Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 6, No.1, January 2024, pp. 02401025-01 ~ 02401025-08
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v6i1.17951

XGBoost and Random Forest Optimization using SMOTE to Classify Air Quality

Fidela Putri Arifianti*, Abu Salam

Faculty of Computer Science, Universitas Dian Nuswantoro, Jl. Imam Bonjol No.207 Semarang 50131, Central Java, Indonesia

*111202012533@mhs.dinus.ac.id

Abstract. Air pollution due to the growth of industry and motorized vehicles seriously threatens human health. Clean air is essential, but pollutant contamination can cause acute respiratory illnesses and other illnesses. Several studies have been carried out to anticipate this air pollution. Various algorithms, methods, and data balancing techniques have been implemented, but still need to be done to obtain better accuracy results. Therefore, this study aims to classify heart disease using the XGBoost and Random Forest algorithms and implement the SMOTE technique to overcome data imbalance. This research produces a Random Forest algorithm with SMOTE implementation with splitting 80:20, which has the best accuracy with an accuracy of 92.4%, an average AUC of 0.98, and a log loss of 0.2366, which shows that SMOTE has succeeded in improving model performance in classifying minority classes. Based on the results obtained, the XGBoost and Random Forest algorithms after SMOTE are superior to the model with SMOTE, with accuracy above 90%.

Keywords: Air Quality, Classification, XGBoost, Random Forest, SMOTE

(Received 2024-01-11, Accepted 2024-01-28, Available Online by 2024-01-29)

1. Introduction

Human health is closely related to air quality, which can be affected by economic growth and urban development, increasing air pollution [1]. Air pollution, especially in Indonesian cities due to the growth of industry and motorized vehicles, seriously threatens human health [2]. Clean air is very important, but contamination by pollutants can cause acute respiratory and other illnesses. Good air quality is a key factor, especially in open areas and sectors with direct interaction [3].

The field of artificial intelligence (AI) continues to develop and has a key role in digital transformation, especially in classification using data mining methods [4]. Machine learning through models such as XGBoost and Random Forest can assist with information extraction and simulation during the early design stages[5]. The challenge of class imbalance in machine learning can be overcome by using the Synthetic Minority Oversampling Technique (SMOTE), which utilizes original data to create additional minority data with different characteristics, aiming to reduce the risk of overfitting [6].

Research conducted by Cosmas Haryawan and Yosef Muria Kusuma Ardhana[7] aims to compare the results of synthetic data measurements produced by SMOTE and K-Means SMOTE with the results

of real data measurements in balanced conditions. The results in the research data resulted in a balanced condition, obtaining accuracy results of 76.85%, sensitivity of 82.05%, and specificity of 71.13%. SMOTE in its application tends to increase accuracy in classification tasks. The use of oversampling techniques such as SMOTE can significantly increase accuracy compared to accuracy without using SMOTE. K-Means SMOTE has been used in various studies and has shown improved accuracy for classification tasks. Overall, the use of SMOTE and K-Means SMOTE can lead to higher accuracy in imbalanced data scenarios.

Research conducted by Adli A. Nababan, Miftahul Jannah, Mia Aulina, and Dwiki Andrian[8] aims to determine the factors that influence air quality using the XGBoost algorithm and the synthetic minority oversampling method (SMOTE) based on the Air Pollution Standard Index (ISPU) category. The proposed SMOTE method is 48%. The XGBoost algorithm is an extension of the Gradient Boosting method, performing better in predicting air quality.

Based on the background that has been presented, air quality problems, especially air pollution in Indonesian cities, are the main focus in efforts to maintain human health. The increasing growth of industry and motorized vehicles has caused a decline in air quality, which poses a serious threat to health. In order to overcome the class imbalance in air quality datasets, this research proposes the use of oversampling techniques, such as SMOTE, and machine learning algorithms, such as XGBoost and Random Forest to develop accurate classification models. Several previous studies have also tested and compared various methods, such as the K-Nearest Neighbors (KNN) algorithm and the SMOTE oversampling technique, with results showing increased accuracy and effectiveness in classifying air pollution. Thus, it is hoped that the results of this research can contribute to the understanding and handling of air quality problems, as well as provide solutions that can help society and the government in facing the threat of air pollution..

2. Methods

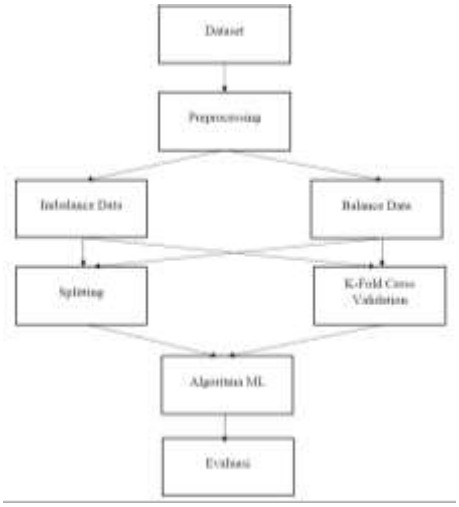

Figure 1. Research Design Flow

This research involves several methodological stages, starting with data collection from open data sites. The data then undergoes a preprocessing stage, including handling missing values and normalization. The next process involves balancing the data with oversampling using the SMOTE method. The next stage includes data sharing through data splitting and applying k-fold cross-validation. The next step involves classification using the XGBoost and Random Forest algorithm models. In the final stage, the research is evaluated to ensure the achievement of the initial objectives.

2.1. Dataset Source

The dataset used in this research comes from the open data site Kaggle.com in the form of Air Quality data in South Tangerang, Indonesia 20-22. The data used is in CSV (comma-separated value) form. The data used is imbalance data containing 1096 data with details of 393 data in the healthy air class, 639 data in the moderate air class, and 64 data in the unhealthy air class. In the dataset used, there are also missing values in the features.Preprocessing Data

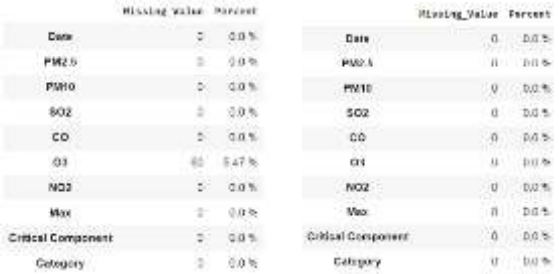

Figure 2. Handling Missing Value

Data Preprocessing is a method applied to obtain information from unprocessed data. This technique aims to eliminate noise that may be present in raw data, making it easier to process data in subsequent processes[9]. This research's preprocessing stage involves handling missing values and normalizing the data. Missing value handling is carried out to ensure data is not lost due to deletion, which can eliminate information in the data. Missing value handling in preprocessing needs to be done because there are 60 missing data in the O3 feature. Missing values in the features are filled with the average value of the O3 feature. This normalization process aims to reduce data redundancy and improve data integrity. In the context of this research, MinMax Scaler is used as a normalization method. The choice of MinMaxScaler was considered because this method effectively adjusts the data scale to contain values in the range 0 to 1.

2.2. Oversampling SMOTE

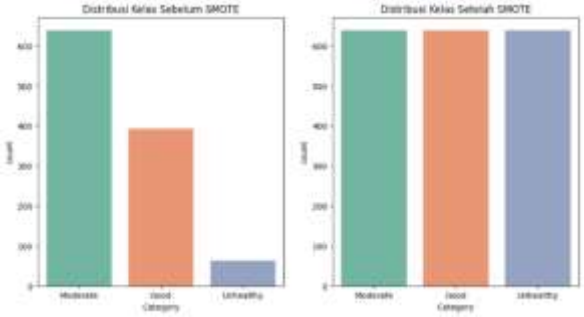

Figure 3. Data before and after SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) method is an oversampling technique used to overcome the imbalance in the number of datasets between minority and majority classes[10]. SMOTE is used to avoid degradation of classification performance[15]. SMOTE achieves balance by synthesizing datasets in the minority class until the number is equal to those in the majority class. Although Oversampling techniques can cause overfitting, using SMOTE is specifically designed to overcome this problem.

2.3. Data Sharing

Data sharing is the stage of dividing data into data for training and data for testing[11]. The division of data into training and testing is important in classification algorithms. The training data forms a knowledge model that is applied to predict new data classes. Data testing measures the accuracy of the classifier. In this research, the dataset is divided with a ratio of 80:20 (80% training, 20% testing) and 70:30 (70% training, 30% testing). K-fold cross-validation is also used, dividing the data into k parts for model evaluation without overlap. This research applies 5-fold and 10-fold cross-validation.

2.4. XGBoost and Random Forest Classification

XGBoost is used for regression and classification. With the loss function as the evaluation criterion, the algorithm builds a weak learner at each iteration to improve prediction accuracy. Additional features in XGBoost help prevent overfitting and increase computing speed[12]. Random Forest is efficient for classification on large datasets. Using several decision trees and a selection process[13], a random forest divides the dataset based on features, and the prediction results from the decision trees are combined. This approach provides good performance in classifying large data.

2.5. Evaluation

Classification assessment uses evaluation to test the truth and error of objects. In this research, the evaluation utilizes a confusion matrix, providing information about actual and predicted results by the classification system. The components of the confusion matrix provide insight into model performance. From the confusion matrix, model performance is measured by the Accuracy, Precision, Recall, F1-Score, and ROC AUC metrics, providing a complete picture of the model's prediction accuracy[14].

3. Results and Discussion

The research focuses on air quality classification analysis using the XGBoost and Random Forest algorithms with SMOTE optimization. The goal is to compare the two, integrate XGBoost as a comparison with Random Forest, and perform SMOTE optimization for optimal accuracy. Data was collected from Kaggle.com, and then preprocessing was carried out, including handling missing values and normalization. Continues with balancing the dataset using SMOTE, splitting dataset, and k-fold cross-validation. The classification process involves XGBoost and Random Forest models, ending with a thorough evaluation using confusion matrices and metrics such as accuracy, log loss, and ROC AUC.

3.1. XGBoost

The XGBoost algorithm shows excellent performance in this research when applied with data splitting and K-Fold Cross-Validation techniques, especially with the utilization of the SMOTE method, which has been proven to improve the performance of classification models. Achieving the best accuracy ratio of up to 92%, the XGBoost algorithm with SMOTE consistently shows commendable performance in air quality data classification.

Table 1. XGBoost Experiment Results with Splitting Data

Num	Splitting	Data	Model	Accuracy	Avg AUC	Log loss
1	80:20	Imbalance	XGBoost	83.2	0.96	0.3996
2	80:20	Balance	XGBoost	91.7	0.98	0.2382
3	70:30	Imbalance	XGBoost	82.1	0.96	0.4311
4	70:30	Balance	XGBoost	90.4	0.98	0.2614

The improvement of experiments with SMOTE can be seen from the comparison of results on balanced and unbalanced datasets. The first experiment (No. 1) at a ratio of 80:20 shows that XGBoost achieves 83.2% accuracy, an average AUC of 0.96, and a log loss of 0.3996. However, class imbalance affects model performance. The second experiment (No. 2) with SMOTE at the same ratio yielded significant improvements: 91.7% accuracy, average AUC 0.98, and logloss 0.2382. The third experiment (No. 3) at a 70:30 ratio without SMOTE gave an accuracy of 82.1%, average AUC 0.96, and logloss 0.4311. The fourth experiment (No. 4) with SMOTE at the same ratio shows improvements: accuracy 90.4%, average AUC 0.98, and logloss 0.2614. Positive SMOTE addresses class imbalance and consistently improves model evaluation, especially for minority classes, improving XGBoost model performance significantly.

Table 2. XGBoost Experiment Results with K-Fold Cross Validation

Num	KFold	Data	Model	Accuracy	Avg AUC	Logloss
1	5	Imbalance	XGBoost	87.3	0.96	0.35

2	5	Balance	XGBoost	91.7	0.98	0.25
3	10	Imbalance	XGBoost	88.0	0.97	0.35
4	10	Balance	XGBoost	92.0	0.99	0.22

Table 2 shows the positive impact of SMOTE on the results of the XGBoost experiment with K-Fold Cross-Validation. The initial experiment (No. 1) showed an accuracy ratio of 87.3%, an average AUC of 0.96, and a log loss of 0.35, but still faced challenges classifying minority classes. The second experiment (No. 2), after implementing SMOTE, achieved an average accuracy of 91.7%, average AUC of 0.98, and log loss of 0.25, showing significant improvements in dealing with class imbalance. The third experiment (No. 3) with K-Fold Cross Validation 10 gave an average accuracy of 88.0%, an average AUC of 0.97, and a log loss of 0.35. In the fourth experiment (No. 4) with SMOTE, the average accuracy reached 92.0%, Average AUC 0.99, and log loss 0.22, showing further improvement. The application of SMOTE to K-Fold Cross Validation provides significant improvements in the performance of the XGBoost model, especially in dealing with class imbalance and increasing the ability to classify minority classes.

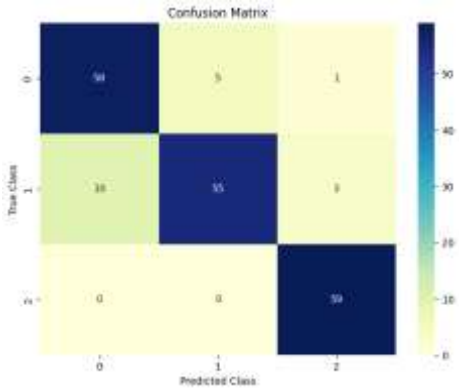

Figure 4. Confusion Matrix Best XGBoost Technique

Evaluating air quality predictions from the XGBoost model with the confusion matrix shows that overall, the model can differentiate between healthy (class 0) and unhealthy (class 2) air quality, with an accuracy of 58 and 59 correct predictions, respectively. However, several weaknesses can be identified. This model has difficulty in separating the moderate air quality class (class 1) from the other classes, as can be seen from the number of prediction errors; there are 10 cases where unhealthy air quality was incorrectly predicted as moderate quality (FP), and 3 cases Air quality is being predicted as healthy quality (FP)

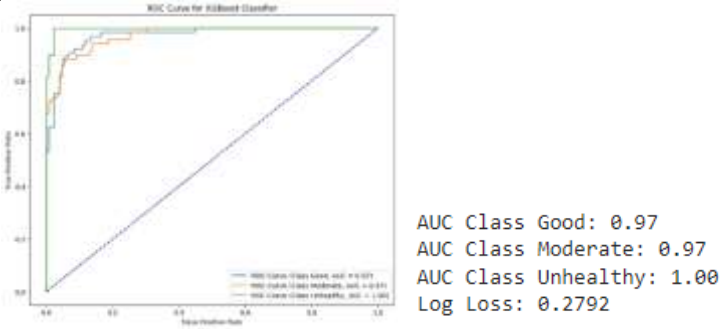

Figure 5. ROC Curve of the Best XGBoost Technique

The results of the Receiver Operating Characteristic (ROC) curve show that the AUC for Class 0 (Healthy Air Quality) is 0.97, the AUC for Class 1 (Medium Air Quality) is 0.97, the AUC for Class 2 (Unhealthy Air Quality) is 1.00, and the log loss value is 0.2792. By looking at the high AUC results for each class, with class 2 AUC reaching a value of 1.00, it can be concluded that the model has very good abilities in separating and classifying data for unhealthy air quality very well. The low log loss value (0.2792) indicates that the model provides predictions with a minimum level of uncertainty, so it is reliable.

3.2. Random Forest

The Random Forest algorithm shows excellent performance in this research when applied with data splitting and K-Fold Cross-Validation techniques, especially with the utilization of the SMOTE method, which has been proven to improve the performance of classification models. Achieving the best accuracy ratio of up to 92.4%, the Random Forest algorithm with SMOTE consistently shows commendable performance in air quality data classification.

Table 3. Results of Random Forest Experiments with Splitting Data

Num	Splitting	Data	Model	Accuracy	Avg AUC	Logloss
1	80:20	Imbalance	Random Forest	85.9	0.96	0.3494
2	80:20	Balance	Random Forest	92.4	0.98	0.2366
3	70:30	Imbalance	Random Forest	83.3	0.94	0.3867
4	70:30	Balance	Random Forest	90.8	0.98	0.2469

Table 3 shows the results of the Random Forest experiment using the data splitting technique at a ratio of 80:20 and 70:30. Initial experiments on unbalanced data (Num 1) yielded 85.9% accuracy, average AUC 0.96, and log loss 0.3494. In the experiment with the application of SMOTE (Num 2), there was a significant increase with an accuracy of 92.4%, average AUC of 0.98, and log loss of 0.2366, indicating that SMOTE succeeded in improving the model's performance in classifying minority classes. Experiments without SMOTE at a ratio of 70:30 (Num 3) gave an accuracy of 83.3%, average AUC 0.94, and logloss 0.3867. Applying SMOTE in the fourth experiment (Num 4) increased accuracy to 90.8% while maintaining a high average AUC of 0.98, and log loss decreased to 0.2469. SMOTE has a positive impact in overcoming class imbalance, as evidenced by better evaluation results on all measured metrics.

Table 4. Random Forest Experiment Results with K-Fold Cross Validation

Num	KFold	Data	Model	Avg Accuracy	Avg AUC	Avg Logloss
1	5	Imbalance	Random Forest	87.7	0.97	0.35
2	5	Balance	Random Forest	91.4	0.98	0.26
3	10	Imbalance	Random Forest	86.4	0.97	0.33
4	10	Balance	Random Forest	91.8	0.98	0.24

The results of the Random Forest experiment with K-Fold Cross-Validation show the positive impact of SMOTE. Initial experiments (Num 1) on unbalanced data and K-Fold Cross Validation 5 resulted in an average accuracy of 87.1%, average AUC of 0.97, and log loss of 0.31. In experiments using SMOTE (Num 2), there was a significant increase, with an average accuracy of 92.2%, average AUC of 0.98, and log loss of 0.23. SMOTE helps models more effectively address class imbalance, providing consistent improvements in performance evaluation. Experiments without SMOTE on K-Fold Cross Validation 10 (Num 3) gave an average accuracy of 88.6%, an average AUC of 0.97, and a log loss of 0.30. Applying SMOTE in the fourth experiment (Num 4) increased the average accuracy to 93.2%, average AUC to 0.99, and log loss to 0.20. The application of SMOTE in K-Fold Cross Validation can significantly improve the performance evaluation of the Random Forest model, especially in dealing with class imbalance and increasing the ability to classify minority classes.

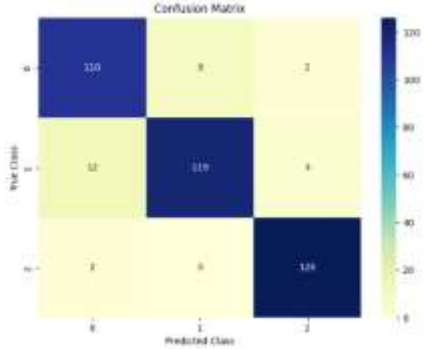

Figure 6. Confusion Matrix Best Random Forest Technique

The confusion matrix results show that the model correctly identified 110 samples as class 0, but there were nine errors where non-class 0 samples were incorrectly identified as class 0. The model also succeeded in identifying 90 samples as not class 0. Class 1 contained 86 True Positives, 14 False Positives, and 85 True Negatives.

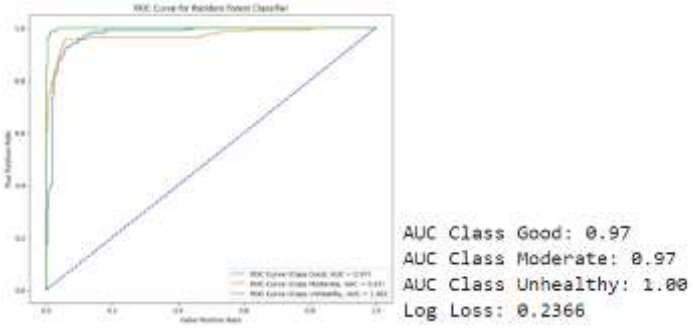

Figure 7. ROC Curve Best Random Forest Technique

Evaluation of the Random Forest model in this study showed very positive results through the Area Under the Receiver Operating Characteristic Curve (AUC) values for each air quality class. With an AUC of 0.97 for the "Good" class, 0.97 for the "Moderate" class, and 1.00 for the "Unhealthy" class, the model can effectively differentiate between different levels of air quality. In addition, the low Log Loss value of 0.2366 indicates that the model provides predictions with a high level of certainty.

4. Conclusion

This research optimizes air quality classification with XGBoost and Random Forest using SMOTE. The results show significant improvements in accuracy, AUC, and reduction in gloss, indicating the effectiveness of minority class oversampling. Even though there were initial weaknesses, implementing SMOTE consistently provided good improvements. Evaluation using the confusion matrix and ROC curve shows XGBoost's good ability in classifying air quality. SMOTE optimization can be relied on to improve model predictions, especially in the face of class imbalance. In the XGBoost experiment, applying SMOTE to K-Fold Cross Validation 10 improved accuracy to 92.0%, AUC 0.99, and log loss decreased to 0.22. Random Forest experiments in this study, both from splitting data and k-fold cross-validation, showed accuracy above 83%. Applying SMOTE at an 80:20 ratio initially provided an accuracy of 85.9%, an average AUC of 0.96, and a loss of 0.3494. Despite the class imbalance, SMOTE improved accuracy to 92.4%, AUC to 0.98, and log loss to 0.2366. Research shows the effectiveness of SMOTE in addressing class imbalance and improving model performance, although there is still room for improvement in further model development. Future research requires exploring the use of other algorithms, evaluating the use of oversampling methods with other techniques, and considering the use of preprocessing techniques to improve the quality of data that will be used in air quality classification. It is hoped that these suggestions can become useful input in developing air quality classification in future.

References

[1] M. Méndez, M. G. Merayo, dan M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artif Intell Rev*, vol. 56, no. 9, hlm. 10031–10066, Sep 2023, doi: 10.1007/s10462-023-10424-4.

[2] B. V. Jayadi, T. Handhayani, dan M. D. Lauro, "Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta," *Jurnal Ilmu Komputer dan Sistem Informasi*, hlm. 1–7, 2023.

[3] S. S. A. Umri, M. S. Firdaus, dan Primajaya A, "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta," *JIKO (Jurnal Informatika dan Komputer)*, vol. 4, no. 2, hlm. 98–104, 2021.

[4] Y. Devianto dan S. Dwiasnati, "Kerangka Kerja Sistem Kecerdasan Buatan dalam Meningkatkan Kompetensi Sumber Daya Manusia Indonesia," *Jurnal Telekomunikasi dan Komputer*, vol. 10, no. 1, hlm. 19, Apr 2020, doi: 10.22441/incomtech.v10i1.7460.

[5] S. B. Nadkarni, G. S. Vijay, dan R. C. Kamath, "Comparative Study of Random Forest and Gradient Boosting Algorithms to Predict Airfoil Self-Noise," dalam *RAiSE-2023*, Basel Switzerland: MDPI, Des 2023, hlm. 24. doi: 10.3390/engproc2023059024.

[6] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, dan F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, hlm. 677–690, Jul 2022, doi: 10.30812/matrik.v21i3.1726.

[7] C. Haryawan dan Y. M. K. Ardhana, "Analisa Perbandingan Teknik Oversampling Smote Pada Imbalanced Data," *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 6, no. 1, hlm. 73–78, Apr 2023.

[8] A. A. Nababan, M. Jannah, M. Aulina, dan D. Andrian, "Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (SMOTE) Berdasarkan Indeks Standar Pencemaran Udara (ISPU)," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 7, no. 1, hlm. 214–219, 2023.

[9] M. Fahmi dan I. Suhartana, "Perbandingan Algoritma Decision Tree Dan Support Vector Machine Dalam Prediksi Kualitas Udara," Jurnal Nasional Teknologi Informasi dan Aplikasinya, vol. 1, no. 1, hlm. 21–30, Nov 2022.

[10] M. Mustaqim, B. Warsito, dan B. Surarso, "Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan," Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 5, no. 2, hlm. 128, Jul 2019, doi: 10.26594/register.v5i2.1705.

[11] A. A. H. Kirono, I. Asror, dan Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma Naive Bayes," e-Proceeding of Engineering, vol. 9, no. 3, hlm. 1962–1969, Jun 2022.

[12] G. A. Mursianto, I. M. Falih, M. Irfan, T. Sakinah, dan D. S. Prasvita, "Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan," dalam *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, Jakarta, Sep 2021, hlm. 41–50.

[13] N. B. Putri dan A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing," *Komputika : Jurnal Sistem Komputer*, vol. 11, no. 1, hlm. 59–66, Jan 2022, doi: 10.34010/komputika.v11i1.4350.

[14] A. Nugroho, I. Asror, dan Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest," *e-Proceeding of Engineering*, vol. 10, no. 2, hlm. 1824–1834, Apr 2023.

[15] A. N. Cahyani, J. Zeniarja, S. Winarno, R. T. E. Putri, and A. A. Maulani, "Heart Disease Classification Using Deep Neural Network with SMOTE Technique for Balancing Data," Advance Sustainable Science, Engineering and Technology, vol. 6, no. 1, p. 0240108, Dec. 2023, doi: 10.26877/asset.v6i1.17521.