E-ISSN : 2580-2879
JUSIKOM
Jurnal Sistem Informasi Komputer

Medan, 06 September 2024

No. : 416/LOA/JSKM/UNPRI/VIII/2024
About : **Letter of Acceptance** (LOA)

Kepada Yth. Bapak/I/Sdr

Muhammad Jodi Noviandri
Ferdiansyah
Andri
Novri Hadinata

Di
Tempat

Dengan ini kami sampaikan bahwa artikel dengan rincian berikut dinyatakan diterima untuk di terbitkan dalam **JURNAL SISTEM INFORMASI DAN ILMU KOMPUTER PRIMA (JUSIKOM)** Universitas Prima Indonesia Medan Pada **Vol 8, No 2**, Untuk Publish **Februari, 2025**.

Penulis Correspondent Email	Judul
Muhammad Jodi Noviandri E-Mail: Tegas.Hadiyanto82@gmail.com	**Prediction Of Air Quality In South Sumatra Based On Air Pollutant Standard Index Using Extreme Gradient Boosting Algorithm**

Demikian surat keterangan ini kami buat untuk dapat digunakan seperlunya

Hormat Kami,

Evta **Indra S.Kom, M.Kom**
Editor In Chief

Prediction Of Air Quality In South Sumatra Based On Air Pollutant Standard Index Using Extreme Gradient Boosting Algorithm

Muhammad Jodi Noviandri[1], Ferdiansyah[2], Andri[3], Novri Hadinata[4]
[1,2,3,4]Information System, Faculty Science and Technology, Bina Darma University, Palembang, Indonesia
Email: muhammadjodinoviandri30@gmail.com, ferdi@binadarma.ac.id, andri@binadarma.ac.id, novri_hadinata@binadarma.ac.id

ABSTRACT

In Indonesia, especially in the province of South Sumatra, forest and land fires have become an annual "tradition" but with different levels of pollution each year. The data obtained is a timeseries for four years. The collected data will be applied to the XGBoost algorithm for air quality prediction. The existing data is divided into training data and test data through data composition trials and then the existing data is divided into three sets of data division, namely the division of training data by 70%, testing data by 30%, training data by 80%, testing data by 20% and training data by 90%, testing data by 10%. The accuracy result for the proportion of 70:30 data is 98%, the recall value is 94%, the F-1 Score value is 95% and the average AUC value is 0.92. The accuracy result for the proportion of 80:20 data is 98%, the recall value is 99%, the F-1 Score value is 99% and the average AUC value is 0.91. The accuracy result for the 90:10 data proportion is 99%, the recall value is 99%, the F-1 Score value is 99% and the average AUC value is 0.91. The performance results in each data proportion do not occur overfitting and produce goodfitting. Feature Importance in this dataset is the PM2.5 parameter which gets the highest value among other ISPU parameters.

Keywords: Prediction, Air Quality, Air Pollutant Standard Index, Extreme Gradient Boosting

INTRODUCTION

At the time of October 11, 2023 based on data from the IQAir website accessed at 10.00 PM that Palembang City, South Sumatra Province, occupies the first position of the 10 cities in Indonesia with the worst air pollution levels with an AQI value of US or average unit 259 based on ISPU parameters (K. Hasan, 2024). Based on the South Sumatra Health Office, on October 1, 2023, the ISPU value was 345 with the PM2 parameter. 5 which means that the air quality is dangerous.[2] This happened because of the haze of forest and land burning that occurred in the South Sumatra area, especially in the Ogan Komering Ilir (OKI) Regency area which has recorded 116 Ha of burned land, this figure is the highest number in South Sumatra(Herda Sabriyah et al., 2023). Palembang City became the city with the worst air quality with the first rank of eight other cities in Indonesia. The main pollutant that causes a decrease in air quality is PM2.5 particulates, where the amount of this pollutant should not exceed a value of 10 microns when the particulates are in the air(Kusnandar, M., 2020).

Extreme Gradient Boosting is a developmental algorithm of gradient tree boosting based on ensemble algorithm, which can effectively handle large-scale machine learning cases(Herni Yulianti et al., nd, 2022). The XGBoost method was chosen because it has several additional features that are useful for speeding up the calculation system and preventing overfitting. XGBoost can solve various instances of classification, regression, and ranking. XGBoost is a tree collection calculation consisting of an assortment of previous trees (CART)(Shafila, G.A, 2020).

The application of the algorithm was also carried out in a study entitled "Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam" by Gokulan Ravindiran et al. In this study the researchers compared boosting algorithms including LightGBM, CatBoost, AdaBoost and XGBoost and Random Forest on air quality data in Visakhapatnam City. This study shows the results that the CatBoost model gets better results than other algorithms with a correlation coefficient R2 value of 0.9998, for Mean Absolute Error (MAE) with a value of 0.60, for Mean Square Error value of 0.58 and Root Mean Square Error (RMSE) value of 0.76(Ravindiran, G., et al, 2023).

Research conducted by Ishan Ayus, et al. entitled "Comparison of Machine Learning and Deep Learning Techniques for The Prediction of Air Pollution: A Case Study From China". The research explains that the XGBoost machine learning model is the most efficient deep learning model compared to other machine learning, namely Recurrent Neural Network (RNN), Bidirectional Gated Recurrent Unit (Bi-GRU), Bidirectional Long Short Term Memory (BiLSTM), Convolutional Neural Network BiLSTM (CNN-BiLSTM) and Convolutional BiLSTM (Conv1D-BiLSTM) at 10 air quality monitoring stations in China(Ayus. I, et al, 2023).

In 2020 the Ministry of Environment and Forestry passed the Minister of Environment and Forestry Regulation number 14 of 2020 concerning the Air Pollutant Standard Index as a replacement for the Minister of Environment Decree No.45 of 1997 concerning Calculation and Reporting and Information on the Air Pollutant Standard Index. In Permen LHK Number 14 of 2020 Article 2 Paragraph 2 states that the ISPU parameters include PM10 particulates, PM2.5 particulates, carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3) and hydrocarbons (HC)(Kusnandar. M, 2020).

METHODS

This research is a type of descriptive research that describes facts and information systematically based on historical data (Syafrida, Hafni, Sahir. 2022) by applying the eXtreme Gradient Boosting algorithm. The research stages carried out are as follows:

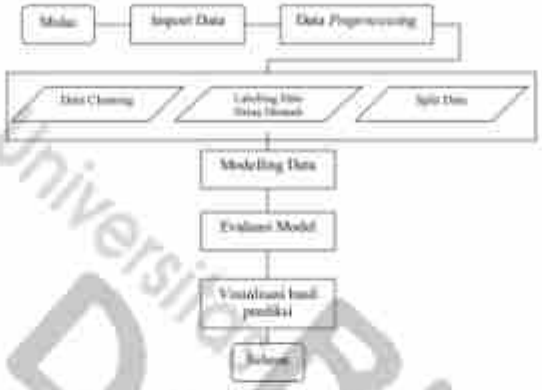


Figure 1. Research Flowchart of XGBoost Application on ISPU Palembang City

Data Collection

The dataset used in this study is air quality data in Palembang City with a time span of the last 5 years starting from 2019 to 2024. This data is secondary data obtained from the Ministry of Environment and Forestry. This data contains the value and category of the Air Pollutant Standard Index which consists of: time, PM10, PM2.5, SO2, CO, O3, NO2, HC, Critical Component, and Category. Import Data is the initial stage of importing datasets into data processing software. For example, Google Colab or Jupyter Notebook.

Data Pre-Processing

In the preprocessing stage, the data goes through the following stages:

1. Data Cleaning: Filling in missing data, removing unused table columns and removing unexplained outliers.
2. Data Transformation: Correcting the data type in each column and converting categorical data to binary data.
3. Labeling string data into integer, This process changes the column that has string data type into integer data type, because artificial intelligence can only read the contents of integer type data consisting of numbers "0" and "1".

Splitting Data

The dataset will be divided into two sets of testing data and training data with three proportions, the first proportion with 70% training data and 30% testing data, the second proportion with 80% training data and 20% testing data, and the third proportion with 90% training data and 10% testing data. This is intended as a comparison between data proportions.

Implementation Model of Extreme Gradient Boosting

Model training is the third stage to enter the method as a prediction pattern from the dataset used. In this study using the XGBoost Classifier method. XGBoost is one of the advanced Gradient Tree Boosting-based methods that can work efficiently in handling large-scale problems with very limited computing resources(Chen. T, et al, 2016). XGBoost is basically a Decision Tree algorithm which is known as Classificattion and Regression Tree (CART)(Shafila. G. A, 202)

Model Evaluation

At the evaluation stage, the extreme gradient boosting algorithm model is tested from the results of making the training model using test data. To find out the performance of the model that has been made, it will then be evaluated using the classification report, confusion matrix, xgboost tree and xgboost logloss.

1. Confusion Matrix

Confusion Matrix basically provides information on the comparison of the classification results performed by the model with the actual classification results. Confusion Matrix is a matrix table that describes the performance of the classification model on a set of test data whose actual values are known.(Kristiawan, et al. 2021)

Tabel 1. Confusion Matrix

	Predicted (0)	Predicted (1)
Actual (0)	True Negative (TN)	False Positive (FP)
Actual (1)	False Negative (FN)	True Positive (TP)

2. Learning Curves

A learning curve is a plot that shows time or experience on the x-axis and learning or improvement on the y-axis. learning curves are used to diagnose overfitting behavior of a model that can be addressed by tuning the hyperparameters of the model(Jason Brownlee,

2021). there are three common dynamic observe in learning curves, follow as: Underfir, Overfit and Goodfit.

3. ROC and AUC Curve

An evaluation metric used to measure the performance of classification models, this metric focuses on the model's ability to distinguish between positive and negative classes with respect to the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC curve is made based on the value that has been obtained from the calculation with the confusion matrix, which is between the False Positive Rate and True Positive Rate (Danang Dwi Nugroho, 2020). AUC is used to calculate the area under the ROC curve. The AUC value can be calculated by adding the trapezoidal area of the AUC measure with the AUC value between 0 and 1. If the AUC value is closer to 1, the better the model will be in classifying the data(Dian Tri Wilujeng, et al. 2023). AUC values can be divided into five categories as follows:

Table 2. Category and values of AUC

AUC Values	AUC Category
0.90 - 1.00	Excellent Classification
0.80 - 0.90	Good Classification
0.70 - 0.80	Fair Classification
0.60 - 0.70	Poor Classification
0.50 - 0.60	Failure

Visualization of Prediction Results

Visualization of prediction results, at this stage will display the prediction results in the form of time series forecasting for one of the most influential ISPU parameters. parameter based on feature importance.

RESULT

Classification Report and Learning Curves

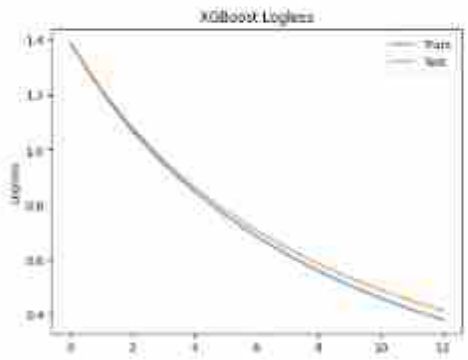
This model evaluation is assessed based on the accuracy, recall, and f-measure levels. This model evaluation is carried out based on the results of each label from the Kategory_ISPU column. Each accuracy, recall, and f-measure value obtained in the previous stage will be described in the form of a comparison table between the three sets of data proportions. Table 1 shows the classification report comparison results from each data proportion.

Table 1. Performance results of each data proportion.

Splitting Data	Accuracy	Recall	F-1 Score	Avg AUC
70:30	98%	94%	95%	0.92
80:20	98%	99%	99%	0.91
90:10	99%	99%	99%	0.91

The model was evaluated using three different data splitting strategies: 70:30, 80:20, and 90:10. Across these splits, the model consistently demonstrated high performance, as indicated by the Accuracy, Recall, F-1 Score, and Average AUC metrics. When the data was split 70:30, the model achieved an accuracy of 98%, with a recall of 94%, an F-1 Score of 95%, and an average AUC of 0.92. As the test set size decreased in the 80:20 split, the recall improved significantly to 99%, and both the F-1 Score and Accuracy remained at 99%, with a slightly lower average AUC of 0.91. A similar trend was observed in the 90 10 split, where the model's accuracy, recall, and F-1 Score were all at 99%, with the average AUC remaining at 0.91. These results suggest that the model's performance is robust across different data splitting ratios, with consistently high accuracy and recall. The slight variation in the Average AUC may reflect minor fluctuations in the model's ability to distinguish between classes as the test set size decreases, but overall, the model maintains strong predictive power.

Based on the results of table 1, it shows that there has been an indication of overfitting in each proportion of the data. We plotted the XGBoost log loss against the XGBoost performance results to see the patterns that occur in the train and test data.



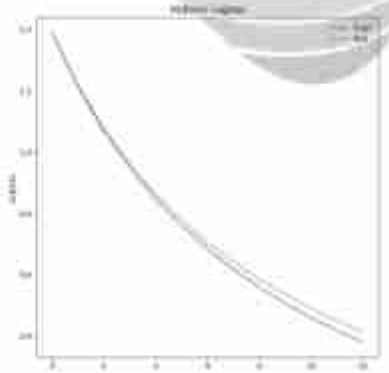


Figure 1(a) Logloss Data Proportion 70:30 **Figure 1(b)** Logloss Data Proportion 80:20

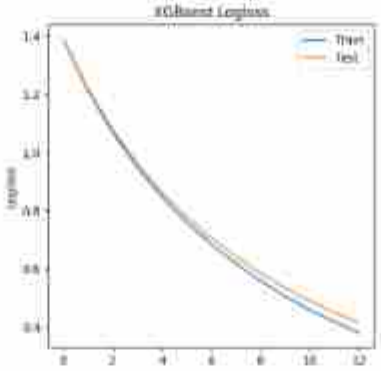
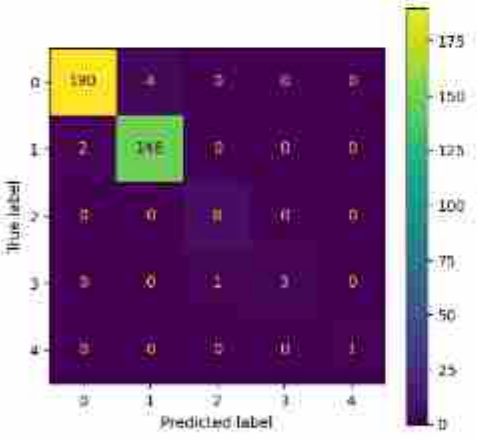


Figure 1(c) Logloss Data Proportion 90:10

The small gap between training and testing Logloss indicates that the model can generalize well to unseen data. This indicates that the model is neither underfitting (where the two curves will be high and far apart) nor overfitting (where the training curve will be much lower than the testing curve)(L.M. Patel et al, 2021). with this that the performance results at each proportion in table 1 are good performance results or can be called goodfitting.

Confusion Matrix

Confusion matrix is a method that can be used to measure the performance of a classification method and Confusion matrix contains information that compares the classification results performed by the system to measure its accuracy(R.A. Smith, 2021). The following are the results of the confusion matrix on each proportion of data.



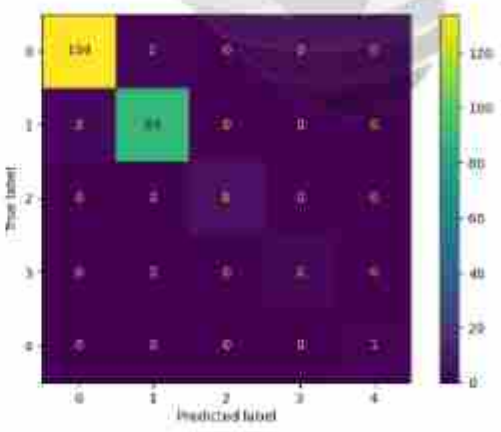


Figure 2(a) Confusion Matrix Data 70:30

Figure 2(b) Confusion Matrix Data 80:20

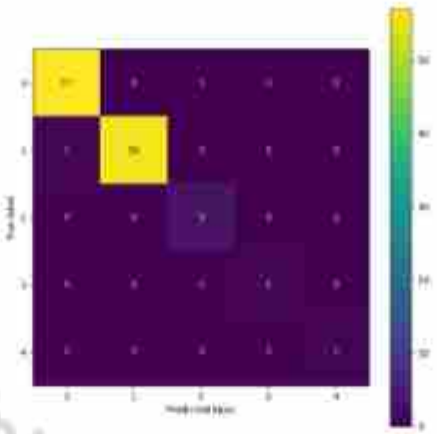


Figure 2 (c) Confusion Matrix Data 90:10

Figure 2(a), Figure 2(b) and Figure 2(c) show that the model performs well on classes 0 and 1 with a high number of correct predictions. The bottom right side of the matrix (classes 2, 3, and 4) has fewer observations, indicating either underrepresentation of these classes in the dataset or less accurate predictions by the model. Misclassification occurs mostly between classes 0 and 1 and slightly between classes 2 and 3.

ROC-AUC Curves

Here is a comparison of the curves of the three sets of data division that have been modeled.

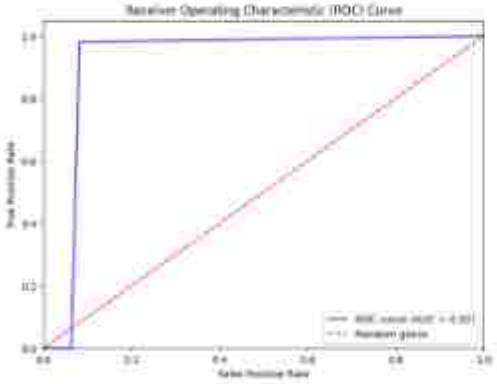


Figure 3 (a) ROC Curve Data 70:30

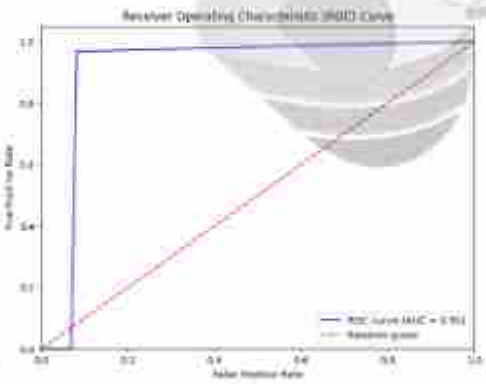


Figure 3(b) ROC Curve Data 80:20

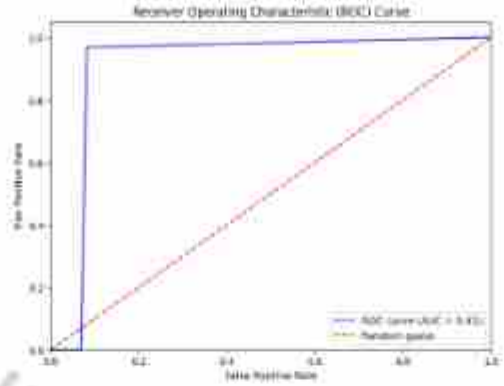


Figure 3(c) ROC Curve Data 90:10

Figures 3(a), 3(b) and 3(c) show the ROC curve that describes the test data and training data. From the curve, the value of Area Under the Curve (AUC) is obtained, which is the area under the ROC curve. The AUC value for the 70:30 data proportion is 0.92, the AUC value for the 80:20 data proportion is 0.91 and the AUC value for the 90:10 data proportion is 0.91. These three sets of data division can be concluded that the classification results obtained are good categories.

Feature Importance

In general, Feature Importance provides a value that indicates how valuable or impactful each feature is in the stages of creating a boosted decision tree model. The more inputs used to make the decision key with the decision tree, the higher the relative importance level(Ichwanul. Muslim. Karo.Karo, 2020). Based on the results of the importance feature on the Palembang City ISPU dataset, the PM2.5 parameter has a significant impact on the model used to predict each variable. In the prediction results in the application of the XGBoost algorithm, it can be seen that PM2.5 has a significant influence on the prediction results.

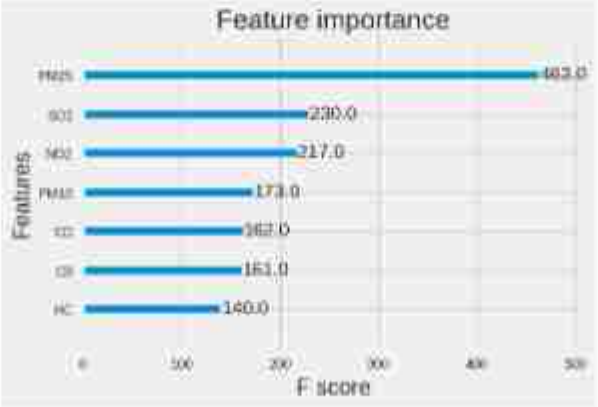


Figure 4. Result of Feature Importance

Visualization of ISPU parameter prediction results

Visualization of prediction results, at this stage will display the prediction results in the form of time series forecasting for one of the most influential ISPU parameters based on feature importance.

Figure 5. Merge PM2.5 dataset and forecast results

It can be seen that the ISPU value of the PM2.5 parameter throughout 2023 gets an ISPU value between 201-300 and an ISPU value of more than 301+ which is where the air quality conditions in Palembang city have had a significant negative impact on a number of exposed population segments and require rapid handling. Based on the results of the prediction of ISPU on the PM2.5 parameter for the next three years, it shows that the ISPU value of the PM2.5 parameter is at a value between the values of 51-100 and the value between 101-200 which is where the air conditions in Palembang City when the ISPU value is between 101-200 that is detrimental and causes insignificant negative impacts on human, animal and plant health.

CONCLUSION

From the results and discussion that has been described, it is concluded from the analysis of the accuracy level of ISPU air quality prediction in Palembang City using the Extreme Gradient Boosting algorithm method. The accuracy result for the proportion of 70:30 data is 98%, the recall value is 94%, the F-1 Score value is 95% and the average AUC value is 0.92. The accuracy result for the proportion of 80:20 data is 98%, the recall value is 99%, the F-1 Score value is 99% and the average AUC value is 0.91. The accuracy result for the 90:10 data proportion is 99%, the recall value is 99%, the F-1 Score value is 99% and the average AUC value is 0.91. The performance results in each data proportion do not occur overfitting and produce goodfitting. Feature Importance in this dataset is PM2.5 which gets the highest value

among other ISPU parameters which means that PM2.5 has a significant influence on the prediction modeling results.

REFERENCES

Ayus, I., Natarajan, N., & Gupta, D. (2023). Perbandingan teknik pembelajaran mesin dan pembelajaran mendalam untuk prediksi polusi udara: studi kasus dari Cina. Asian Journal of Atmospheric Environment, 17(1). https://doi.org/10.1007/s44273-023-00005-w

Danang Dwi Nugroho, 2020, "ANALISIS KERENTANAN TANAH LONGSOR MENGGUNAKAN METODE FREQUENCY RATIO DI KABUPATEN BANDUNG BARAT JAWA BARAT", S.T., Skripsi, Departement Teknik Geodesi, Institut Teknologi Nasional Bandung, Bandung, 2020.

Dian Tri Wilujeng, Mohamat Fatekurohman, I Made Tirta. "Analisis Risiko Kredit Perbankan Menggunakan Algoritma K-Nearest Neighbor dan Nearest Weighted K-Nearest Neighbor" in Indonesian Journal of Applied Statistics.2023. https://doi.org/10.13057/ijas.v5i2.58426

Chen, T., & Guestrin, C. (2016). "Xgboost: A scalable tree boosting system" Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, (pp. 785–794).

South Sumatra Provincial Health Office. "ISPU 345 in Palembang City, South Sumatra Provincial Health Office Socializes Health Education to the Public at Road Junctions". Accessed: November 29, 2023. Available: https://dinkes.sumselprov.go.id/2023/10/ispu-345-di-kota-palembang-dinkes-provinsi-sumsel-sosialisasikan-edukasi-kesehatan-ke-masyarakat-di-persimpangan-jalan/.

Herda Sabriyah Dara Kospa, Awaluddin A Praja. "Evaluation of Forest and Peatland Fire Prevention in Ogan Komering Ilir Regency, South Sumatra", Engineering Journal. Vol.13, No. 01.(1-9). pp.2, 2023.

Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Application of Extreme Gradient Boosting (XGBOOST) Method on Credit Card Customer Classification. Journal of Mathematics: Theory and Applications, 4(1), 21-26. https://doi.org/10.31605/jomta.v4i1.1792

Ichwanul Muslim Karo Karo. "Implementation of XGBoost and Feature Importance Methods for Classification on Forest and Land Fires". Journal of Software Engineering, Information and Communication Technology Vol. 1 No. 1, pp. 11-18, November 2020

Jason Brownlee, "Tune XGBoost Performance With Learning Curves", Accessed: 20 May 2024. Available: https://machinelearningmastery.com/tune-xgboost-performance-with-learning-curves/

K. Hasan, "Indonesia's air quality: Decline in 2023 due to lack of intervention and El Niño. what about 2024? ," Centre for Research on Energy and Clean Air, https://energyandcleanair.org/publication/indonesias-air-quality-decline-in-2023-due-to-lack-of-intervention-and-el-nino/ (accessed May. 4, 2024).

Kristiawan, Andreas Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel" in Jurnal Teknik Informatik dan Sistem Informasi. April 2021.

Kusnandar, M. (2020). Permen LHK Number 14 of 2020. Permen LHK Number 14 of 2020 concerning Air Pollutant Standard Index (ISPU), 1-16.

L.M. Patel, R.P. Sen, "Evaluating Model Performance in Machine Learning: Techniques and Tools", IEEE Transactions on Neural Networks and Learning Systems, Vol. 36, pp 10-12, Maret 2021.

R. A. Smith, H. S. Gupta, "Confusion Matrix-Based Performance Evaluation of Classification Algorithms". IEEE Access. 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP). Maret 2021.

Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Prediksi kualitas udara dengan model pembelajaran mesin: Sebuah studi prediktif di kota pesisir India, Visakhapatnam. Chemosphere, 338(May), 139518. https://doi.org/10.1016/j.chemosphere.2023.139518

Shafila, G. A. (2020). Implementation of Extreme Gradient Boosting (Xgboost) Method for Classification on Bioinformatics Data (Case Study: Ebola Disease, GSE 122692). Dspace.Uii.Ac.Id,1-77. https://dspace.uii.ac.id/bitstream/handle/123456789/29276/16611022 Gregy Addis Shafila.pdf?sequence=1&isAllowed=y

Syafrida Hafni Sahir. "Research Methodology". University of Medan Area. ISBN:978-623-6155-06-6. KBM Indonesia Publisher: Yogyakarta. pp 5-11. January 2022Summers, J. O., 2001. Guideline for conducting research and publishing in marketing: From conceptualization through the review process. *Journal of the Academy of Marketing Science* 29 (4): 405-415.

Shafila, G. A. (2020). Implementation of Extreme Gradient Boosting (Xgboost) Method for Classification on Bioinformatics Data (Case Study: Ebola Disease, GSE 122692). Dspace.Uii.Ac.Id, 1-77. https://dspace.uii.ac.id/bitstream/handle/123456789/29276/16611022 Gregy Addis Shafila.pdf?sequence=1&isAllowed=y