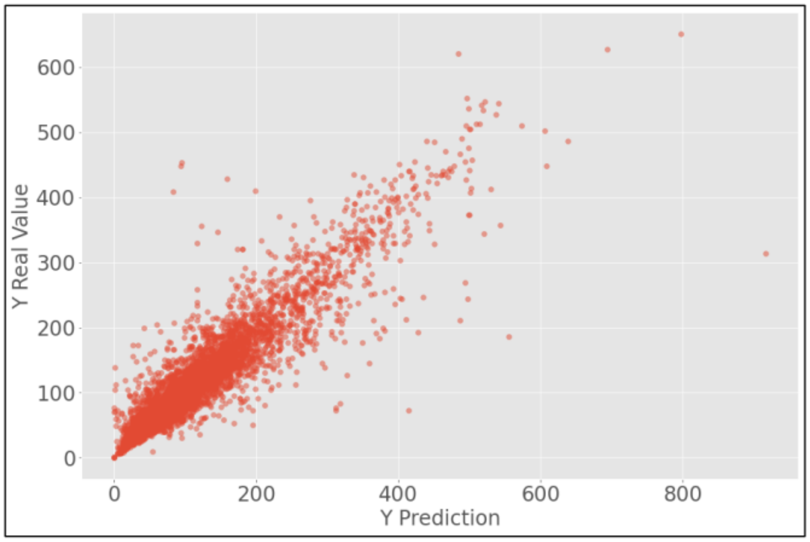
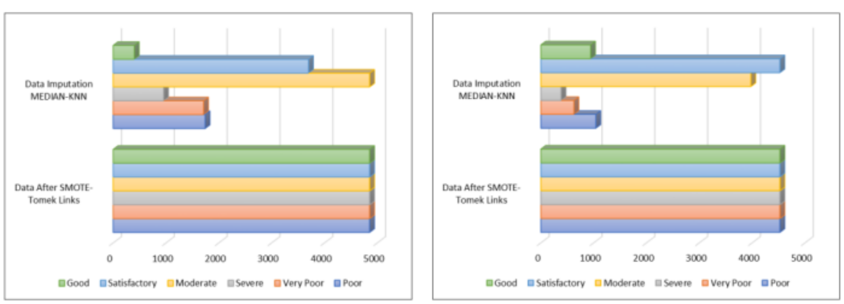


Figure 5. The graph of the comparison of the predicted value and the actual value using the KNN regressor.

In Figure 5, the predicted initial value is still on a straight diagonal line, which means there is a match between the expected and actual values. However, towards the end of the predicted value, some values slightly deviate from the actual value. To see how successful the KNN regressor method is in filling in the missing values on the PM_{10} and NH_3 variables, measurements are made using the accuracy value with the obtained value of 0.8412.

3.2. Handling Imbalance Data

The amount of data in each class in the dataset is not balanced. In this stage, SMOTE-Tomek Links is carried out using the oversampling technique. SMOTE-Tomek Links is used to maintain a balance between the number of classes by increasing the amount of data in the minority class. The class with the most minority data of the training data is the Good class, with a total of 400 data, while the class with the most majority data is the Moderate class, with a total of 4854. Meanwhile, for data testing, the class with the most minority data is the Severe class, with a total of 383 data, while the class with the most majority data is the Satisfactory class, with a total of 4526. A comparison of the data amounts before SMOTE-Tomek Links and after SMOTE-Tomek Links can be seen in Figure 6a,b.



(**a**) (**b**)

Figure 6. Comparison of the amount of data before SMOTE-Tomek Links and after SMOTE-Tomek Links: (**a**) training data; (**b**) testing data.

It can be seen in Figure 6a for the SMOTE-Tomek Links process that the training dataset from the Good class, which initially had 400 data, was replicated for a total of 4854 data, so that the number was balanced with the majority class. In the same way, the SMOTE-Tomek Links process was also carried out for other classes, Poor, Very Poor, Severe, and Satisfactory, so that the final amount of data in the training dataset after SMOTE-Tomek Links was 29,124 data, with 4854 data for each class. On the other hand, as shown in Figure 6b for the test data, the SMOTE-Tomek Links process was also carried out for the Severe class, which was replicated from 382 data to 4526 data. For other classes, the same process was also carried out so that the final amount of data in the test dataset after SMOTE-Tomek Links was 27,156 data, with 4526 data for each class.

3.3. Performance of Prediction Model

The results of applying Median-KNN-SMOTE-Tomek Links to handle missing data and class imbalances are applied to predict air quality. The methods used in this study are Naive Bayes (NB), KNN, and C4.5. The performance metrics used to measure the results of the application of Median-KNN-SMOTE-Tomek Links are accuracy, precision, recall, and *F*1-score. A comparison of the prediction results using the proposed methods can be seen in Table 3. The accuracy of the prediction data after being imputed using Median-KNN and balanced using SMOTE-Tomek Links (Median-KNN-SMOTE-Tomek Links) increases significantly compared with data that are only imputed with Median-KNN, not balanced using SMOTE-Tomek Links (Median-KNN). The increase in accuracy from highest to lowest was 29.16% (C4.5), 19.75% (KNN), and 5.16% (NB). In line with this increase, C4.5 has the highest accuracy compared with the other two methods, with a metric value of 100%.

Table 3. The result of the accuracy of the classification process on research data using the NB, KNN, and C4.5 algorithms.

Methods	Data Imputation Median-KNN (%)	Data after SMOTE-Tomek Links (%)
NB	68.80	73.96
KNN	76.25	96.45
C4.5	70.84	100

The other performance metrics measured in each class, such as precision, recall, and *F*1-score, for each prediction method can be seen in Figures 7–9.

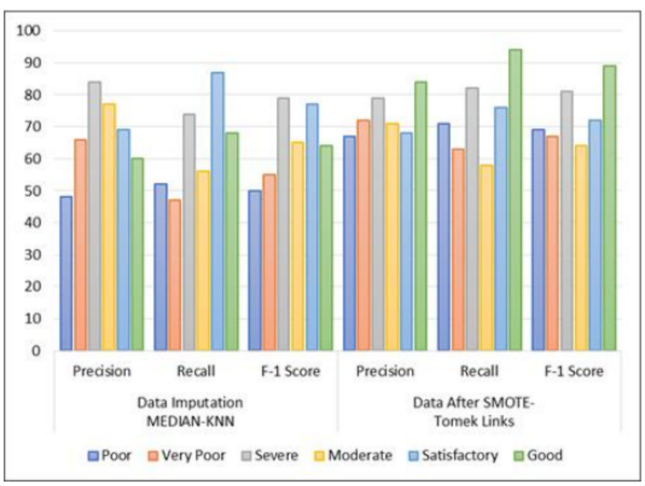


Figure 7. The precision, recall, and *F*1-score using the NB method.

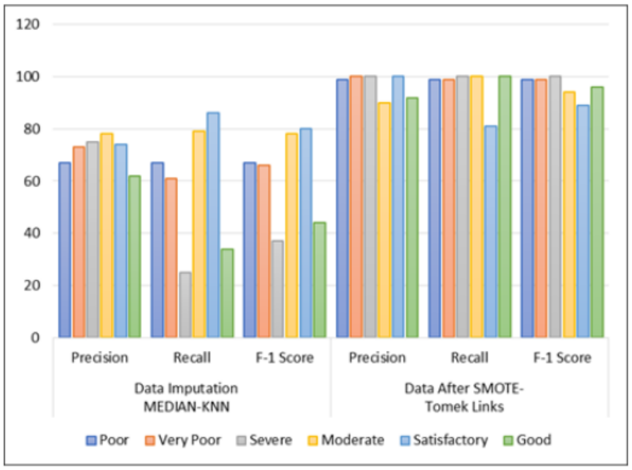


Figure 8. The precision, recall, and $F1$-score using the KNN method.

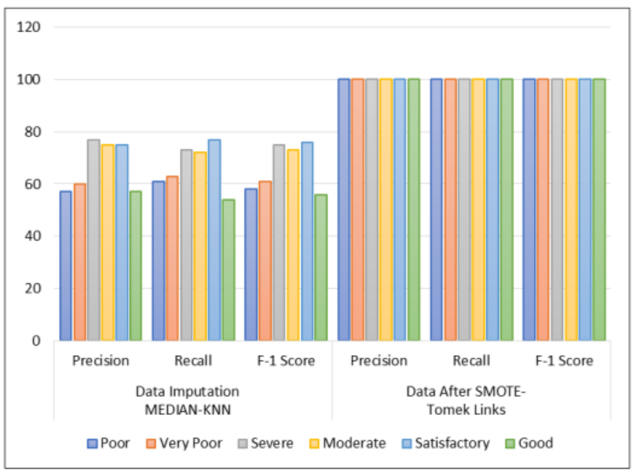


Figure 9. The precision, recall, and $F1$-score using the C.45 algorithm.

Figures 7–9 show that the prediction results using data after SMOTE-Tomek Links (Median-KNN-SMOTE-Tomek Links) exhibit an increase in the other metrics, precision, recall, and $F1$-score, in all three methods compared with the data that were only imputed using Median-KNN. In data imputation using Median-KNN, the average precision ranges from 48% to 89%, the average recall ranges from 25% to 87%, and the average $F1$-score ranges from 37% to 80%. Meanwhile, the data after SMOTE-Tomek Links produce an average precision ranging from 67% to 100%, average recall ranging from 58% to 100%, and average $F1$-score ranging from 64% to 100%.

Furthermore, we used five treatments to show the effect of Median-KNN regressor imputation and SMOTE-Tomek Links resampling in dealing with missing data and class imbalance, respectively (Table 4). For the first treatment, all missing values are discarded, and class imbalances are ignored. All missing values are discarded in the second treatment, and class imbalances are handled with SMOTE. All missing values are discarded in the third treatment, and class imbalances are handled with SMOTE-Tomek Links. The missing values are handled with the Median-KNN regressor in the fourth treatment, and class imbalances are handled with SMOTE. For the fourth treatment, the missing values are handled with the Median-KNN regressor, and class imbalances are handled with SMOTE-Tomek Links.

Table 4. Comparison of the predictive performance of the five proposed treatments.

Method	Treatment	Precision (%)	Recall (%)	*F*1-Score (%)	Accuracy (%)
NB	1	70.14	70.42	69.89	74.99
	2	74.13	73.59	73.48	73.59
	3	74.76	73.99	73.94	73.97
	4	67.35	66.67	66.22	66.67
	5	67.17	66.67	66.33	73.96
KNN	1	74.34	65.29	67.35	75.97
	2	76.62	76.09	75.9	76.09
	3	76.28	75.77	75.67	75.75
	4	74.87	74.53	74.46	74.53
	5	96.83	96.50	96.17	96.64
C4.5	1	65.40	66.41	65.71	69.5
	2	68.38	64.55	64.95	64.55
	3	67.81	64.88	65.27	64.9
	4	70.52	69.45	69.67	69.45
	5	100.00	100.00	100.00	100.00

We show that the increase in predictive performance with Naive Bayes from treatment 1 to treatment 3 is due to the weakening of the correlations of most of the predictor variables (Tables 5–7). In these three datasets, the more the variable correlations are weakened, the more predictive performance increases with Naive Bayes. In other words, fulfilling more naive assumptions between predictor variables can further improve the predictive performance.

Table 5. Correlation in the first treatment.

Correlation	$PM_{2.5}$	PM_{10}	NO_x	NH_3	CO	SO_2	O_3
$PM_{2.5}$	1	0.8599	0.5861	0.5855	0.2851	0.3705	0.3620
PM_{10}		1	0.6086	0.5844	0.2870	0.4317	0.3572
NO_x			1	0.4394	0.2499	0.2970	0.2980
NH_3				1	0.3247	0.2136	0.2776
CO					1	0.0809	0.0371
SO_2						1	0.1683
O_3							1

Table 6. Correlation in the second treatment.

Correlation	$PM_{2.5}$	PM_{10}	NO_x	NH_3	CO	SO_2	O_3
$PM_{2.5}$	1	−0.0042	−0.0006	−0.0055	0.0076	0.0061	−0.0105
PM_{10}		1	−0.0095	−0.0207	−0.0012	0.0219	−0.0312
NO_x			1	−0.0043	−0.0147	−0.0168	0.0176
NH_3				1	−0.0178	−0.0057	0.0068
CO					1	0.0086	−0.0260
SO_2						1	−0.0102
O_3							1

The same thing happened to the data of the fourth and fifth treatments, where the missing data were imputed with the Median-KNN regressor and then the imbalanced classes were resampled using SMOTE and SMOTE-Tomek Links. The correlation between predictor variables increased from the fourth treatment to the fifth treatment, causing the prediction performance using the NB method to decrease (Tables 8 and 9).

Table 7. Correlation in the third treatment.

Correlation	$PM_{2.5}$	PM_{10}	NO_x	NH_3	CO	SO_2	O_3
$PM_{2.5}$	1	−0.0083	−0.0206	−0.0068	0.0482	0.0170	−0.0071
PM_{10}		1	−0.0071	−0.0270	0.0366	0.0131	−0.0075
NO_x			1	−0.0157	−0.0201	−0.0087	−0.0105
NH_3				1	−0.0010	0.0216	−0.0015
CO					1	0.0228	−0.0111
SO_2						1	0.0032
O_3		0.0143					1

Table 8. Correlation in the fourth treatment.

Correlation	$PM_{2.5}$	PM_{10}	NO_x	NH_3	CO	SO_2	O_3
$PM_{2.5}$	1	0.0143	0.0039	0.0240	−0.0046	0.0158	0.0104
PM_{10}		1	−0.0049	−0.0067	0.0096	−0.0059	−0.0068
NO_x			1	0.0025	−0.0055	−0.0014	0.0036
NH_3				1	0.0000	−0.0262	−0.0035
CO					1	0.0228	−0.0128
SO_2						1	−0.0148
O_3							1

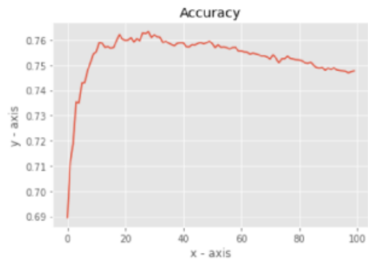
Table 9. Correlation in the fifth treatment.

Correlation	$PM_{2.5}$	PM_{10}	NO_x	NH_3	CO	SO_2	O_3
$PM_{2.5}$	1	0.8939	0.4380	0.5018	0.0190	0.0450	0.0648
PM_{10}		1	0.4468	0.4304	0.0673	0.0678	0.0834
NO_x			1	0.1213	0.3086	0.1951	0.0759
NH_3				1	0.0338	−0.0437	0.1066
CO					1	0.3535	0.1563
SO_2						1	0.1627
O_3							1

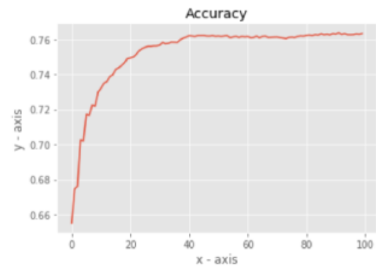
In implementing the KNN method for the five treatments, we explored a number of values of k in the range 1–100, determining which produces the highest accuracy. The highest k values in the training dataset for each treatment were 30, 93, 61, 55, and 2, respectively (Figure 10). Especially for the fifth treatment, the accuracy value was stable at 100% for $k = 1$ to 100.

The use of Median-KNN regressor imputation and SMOTE-Tomek Links resampling, proposed in this work to improve air quality prediction performance, obtained significant results using the KNN and C4.5 methods. Even with the C4.5 method, the model performance reached 100% on all metrics. As non-parametric methods, both the KNN and C4.5 methods do not consider the effect of the correlation between predictor variables, so adding observations to balance classes positively affects predictive performance.

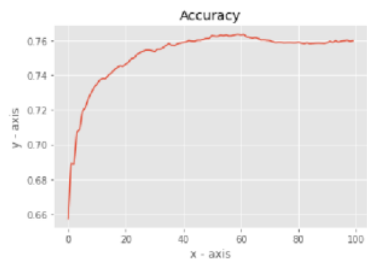
Furthermore, improved evaluation results for the three proposed methods with data obtained before SMOTE-Tomek Links and after SMOTE-Tomek Links are compared with those of other studies. Table 10 shows that not all studies that handled imbalanced classes before using SMOTE-Tomek Links had improved performance on all metrics. An increase in performance is marked with a positive value, while a decrease is marked with a negative value in accuracy, precision, recall and $F1$-Score. In credit card fraud detection [20], with a class ratio of 0.17:99.83, the class imbalance handled with SMOTE-Tomek Links resulted in a recall of 94.94% from the previous 74.83%. The increase in this metric by 20.11% was not followed by an increase in accuracy, which actually decreased from 99.17% to 98.32%. However, the decline did not reach 1%.



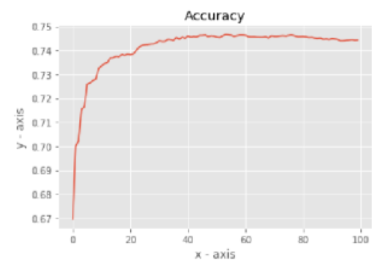
(**a**) Treatment 1



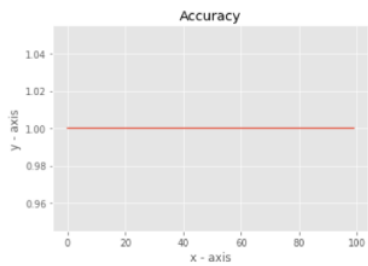
(**b**) Treatment 2



(**c**) Treatment 3



(**d**) Treatment 4



(**e**) Treatment 5

Figure 10. The highest *k* value using KNN for prediction.

Table 10. Comparison of the improvement in performance metrics before and after SMOTE-Tomek Links between the proposed method and other studies.

Method, Dataset, Class Ratio	Increase Precision (%)	Recall (%)	*F*1-Score (%)	Accuracy (%)
KNN, Credit Card Fraud, 0.17: 99.83 [20]	-	20.11	-	−0.85
KNN, Electrical Rotating Machine, 37.5: 62.5 [19]	43.00	41.00	45.00	40.00
KNN, Electrical Rotating Machine,16.7: 83.3 [19]	60.00	38.00	47.00	38.70
NB, DNA Methylation, [25]	5.00	24.00	12.00	-
LR, DNA Methylation, [25]	−14.00	7.00	−5.00	-
RF, DNA Methylation, [25]	−1.00	7.00	3.00	-
Proposed Method (NB, multiclass with Median-KNN regressor imputation)	0	2.67	1.33	5.16
Proposed Method (KNN, multiclass with Median-KNN regressor imputation)	25.33	37.83	34.17	20.2
Proposed Method (C4.5, multiclass with Median-KNN regressor imputation)	33.50	33.33	33.50	29.16

High accuracy does not always mean that the algorithm performs better in all situations. It is sometimes misleading in situations such as imbalanced class datasets, so it is not always considered to be accurate. In credit card fraud detection [20], if you only predict that the transaction is genuine and not fraudulent, it will cause big problems for credit card companies. As a result, the proportion of transaction cases that are predicted to be fraud but are not fraud must be higher than the proportion of transaction cases that are predicted not to be a fraud but are fraud. Thus, a high increase in a recall is needed compared to a high increase in accuracy.

In the case of DNA methylation classification [25], only NB had improved performance on all three metrics—precision, recall, and $F1$-Score—while the classification using random forest (RF) and logistic regression (LR) for data that are balanced with SMOTE-Tomek Links each have performance improvements in recall and $F1$-score or recall only. The highest increase was only achieved for recall using Naive Bayes. Overall, the increase achieved was less than 25% and the decline did not reach 15%.

A recall is a more significant evaluation measure than precision in most high-risk disease (such as cancer) detection situations. The recall represents the percentage of all cancer cases that the model correctly predicted, whereas the precision represents the percentage of predictions made by the cancer model where cancer is truly present. Similar to the detection of credit card fraud, DNA methylation also requires a more significant increase in recall compared to other evaluation measures.

The most significant performance increase (38%–60%) was obtained from monitoring an electrical rotating machine dataset [19] with class ratios of 37.5:62.5 and 16.7:83.3. In the first ratio, the metric that has the highest increase is precision, while in the second ratio, this metric is the $F1$-score. All of these studies predict cases using KNN. The increase in performance metrics, which reached 60% in [19], could also be due to the small number of samples. Further exploration is needed for large samples.

In our proposed method, the highest increase in prediction performance was achieved by C4.5, for which the metrics ranged from 29.16% to 33.5%. Generally, the resampling technique of SMOTE-Tomek Links has a positive effect on improving the prediction performance of the KNN and C4.5 methods, especially for India air quality prediction, where the amount of missing data is less than 10% or more than or equal to 10%, handled using the median and KNN regressor, respectively.

Furthermore, Table 11 present the performance metrics obtained by this proposed study using the SMOTE-Tomek Links technique and the three proposed methods are compared with previous research using this air quality dataset. Sethi and Mittal [32] obtained the lowest accuracy value, which only calculated accuracy and precision values. The lowest precision, recall, and $F1$-scores were obtained in this study using the Median-KNN-SMOTE-Tomek Links with NB method. This method of handling lost data and class imbalance that combines the Median-KNN and SMOTE-Tomek Link methods, then predicts quality with C4.5, has a very satisfying performance, where all performance metrics reach 100%. Sufficient experimentation is required to obtain a satisfactory predictive model performance.

Table 11. Comparison of the performance evaluation results in this study with previous studies using the Air Quality dataset.

Methods	Precision (%)	Recall (%)	$F1$-Score (%)	Accuracy (%)
Cloud Model Granulation [33]	71.43	-	-	-
KNN [34]	95.2	93.46	91.87	-
SVM [32]	74.3	-	-	70.66
Multilayer Perceptron [35]	85.4	-	-	85.79
Proposed Method (NB, multiclass with Median-KNN regressor imputation)	67.17	66.67	66.33	73.96
Proposed Method (KNN, multiclass with Median-KNN regressor imputation)	96.83	96.50	96.17	96.64
Proposed Method (C4.5, multiclass with Median-KNN regressor imputation)	100.00	100.00	100.00	100.00

4. Conclusions

This work predicts air quality using the India AQI dataset, which has many missing observations and imbalanced classes. Handling these two problems is important because they may give biased results and cause inaccurate predictions, respectively. Inaccurate predictions for the minority class can be fatal or cause big losses. The median and the KNN regressor are proposed to handle missing values of less than or equal to 10% and more than 10%, respectively. At the same time, the SMOTE-Tomek Links method addresses the class imbalance. These proposed approaches to handle both issues are then used to assess the air quality prediction of the India AQI dataset using NB, KNN, and C4.5. Five treatments are created to show the effect of Median-KNN regressor imputation and SMOTE-Tomek Links on the air quality prediction performance of the India AQI dataset. The five treatments are a combination of removing missing data and imputing missing data with SMOTE resampling and SMOTE-Tomek Links, respectively. The results show that the proposed method using the Median-KNN regressor and SMOTE-Tomek Links is able to improve the performance of the India air quality prediction model. In other words, the proposed method has succeeded in overcoming the problem of missing values and class imbalance. Even the predictions from the proposed model using C4.5 have values for the performance metrics of accuracy, precision, recall, and $F1$-score each of 100.

Author Contributions: Conceptualization, W.C. and Y.R.; methodology, W.C. and Y.R.; software, W.C. and Y.R.; validation, W.C., B.S. and Y.R.; formal analysis, W.C., B.S. and Y.R.; investigation, W.C. and Y.R.; resources, W.C. and Y.R.; data curation, W.C. and Y.R.; writing—original draft preparation, W.C. and Y.R.; writing—review and editing, W.C., B.S. and Y.R.; visualization, W.C. and Y.R; supervision, Y.R.; project administration, Y.R.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors thank to Rector of Universitas Sriwijaya and Rector of Universitas Bina Darma for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, F.; Du, J.; Lang, J.; Lu, W.; Liu, L.; Jin, C.; Kang, Q. Missing Value Estimation Methods Research for Arrhythmia Classification Using the Modified Kernel Difference-Weighted KNN Algorithms. *BioMed Res. Int.* **2020**, *2020*, 7141725. [CrossRef]
2. Cheng, C.-H.; Kao, Y.-F.; Lin, H.-P. A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. *Appl. Soft Comput.* **2021**, *108*, 107487. [CrossRef]
3. Rafsunjani, S.; Safa, R.S.; Imran, A.A.; Rahim, S.; Nandi, D. An Empirical Comparison of Missing Value Imputation Techniques on APS Failure Prediction. *Int. J. Inf. Technol. Comput. Sci.* **2019**, *11*, 21–29. [CrossRef]
4. Roy, K.; Ahmad, M.; Waqar, K.; Priyaah, K.; Nebhen, J.; Alshamrani, S.S.; Raza, M.A.; Ali, I. An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values. *Complexity* **2021**, *2021*, 9953314. [CrossRef]
5. Al Khaldy, M.; Kambhampati, C. Performance Analysis of Various Missing Value Imputation Methods on Heart Failure Dataset. *Lect. Notes Netw. Syst.* **2018**, *16*, 415–425. [CrossRef]
6. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
7. Ayilara, O.F.; Zhang, L.; Sajobi, T.T.; Sawatzky, R.; Bohm, E.; Lix, L.M. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health Qual. Life Outcomes* **2019**, *17*, 106. [CrossRef]
8. van Buuren, S. *Flexible Imputation of Missing Data*; Chapman & Hall/CRC Interdisciplinary Statistics; CRC Press: Boca Raton, FL, USA, 2012.
9. Sim, J.; Lee, J.S.; Kwon, O. Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications. *Math. Probl. Eng.* **2015**, *2015*, 538613. [CrossRef]
10. Xia, J.; Zhang, S.; Cai, G.; Li, L.; Pan, Q.; Yan, J.; Ning, G. A Adjusted weight voting algorithm for random forests in handling missing values. *Pattern Recognit.* **2017**, *69*, 52–60. [CrossRef]
11. Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; Tabona, O. A Survey on Missing Data in Machine Learning. *J. Big Data* **2021**, *8*, 140. [CrossRef]
12. Salgado, C.M.; Azevedo, C.; Proença, H.; Vieira, M.S. *Missing Data. Secondary Analysis of Electronic Health Records*; Spriner: Berlin/Heidelberg, Germany, 2016.
13. Razavi-Far, R.; Farajzadeh-Zanajni, M.; Wang, B.; Saif, M.; Chakrabarti, S. Imputation-Based Ensemble Techniques for Class Imbalance Learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 1988–2001. [CrossRef]
14. Huang, J.; Keung, J.W.; Sarro, F.; Li, Y.F.; Yu, Y.T.; Chan, W.K.; Sun, H. Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study. *J. Syst. Softw.* **2017**, *132*, 226–252. [CrossRef]
15. Zhang, S.; Cheng, D.; Deng, Z.; Zong, M.; Deng, X. A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognit. Lett.* **2018**, *109*, 44–54. [CrossRef]

16. Manimekalai, K.; Kavitha, A. Missing Value Imputation and Normalization Techniques in Myocardical Infarction. *ICTACT J. SOFT Comput.* **2018**, *8*, 1655–1662. [CrossRef]
17. Upadhyay, K.; Kaur, P. A Review on Data level Approaches to address the Class Imbalance Problem. In Proceedings of the International Conference on Challenges in Engineering Science and Technology, Babylon, Iraq, 6–8 April 2021.
18. Kaur, P.; Gosain, A. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. *Adv. Intell. Syst. Comput.* **2018**, *653*, 23–30. [CrossRef]
19. Swana, E.F.; Doorsamy, W.; Bokoro, P. Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors* **2022**, *22*, 3246. [CrossRef]
20. Lin, T.H.; Jiang, J.R. Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics* **2021**, *9*, 2683. [CrossRef]
21. Imran, M.; Hina, S.; Baig, M.M. Analysis of Learner's Sentiments to Evaluate Sustainability of Online Education System during COVID-19 Pandemic. *Sustainability* **2022**, *14*, 4529. [CrossRef]
22. Walsh, R.; Tardy, M. A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer. *Diagnostics* **2023**, *13*, 67. [CrossRef]
23. Ai-Jun, L.; Peng, Z. Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote. *ACM Int. Conf. Proc. Ser.* **2020**, 13–17. [CrossRef]
24. Zeng, M.; Zou, B.; Wei, F.; Liu, X.; Wang, L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In Proceedings of the 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, China, 28–29 May 2016.
25. Liu, C.; Wu, J.; Mirador, L.; Song, Y.; Hou, W. Classifying DNA methylation imbalance data in cancer risk prediction using SMOTE and Tomek link methods. In *Data Science*; Springer: Singapore, 2018. [CrossRef]
26. Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change, Government of India. National Air Quality Index. Available online: https://cpcb.nic.in/National-Air-Quality-Index/ (accessed on 12 September 2022).
27. Khazaee Poul, A.; Shourian, M.; Ebrahimi, H. A Comparative Study of MLR, KNN, ANN and ANFIS Models with Wavelet Transform in Monthly Stream Flow Prediction. *Water Resour. Manag.* **2019**, *33*, 2907–2923. [CrossRef]
28. Mahboob, T.; Ijaz, A.; Shahzad, A.; Kalsoom, M. Handling Missing Values in Chronic Kidney Disease Datasets Using KNN, K-Means and K-Medoids Algorithms. *Syst. Technol. Proc.* **2019**, 76–81. [CrossRef]
29. Skryjomski, P.; Krawczyk, B. Influence of Minority Class Instance Types on SMOTE Imbalanced Data Oversampling. *Proc. Mach. Learn. Res.* **2017**, *74*, 7–21.
30. Fernández, A.; García, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
31. Alzoman, R.M.; Alenazi, M.J.F. A comparative study of traffic classification techniques for smart city networks. *Sensors* **2021**, *21*, 4677. [CrossRef]
32. Sethi, J.K.; Mittal, M. Ambient Air Quality Estimation Using Supervised Learning Techniques. *EAI Endorsed Trans. Scalable Inf. Syst.* **2019**, *6*, e8. [CrossRef]
33. Lin, Y.; Zhao, L.; Li, H.; Sun, Y. Air Quality Forecasting Based on Cloud Model Granulation. *Eurasip J. Wirel. Commun. Netw.* **2018**, *2018*, 106. [CrossRef]
34. Haq, M.A. Smotednn: A novel model for air pollution forecasting and aqi classification. *Comput. Mater. Contin.* **2022**, *71*, 1403–1425. [CrossRef]
35. Chowdhury, A.S.; Uddin, M.S.; Tanjim, M.R.; Noor, F.; Rahman, R.M. Application of Data Mining Techniques on Air Pollution of Dhaka City. In Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS), Varna, Bulgaria, 28–30 August 2020. [CrossRef]

Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction

ORIGINALITY REPORT

11%	7%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

2%

★ mdpi-res.com

Internet Source

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		