TECHNOMEDIA : Informatics and Computer Science, Vol. 2 No. 1, Januari 2025, 15 - 20
e-ISSN: 3047-2180 | DOI: 10.58641

IMPLEMENTASI BIG DATA ANALYTICS DALAM KLASIFIKASI KUALITAS UDARA MENGGUNAKAN ALGORITMA GRADIENT-BOOSTED TREE CLASSIFIER PADA PYSPARK

[1)]Muhamad Fuat Asnawi, [2)]Nur Fitriyanto, [3)] M. Agoeng Pamoengkas
[1)]Universitas Sains Al-Qur'an
[1,2,3)]Mahasiswa S2 PJJ Informatika, Universitas Amikom Yogyakarta
[1)]fuatasnawi@unsiq.ac.id, [2)]nur.fitriyanto@students.amikom.ac.id, [3)]agoeng@students.amikom.ac.id

INFO ARTIKEL

Riwayat Artikel :
Diterima : 10 Januari 2025
Disetujui : 29 Januari 2025

Kata Kunci :
Big Data Analytics, Gradient-Boosted Tree, Kualitas Udara, PySpark

ABSTRAK

Penelitian ini bertujuan untuk mengklasifikasikan kualitas udara berdasarkan parameter PM1.0, PM2.5, dan PM10 dengan memanfaatkan pendekatan Big Data Analytics menggunakan algoritma Gradient-Boosted Tree Classifier (GBT) yang diimplementasikan pada framework PySpark. Dataset yang digunakan diunduh dari OpenAQ, mencakup periode 14 April 2021 hingga 16 April 2023, dengan total 1.048.154 entri, menunjukkan volume data yang besar dan kompleks. Proses penelitian meliputi pra-pemrosesan data untuk menangani ketidakseimbangan data, pembagian dataset untuk pelatihan dan pengujian, serta tuning hyperparameter menggunakan grid search dan cross-validation untuk mengoptimalkan kinerja model. Dengan memanfaatkan keunggulan PySpark dalam pemrosesan data besar secara paralel, model GBT berhasil mencapai akurasi 98.87%, presisi 99.00%, recall 98.87%, dan F1-Score 98.90%. Penelitian ini menunjukkan bagaimana Big Data Analytics dapat meningkatkan efisiensi dan akurasi dalam klasifikasi kualitas udara, memberikan kontribusi signifikan dalam pengembangan sistem pemantauan real-time yang mendukung mitigasi polusi udara dan pengambilan kebijakan berbasis data.

ARTICLE INFO

Article History :
Received : January 10, 2025
Accepted : January 29, 2025

Keywords:
Big Data Analytics, Gradient-Boosted Tree, Air Quality, PySpark

ABSTRACT

This study aims to classify air quality based on PM1.0, PM2.5, and PM10 parameters using a Big Data Analytics approach with the Gradient-Boosted Tree Classifier (GBT) algorithm implemented on the PySpark framework. The dataset used was downloaded from OpenAQ, covering the period from April 14, 2021, to April 16, 2023, with a total of 1,048,154 entries, representing a large and complex volume of data. The research process includes data preprocessing to address data imbalance, dataset splitting for training and testing, and hyperparameter tuning using grid search and cross-validation to optimize model performance. By leveraging PySpark's advantage in parallel processing of large data, the GBT model achieved an accuracy of 98.87%, precision of 99.00%, recall of 98.87%, and an F1-Score of 98.90%. This study demonstrates how Big Data Analytics can enhance efficiency and accuracy in air quality classification, contributing significantly to the development of real-time monitoring systems that support air pollution mitigation and data-driven policy-making.

1. PENDAHULUAN

Polusi udara menjadi ancaman besar terhadap kesehatan manusia dan ekosistem global. Data menunjukkan bahwa paparan terhadap polusi udara, terutama partikel partikulat seperti PM1.0, PM2.5, dan PM10, dapat menyebabkan berbagai masalah kesehatan, mulai dari gangguan pernapasan, penyakit kardiovaskular, hingga kematian dini. Menurut World Health Organization (WHO), sekitar 92% populasi dunia menghirup udara yang tidak memenuhi standar kualitas udara, yang mengakibatkan lebih dari 7 juta kematian setiap tahunnya akibat polusi udara ambien (Muthukumar, Cocom, et al., 2022). Selain dampak kesehatan, polusi udara memberikan beban ekonomi yang besar, dengan estimasi biaya global sebesar $5 triliun per tahun, termasuk biaya perawatan kesehatan dan kehilangan produktivitas (Ren et al., 2020).

Kualitas udara dipengaruhi oleh berbagai faktor, termasuk aktivitas manusia seperti emisi kendaraan bermotor, pembakaran biomassa, dan proses industri, serta fenomena alam seperti transportasi jarak jauh partikel atmosfer. Parameter seperti PM2.5 digunakan sebagai indikator utama untuk mengukur kualitas udara, mengingat partikel ini memiliki kemampuan menembus jauh ke dalam sistem pernapasan manusia dan menimbulkan dampak kesehatan yang signifikan (Koo et al., 2023).

Metode tradisional untuk memprediksi polusi udara, seperti model transportasi kimia, sering kali menghadapi keterbatasan struktural dan ketidaksesuaian data meteorologi dan emisi (Koo et al., 2023). Alternatif modern seperti deep learning dan algoritma machine learning telah digunakan untuk menangkap pola spasial dan temporal dalam data polusi udara. Model seperti Graph Convolutional Network (GCN) dan Convolutional Long Short-Term Memory (ConvLSTM) menawarkan kemampuan untuk menganalisis hubungan non-linear dalam data kualitas udara, namun implementasinya masih menghadapi kendala efisiensi dan keterbatasan data spasial-temporal (Muthukumar, Pathak, et al., 2022).

Meskipun berbagai pendekatan modern telah dikembangkan, prediksi kualitas udara masih menghadapi tantangan besar. Keterbatasan utama terletak pada kualitas dan cakupan data, seperti distribusi spasial yang tidak merata dari sensor pemantauan, serta kebutuhan untuk memproses volume data besar secara efisien. Selain itu, masih kurangnya integrasi antara data satelit, sensor darat, dan model prediktif yang dapat menghasilkan informasi yang lebih akurat dan kontekstual (Ren et al., 2020).

Idealnya, sebuah sistem pemantauan kualitas udara harus mampu menangani data besar, mencakup wilayah yang luas, dan memberikan hasil prediksi secara real-time. Namun, kenyataannya, banyak model yang ada masih terbatas dalam menangkap pola spasial-temporal secara holistik dan kurang efisien dalam menangani data yang kompleks. Gap ini menunjukkan perlunya integrasi teknologi Big Data Analytics dan algoritma machine learning untuk menghasilkan sistem prediksi yang lebih andal (Zhou et al., 2023).

Penelitian ini memanfaatkan algoritma Gradient-Boosted Tree Classifier (GBT) dalam framework PySpark untuk mengatasi keterbatasan metode sebelumnya. Dibandingkan penelitian terdahulu seperti yang dilakukan oleh Bai, yang mengintegrasikan data satelit dan model numerik untuk menghasilkan dataset LGHAP, penelitian ini berfokus pada pengolahan data kualitas udara berskala besar dari OpenAQ (Bai et al., 2022). Dengan menggabungkan data dari berbagai sumber dan algoritma machine learning, penelitian ini bertujuan untuk memberikan hasil prediksi yang lebih akurat serta mendukung kebijakan mitigasi polusi udara yang lebih baik (Deng et al., 2020).

Penelitian ini bertujuan untuk mengembangkan model prediksi kualitas udara berbasis algoritma Gradient-Boosted Tree Classifier yang diimplementasikan dalam framework PySpark. Selain itu, penelitian ini bertujuan untuk mengidentifikasi faktor-faktor utama yang memengaruhi polusi udara serta memberikan solusi berbasis data untuk mendukung pengambilan keputusan yang lebih baik dalam pengelolaan kualitas udara.

2. METODE

Penelitian ini merupakan studi kuantitatif yang berfokus pada klasifikasi kualitas udara berdasarkan parameter PM1.0, PM2.5, dan PM10 dengan menggunakan algoritma

Gradient-Boosted Tree Classifier (GBT) yang diimplementasikan pada framework PySpark (Su, 2020). Penelitian ini bersifat eksperimental, di mana data kualitas udara dikumpulkan terlebih dahulu dari platform OpenAQ melalui proses pengunduhan manual (Jin et al., 2022). Data yang digunakan mencakup periode waktu dari 14 April 2021 hingga 16 April 2023 dengan total 1.048.154 data. Pendekatan yang digunakan dalam penelitian ini adalah pendekatan analitik berbasis data besar (Big Data Analytics), yang memungkinkan pengolahan dan analisis data secara terdistribusi dan efisien. Metode analisis data dilakukan dalam beberapa tahap, dimulai dengan proses pra-pemrosesan untuk membersihkan dan mempersiapkan data, dilanjutkan dengan tuning hyperparameter algoritma GBT menggunakan grid search dan cross-validation untuk memperoleh model yang optimal (Cheng et al., 2021). Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score

3. HASIL DAN PEMBAHASAN

1) Dataset

Dataset yang digunakan berjumlah 1.048.154 data, mencakup periode 14 April 2021 hingga 16 April 2023. Data ini diproses untuk memastikan integritas, termasuk pembersihan nilai yang hilang dan transformasi format waktu. Berikut ini gambaran dari dataset yang digunakan:

location_id	sensors_id	location	datetime	lat	lon	parameter	units	value
222830	1292732	Sens6_855b-235252	2021-04-14 08:34:16	35.147285	33.415157	pm10	µg/m³	57.2
222830	1292732	Sens6_855b-235252	2021-04-14 08:38:16	35.147285	33.415157	pm10	µg/m³	34.2
222830	1292732	Sens6_855b-235252	2021-04-14 08:46:16	35.147285	33.415157	pm10	µg/m³	40.7
222830	1292732	Sens6_855b-235252	2021-04-14 08:52:18	35.147285	33.415157	pm10	µg/m³	24.7
222830	1292732	Sens6_855b-235252	2021-04-14 08:54:21	35.147285	33.415157	pm10	µg/m³	44.0

only showing top 5 rows

Gambar 2. Gambaran dataset 5 baris pertama

2) Data Preprocessing

Tahapan preprocessing pada penelitian mencakup pengklasifikasian data kualitas udara menjadi dua kelas biner, yaitu "Sehat" (label 0) untuk nilai parameter kurang dari 29 dan "Tidak Sehat" (label 1) untuk nilai parameter sama dengan atau lebih dari 29. Setelah itu, dilakukan verifikasi distribusi label, yang menunjukkan ketidakseimbangan data dengan mayoritas data pada label "Tidak Sehat". Selanjutnya, fitur-fitur seperti location_id, lat, lon, dan value digabungkan menjadi satu vektor menggunakan VectorAssembler untuk mempersiapkan data bagi model machine learning. Fitur yang telah digabungkan kemudian dinormalisasi menggunakan StandardScaler untuk memastikan data memiliki rata-rata nol dan standar deviasi satu, sehingga meningkatkan kinerja algoritma pada tahap pemodelan.

3) Pembagian dataset

Pembagian data pada penelitian ini dilakukan menggunakan metode randomSplit, di mana dataset dibagi menjadi dua subset utama: train_data (80% dari total data) dan test_data (20% dari total data). Proporsi pembagian ini bertujuan untuk memastikan bahwa sebagian besar data digunakan untuk melatih model (training), sementara sisanya digunakan untuk menguji kinerja model (testing). Parameter seed diatur ke nilai tetap (42) untuk memastikan pembagian data bersifat reproducible, sehingga menghasilkan hasil yang konsisten setiap kali kode dijalankan(Mishra et al., 2020). Pembagian data seperti ini merupakan langkah penting untuk mengevaluasi generalisasi model pada data yang belum pernah dilihat sebelumnya.

4) Pemodelan dengan GBT

Gradient-Boosted Tree Classifier (GBT) digunakan dalam penelitian ini untuk mengklasifikasikan kualitas udara berdasarkan fitur yang telah dinormalisasi (scaledFeatures) dan label biner (label). Model diinisialisasi dengan parameter dasar seperti jumlah maksimum iterasi (maxIter=20), yang menentukan jumlah pohon dalam ensemble. GBT bekerja dengan membangun pohon keputusan secara iteratif, di mana setiap pohon bertujuan memperbaiki kesalahan dari pohon sebelumnya, sehingga efektif menangkap pola non-linear dalam data. Untuk memastikan performa model optimal, dilakukan hyperparameter tuning menggunakan grid search untuk menguji kombinasi parameter maxDepth (kedalaman pohon) dan stepSize (ukuran langkah pembelajaran) (Le & Thu Hien, 2024; Liu et al., 2021; Xu et al., 2020). Hyperparameter tuning melibatkan pencarian kombinasi parameter terbaik melalui grid search dengan tiga nilai untuk maxDepth (5, 7, dan 10) dan stepSize (0.1, 0.2, 0.3), menghasilkan sembilan kombinasi(S. Wang et al., 2022; Yang et al., 2023). Evaluasi dilakukan menggunakan 5-fold cross-validation dengan metrik akurasi untuk memastikan model dapat bekerja secara

konsisten pada berbagai subset data. Hasil tuning menunjukkan parameter optimal pada maxDepth=5 dan stepSize=0.1, yang memberikan keseimbangan antara akurasi dan generalisasi. Model akhir yang dilatih menggunakan parameter terbaik ini siap untuk digunakan dalam menguji data baru, memastikan prediksi yang robust dan akurat.

```
Best Model Parameters:
  - Max Depth: 5
  - Step Size: 0.1
```

Gambar 3. Hasil tuning

5) Evaluasi Model

```
Accuracy: 0.9887417660460815
Precision: 0.9900188129941272
Recall: 0.9887417660460815
F1-Score: 0.9890452638312383
```

Gambar 4. Nilai evaluasi model

Pada gambar 4. Hasil evaluasi model menunjukkan performa yang sangat baik dengan akurasi sebesar 98.87%, yang berarti hampir semua prediksi model terhadap data pengujian adalah benar. Presisi sebesar 99.00% menunjukkan kemampuan model dalam meminimalkan false positives, memastikan prediksi "Tidak Sehat" sangat akurat. Recall sebesar 98.87% mengindikasikan bahwa model mampu mendeteksi hampir semua data yang benar-benar "Tidak Sehat". Kombinasi presisi dan recall menghasilkan F1-Score sebesar 98.90%, yang mencerminkan keseimbangan performa model dalam mendeteksi data positif tanpa mengorbankan akurasi secara keseluruhan. Hasil ini menunjukkan bahwa model Gradient-Boosted Tree Classifier yang diimplementasikan mampu mengklasifikasikan kualitas udara dengan tingkat keandalan yang tinggi.

6) Hasil dan Analisis

Dibandingkan dengan metode tradisional seperti regresi linear, Random Forest, Graph Convolutional Network (GCN) dan Convolutional Long Short-Term Memory (ConvLSTM) penggunaan GBT dalam penelitian ini menunjukkan peningkatan signifikan dalam akurasi dan efisiensi waktu komputasi (An et al., 2023; Asgari et al., 2022; Karampelas & Sotiropoulos, 2022; Muthukumar et al., 2020; Muthukumar, Pathak, et al., 2022; Sibyan et al., 2022; C. Wang et al., 2021). Selain itu, framework PySpark memungkinkan pemrosesan data besar secara paralel, yang mempercepat waktu pelatihan model.

Model ini memiliki potensi besar untuk digunakan sebagai dasar dalam sistem pemantauan kualitas udara real-time yang mampu memberikan informasi akurat kepada masyarakat dan pembuat kebijakan, sehingga mendukung inisiatif mitigasi polusi udara terutama di daerah dengan tingkat polusi tinggi. Namun, model ini memiliki keterbatasan yang perlu diperhatikan, yaitu ketergantungan pada kualitas data dari OpenAQ, yang mungkin memiliki kekurangan seperti distribusi spasial yang tidak merata. Selain itu, faktor meteorologi seperti kelembaban dan suhu belum dimasukkan ke dalam analisis, padahal faktor-faktor tersebut dapat memengaruhi distribusi dan konsentrasi polutan, sehingga potensi peningkatan akurasi masih dapat dieksplorasi lebih lanjut.

4. PENUTUP

Kesimpulan

Dari hasil penelitian ini dapat disimpulkan bahwa algoritma Gradient-Boosted Tree Classifier (GBT) yang diimplementasikan menggunakan framework PySpark mampu mengklasifikasikan kualitas udara dengan tingkat akurasi tinggi, mencapai 98.87%. Pemodelan ini efektif menangkap hubungan non-linear dalam data kualitas udara, yang diperkuat dengan proses hyperparameter tuning untuk mengoptimalkan performa model. Dataset yang digunakan, mencakup lebih dari satu juta entri dari OpenAQ, memberikan cakupan temporal yang luas, meskipun distribusi spasialnya masih menjadi keterbatasan. Penelitian ini memberikan kontribusi penting dalam mendukung sistem pemantauan kualitas udara real-time yang dapat memberikan informasi akurat kepada pembuat kebijakan dan masyarakat. Namun, penelitian ini masih dapat dikembangkan lebih lanjut dengan memasukkan faktor meteorologi seperti suhu dan kelembaban, serta memperbaiki cakupan data spasial untuk meningkatkan generalisasi dan akurasi prediksi.

5. DAFTAR PUSTAKA

An, Z., Gui, H., Song, Y., & Liu, J. (2023). Prediction of Air Quality Based on Artificial Intelligence Regression Model. *Proceedings - 2023 2nd International Conference on Artificial Intelligence*

and Autonomous Robot Systems, AIARS 2023, 288–292. https://doi.org/10.1109/AIARS59518.2023.00065

Asgari, M., Yang, W., & Farnaghi, M. (2022). Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework. *Environmental Technology and Innovation*, *27*. https://doi.org/10.1016/j.eti.2022.102776

Bai, K., Li, K., Ma, M., Li, K., Li, Z., Guo, J., Chang, N. Bin, Tan, Z., & Han, D. (2022). LGHAP: The Long-Term Gap-free High-resolution Air Pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion. *Earth System Science Data*, *14*(2), 907–927. https://doi.org/10.5194/essd-14-907-2022

Cheng, B., Ma, Y., Feng, F., Zhang, Y., Shen, J., Wang, H., Guo, Y., & Cheng, Y. (2021). Influence of weather and air pollution on concentration change of PM2.5 using a generalized additive model and gradient boosting machine. *Atmospheric Environment*, *255*. https://doi.org/10.1016/j.atmosenv.2021.118437

Deng, F., Lv, Z., Qi, L., Wang, X., Shi, M., & Liu, H. (2020). A big data approach to improving the vehicle emission inventory in China. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16579-w

Jin, C., Wang, Y., Li, T., & Yuan, Q. (2022). Global Validation and Hybrid Calibration of CAMS and MERRA-2 PM2.5 Reanalysis Products Based on OpenAQ platform. *Atmospheric Environment*, *274*. https://doi.org/10.5281/zenodo.5168102

Karampelas, G., & Sotiropoulos, D. N. (2022). Analysis and Prediction of Air Pollutant Indices using Bidirectional-Convolutional LSTMs. *13th International Conference on Information, Intelligence, Systems and Applications, IISA 2022*. https://doi.org/10.1109/IISA56318.2022.9904392

Koo, Y. S., Choi, Y., & Ho, C. -H. (2023). Air Quality Forecasting Using Big Data and Machine Learning Algorithms. In *Asia-Pacific Journal of Atmospheric Sciences* (Vol. 59, Issue 5, pp. 529–530). Korean Meteorological Society. https://doi.org/10.1007/s13143-023-00347-z

Le, X. H., & Thu Hien, L. T. (2024). Predicting maximum scour depth at sluice outlet: a comparative study of machine learning models and empirical equations. *Environmental Research Communications*, *6*(1). https://doi.org/10.1088/2515-7620/ad1f94

Liu, M., Chen, H., Wei, D., Wu, Y., & Li, C. (2021). Nonlinear relationship between urban form and street-level PM2.5 and CO based on mobile measurements and gradient boosting decision tree models. *Building and Environment*, *205*. https://doi.org/10.1016/j.buildenv.2021.108265

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry*, *132*, 116045. https://doi.org/10.1016/J.TRAC.2020.116045

Muthukumar, P., Cocom, E., Nagrecha, K., Comer, D., Burga, I., Taub, J., Calvert, C. F., Holm, J., & Pourhomayoun, M. (2022). Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data. *Air Quality, Atmosphere and Health*, *15*(7), 1221–1234. https://doi.org/10.1007/s11869-021-01126-3

Muthukumar, P., Cocom, E., Nagrecha, K., Holm, J., Comer, D., Lyons, A., Burga, I., Calvert, C. F., & Pourhomayoun, M. (2020). Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM. *Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020*, 521–526. https://doi.org/10.1109/CSCI51800.2020.00094

Muthukumar, P., Pathak, S., Nagrecha, K., Hosseini, H., Comer, D., Amini, N., Holm, J., & Pourhomayoun, M. (2022). Multi-Pollutant Ground-level Air Pollution Prediction through Deep MeteoGCN-ConvLSTM. *Proceedings - 2022 International Conference on Computational Science and Computational Intelligence, CSCI 2022*, 26–34. https://doi.org/10.1109/CSCI58124.2022.00012

Ren, X., Zhao, Y., Wu, L., Jiang, J., Zhang, C., Tang, X., Han, L., Zhu, M., & Wang, Z. (2020). Towards efficient digital governance of city air pollution using technique of big atmospheric environmental data. *IOP Conference Series: Earth and Environmental Science*, *502*(1). https://doi.org/10.1088/1755-1315/502/1/012031

Sibyan, H., Svajlenka, J., Hermawan, H., Faqih, N., & Arrizqi, A. N. (2022). Thermal Comfort Prediction Accuracy with Machine Learning between Regression Analysis and Naïve Bayes Classifier. *Sustainability (Switzerland)*, *14*(23). https://doi.org/10.3390/su142315663

Su, Y. (2020). Prediction of air quality based on Gradient Boosting Machine Method. *Proceedings - 2020 International Conference on Big Data and Informatization Education, ICBDIE 2020*, 395–397. https://doi.org/10.1109/ICBDIE50010.2020.00099

Wang, C., Zhu, Y., Zang, T., Liu, H., & Yu, J. (2021). Modeling Inter-station Relationships with Attentive Temporal Graph Convolutional Network for Air Quality Prediction. *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 616–624. https://doi.org/10.1145/3437963.3441731

Wang, S., Ma, C., Xu, Y., Wang, J., & Wu, W. (2022). A Hyperparameter Optimization Algorithm for the LSTM Temperature Prediction Model in Data Center. *Scientific Programming*, *2022*. https://doi.org/10.1155/2022/6519909

Xu, J., Wang, A., Schmidt, N., Adams, M., & Hatzopoulou, M. (2020). A gradient boost approach for predicting near-road ultrafine particle concentrations using detailed traffic

characterization. *Environmental Pollution*, *265*, 114777. https://doi.org/10.1016/J.ENVPOL.2020.114777

Yang, F., Jiang, X., & Chen, Z. (2023). Air Quality Index Prediction Model Based on Multiple Attention Mechanisms and Hyperparameter Optimization. *2023 3rd International Conference on Electronic Information Engineering and Computer Science, EIECS 2023*, 603–606. https://doi.org/10.1109/EIECS59936.2023.10435483

Zhou, P., Ma, J., Li, X., Zhao, Y., Yu, K., Su, R., Zhou, R., Wang, H., & Wang, G. (2023). The long-term and short-term effects of ambient air pollutants on sleep characteristics in the Chinese population: big data analysis from real world by sleep records of consumer wearable devices. *BMC Medicine*, *21*(1). https://doi.org/10.1186/s12916-023-02801-1