

KLASIFIKASI KUALITAS UDARA MENGGUNAKAN *RANDOM FOREST* DENGAN PENERAPAN REGRESI K-NN DAN SMOTE UNTUK MENGATASI DATA HILANG DAN KELAS TIDAK SEIMBANG

SKRIPSI

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana
di Jurusan Matematika pada Fakultas MIPA**

Oleh :
INDAH LESTARI
08011382025103



**JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SRIWIJAYA
2025**

HALAMAN PENGESAHAN

KLASIFIKASI KUALITAS UDARA MENGGUNAKAN *RANDOM FOREST* DENGAN PENERAPAN REGRESI K-NN DAN SMOTE UNTUK MENGATASI DATA HILANG DAN KELAS TIDAK SEIMBANG

SKRIPSI

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana
di Jurusan Matematika pada Fakultas MIPA**

Oleh :

INDAH LESTARI

08011382025103

Indralaya, 23 April 2025

Pembimbing Kedua

Pembimbing Utama



Dr. Endang Sri Kresnawati, S.Si., M.Si

NIP. 197702082002122003



Prof. Yulia Resti, S.Si., M.Si. Ph.D

NIP. 197307191997022001

**Mengetahui,
Ketua Jurusan Matematika**



Dr. Dian Cahyawati Sukanda, S.Si., M.Si.

NIP. 197303212000122001

PERNYATAAN KEASLIAN KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama Mahasiswa : Indah Lestari

NIM : 08011382025103

Fakultas/Jurusan : Matematika dan Ilmu Pengetahuan Alam/Matematika

Menyatakan bahwa skripsi ini adalah hasil karya ilmiah saya sendiri dan karya ilmiah ini belum pernah diajukan sebagai pemenuhan persyaratan untuk memperoleh gelar kesarjanaan strata satu (S1) dari Universitas Sriwijaya maupun perguruan tinggi lain.

Semua informasi yang dimuat didalam skripsi ini yang berasal dari penulis lain baik yang dipublikasi atau tidak telah diberikan penghargaan dengan mengutip nama sumber penulis baik yang secara benar. Semua isi dari skripsi ini sepenuhnya menjadi tanggung jawab saya sebagai penulis.

Demikianlah surat pernyataan ini saya buat dengan sebenarnya.

Indralaya, 23 April 2025
Penulis



Indah Lestari
NIM. 08011382025103

LEMBAR PERSEMBAHAN

MOTTO

**“Selalu ada harga dalam sebuah proses. Nikmati saja lelah-lelah itu.
Lebarkan lagi rasa sabar itu. Semua yang kau investasikan untuk
menjadikan dirimu serupa yang kau impikan. Mungkin tidak akan selalu
berjalan lancar. Tapi gelombang-gelombang itu yang nanti bisa kau
ceritakan.”**

(Boy Chandra)

**“Aku membahayakan nyawa ibu untuk lahir ke dunia, jadi tidak mungkin
aku tidak ada artinya.”**

Skripsi ini kupersembahkan untuk:

- 1. Allah SWT**
- 2. Rasulullah Muhammad SAW**
- 3. Bapak dan Mamak**
- 4. Saudara dan Keluarga Besar**
- 5. Dosen**
- 6. Almamater**
- 7. Diriku Sendiri**

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh

Puji syukur atas kehadiran Tuhan Yang Maha Esa, Allah *Subhanahu wa ta'ala* yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi berjudul “**Klasifikasi Kualitas Udara Menggunakan *Random Forest* dengan Penerapan Regresi K-NN dan SMOTE untuk mengatasi Data Hilang dan Kelas Tidak Seimbang**” sebagai salah satu syarat memperoleh gelar sarjana sains studi Matematika di Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.

Penulis sadar bahwa dalam proses pembuatan skripsi ini sebagai salah satu proses pembelajaran yang sangat berharga dan masih banyak kekurangan, serta banyaknya rintangan dalam pengerjaannya. Dengan penuh rasa hormat penulis juga mengucapkan terima kasih yang sebesar-besarnya kepada cinta pertama dan panutanku, Ayahanda **Seno Sukadi** dan pintu surgaku Ibunda **Maryati** yang selalu mendoakan dan memberikan dukungan yang tak henti-henti kepada penulis. Terima kasih atas segala pengorbanan dan tulus kasih yang di berikan. Beliau memang tidak sempat merasakan pendidikan bangku perkuliahan, namun mereka mampu senantiasa memberikan yang terbaik hingga penulis mampu menyelesaikan studinya sampai meraih gelar sarjana. Semoga Bapak dan Mamak sehat, panjang umur dan bahagia selalu. Penulisan ini tidak lepas dari bantuan beberapa pihak, oleh karena itu penulis juga menyampaikan rasa terima kasih yang tak terhingga kepada:

1. Bapak **Prof. Hermansyah, S.Si., M.Si., Ph.D.** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.
2. Ibu **Dr. Dian Cahyawati Sukanda, M.Si** selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya, dan selaku Dosen Pembahas Pertama serta Ketua Pelaksanaan Seminar yang telah memberikan tanggapan, masukan, saran, kritikan yang bermanfaat dalam penyelesaian skripsi dan memandu jalannya seminar sehingga seminar dapat dilaksanakan dengan baik
3. Ibu **Dr. Oki Dwipurwani, M.Si** selaku Dosen Pembimbing Akademik yang telah memberikan bimbingan, pengarahan, dan saran selama masa perkuliahan di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.
4. Ibu **Prof. Yulia Resti, M.Si., Ph.D.** selaku Dosen Pembimbing Utama yang telah meluangkan waktu untuk memberikan ide pemikiran, saran, kritikan, masukan, nasehat, dan membimbing penulis sehingga penulis mampu menyelesaikan skripsi dengan baik.
5. Ibu **Dr. Endang Sri Kresnawati, M.Si** selaku Dosen Pembimbing Kedua yang telah meluangkan waktu untuk memberikan ide pemikiran, saran, kritikan, masukan, nasehat, dan membimbing penulis sehingga penulis mampu menyelesaikan skripsi dengan baik.
6. Bapak **Dr. Bambang Suprihatin, M.Si** selaku Dosen Pembahas Kedua yang telah memberikan tanggapan, masukan, saran dan kritikan yang bermanfaat dalam penyelesaian skripsi.

7. **Seluruh Dosen di Jurusan Matematika FMIPA Universitas Sriwijaya** yang telah memberikan ilmu yang bermanfaat kepada penulis selama selama perkuliahan di jurusan Matematika.
8. Bapak **Irwansyah** dan ibu **Hamida** yang telah membantu proses administrasi selama di jurusan Matematika FMIPA Universitas Sriwijaya.
9. Saudara-saudara yang saya sayangi **Reni Aprianti, Evi Oktavia, dan Richi Febrianto** yang selalu memberikan perhatian, kasih sayang, dukungan kepada penulis sampai saat ini.
10. Kepada **Ade Elna Febrianti, Dian Putri Miranda, dan Ria Maharani** yang selalu menemani keseharian penulis dengan penuh kebahagiaan dan memberi dukungan kepada penulis.
11. Sahabat sejak SMA **Deliana Pardede** yang selalu mendukung dan memberi semangat kepada penulis. Terima kasih telah menjadi bagian dari perjalanan penulis sampai saat ini.
12. **Keluarga Matematika 20**, teman seperjuangan, dan rekan-rekan selama perkuliahan.
13. Terima kasih kepada **Indah Lestari**, perempuan sederhana yang memiliki keinginan tinggi namun sulit dimengerti, sang penulis karya tulis. Semoga langkah kebaikan terus berada padamu dan semoga Allah selalu meridhoi setiap perbuatanmu dan selalu dalam lindungan-Nya.
14. **Semua Pihak** yang terlibat dan terkait dalam hidup penulis yang tidak dapat penulis sebutkan satu persatu. Semoga segala kebaikan yang diberikan mendapat balasan dari Allah SWT.

Akhir kata semoga skripsi ini dapat menambah pengetahuan dan bermanfaat bagi mahasiswa/mahasiswi Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya dan seluruh pihak yang membutuhkan.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, 23 April 2025

Penulis

AIR QUALITY CLASSIFICATION USING RANDOM FOREST WITH THE APPLICATION OF K-NN REGRESSION AND SMOTE TO OVERCOME MISSING DATA AND UNBALANCED CLASSES

By:

Indah Lestari

08011382025103

ABSTRACT

The lives of living things are highly dependent on air quality, especially humans. Air quality prediction that classifies air quality into categories can involve many factors including factors that affect pollution levels. This research classifies air quality based on $PM_{2.5}$, PM_{10} , NO_2 , CO , SO_2 , and O_3 which are the main factors that affect pollution. The secondary data used in the study was obtained from Kaggle totaling 108.035. In this data there are more than 10,000 missing values and there is a class imbalance, to fill in the missing values the K-Nearest Neighbor method is used and for balancing class the Synthetic Minority Oversampling Technique (SMOTE) method is used. As for classifying the level of accuracy of air quality, the Random Forest method is used. The results of this study obtained the highest level of accuracy, precision, recall, and F1-score in the Random Forest classification method, respectively 76,192%, 72%, 76%, and 73%.

Keywords: Air Quality, K-Nearest Neighbor, SMOTE, Random Forest

KLASIFIKASI KUALITAS UDARA MENGGUNAKAN *RANDOM FOREST* DENGAN PENERAPAN REGRESI K-NN DAN SMOTE UNTUK MENGATASI DATA HILANG DAN KELAS TIDAK SEIMBANG

Oleh:

Indah Lestari

08011382025103

ABSTRAK

Kehidupan makhluk hidup sangat bergantung pada kualitas udara, terutama manusia. Prediksi kualitas udara yang mengklasifikasi kualitas udara menjadi beberapa kategori dapat melibatkan banyak faktor termasuk faktor-faktor yang mempengaruhi tingkat polusi. Penelitian ini mengklasifikasi kualitas udara berdasarkan $PM_{2.5}$, PM_{10} , NO_2 , CO , SO_2 , dan O_3 yang merupakan faktor-faktor utama yang mempengaruhi polusi. Data sekunder yang digunakan dalam penelitian diperoleh dari *Kaggle* berukuran 108.035. Pada data tersebut terdapat nilai hilang sebanyak lebih dari 10.000 dan terdapat ketidakseimbangan kelas, untuk mengisi nilai hilang digunakan metode *K-Nearest Neighbor* dan untuk penyeimbangan kelas digunakan metode *Synthetic Minority Oversampling Technique* (SMOTE). Sedangkan untuk pengklasifikasian tingkat ketepatan kualitas udara digunakan metode *Random Forest*. Hasil penelitian ini memperoleh tingkat ketepatan akurasi, *precision*, *recall*, dan *F1-score* tertinggi pada metode pengklasifikasian *Random Forest* berturut-turut adalah 76,192%, 72%, 76%, dan 73%.

Kata Kunci: Kualitas Udara, *K-Nearest Neighbor*, SMOTE, *Random Forest*

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN	ii
SURAT PENYATAAN KEASLIAN KARYA ILMIAH	iii
LEMBAR PERSEMBAHAN	iiiv
KATA PENGANTAR.....	iv
ABSTRACT	viii
ABSTRAK	ix
DAFTAR ISI.....	x
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah.....	4
1.3 Batasan Masalah.....	5
1.4 Tujuan.....	5
1.5 Manfaat	5
BAB II TINJAUAN PUSTAKA	6
2.1 Kualitas Udara.....	6
2.2 Indeks Kualitas Udara	8
2.3 K-Nearest Neighbor (K-NN).....	8
2.4 Diskritisasi Data	9
2.5 <i>Synthetic Minority Oversampling Technique (SMOTE)</i>	11
2.5 <i>Random Forest</i>	11
2.5.1 <i>Confusion Matrix</i>	14
2.6 Penelitian Relevan	14
BAB III METODOLOGI PENELITIAN	16
3.1 Tempat	16
3.2 Waktu	16
3.3 Alat.....	16
3.4 Metode Penelitian	16

3.5	Diagram Alur	18
BAB IV HASIL DAN PEMBAHASAN.....		19
4.1	Deskripsi Data.....	19
4.2	K-Nearest Neighbor (KNN).....	20
4.2.1	Standarisasi Data	21
4.2.2	Perhitungan Jarak <i>Euclidean</i>	22
4.2.3	Mengurutkan Data Jarak <i>Euclidean</i>	23
4.3	Diskritisasi Data	24
4.4	Partisi Data	26
4.5	<i>Synthetic Minority Oversampling Technique</i> (SMOTE).....	26
4.6	Klasifikasi dengan <i>Random Forest</i>	29
4.6.1	Menentukan <i>ntree</i>	29
4.6.2	<i>Bootstrap Sampling</i>	30
4.6.3	<i>Confusion Matrix</i> Metode <i>Random Forest</i>	37
4.7	Analisis Hasil.....	48
BAB V KESIMPULAN DAN SARAN		49
5.1	Kesimpulan	49
5.2	Saran.....	49
DAFTAR PUSTAKA.....		50
LAMPIRAN.....		53

DAFTAR TABEL

Tabel 2.1 Confusion Matrix.....	13
Tabel 4.2 Jumlah Data Hilang.....	20
Tabel 4.3 Data Contoh Metode K-Nearest Neighbor	21
Tabel 4.4 Standarisasi Data Kualitas Udara.....	22
Tabel 4.5 Jarak Euclidean Kualitas Udara.....	23
Tabel 4.6 Penentuan Jarak Euclidean Kategori Kualitas Udara.....	24
Tabel 4.7 Data Diskritisasi Kualitas Udara	24
Tabel 4.8 Data Penelitian Kualitas Udara.....	25
Tabel 4.9 Data Latih Contoh Metode SMOTE.....	26
Tabel 4.10 Hasil Distribusi Data Latih Contoh Kualitas Udara	27
Tabel 4.11 Data Latih Baru Kualitas Udara	28
Tabel 4.12 Hasil Distribusi Data Latih Kualitas Udara.....	29
Tabel 4.13 Data Bootstrap Sampling untuk Pohon Pertama	30
Tabel 4.14 Jumlah Data pada Setiap Variabel.....	31
Tabel 4.15 Hasil Perhitungan Entropy dan Gain untuk Root Node.....	33
Tabel 4.16 Hasil Perhitungan Entropy dan Gain untuk Node 1.1	34
Tabel 4.17 Klasifikasi Akhir dengan ntree=5.....	38
Tabel 4.18 Klasifikasi Akhir dengan ntree=10.....	39
Tabel 4.19 Klasifikasi Akhir dengan ntree=25.....	40
Tabel 4.20 Klasifikasi Akhir dengan ntree=50.....	41
Tabel 4.21 Klasifikasi Akhir dengan ntree=100.....	42
Tabel 4.22 Klasifikasi Akhir dengan ntree=150.....	43
Tabel 4.23 Klasifikasi Akhir dengan ntree=200.....	44
Tabel 4.24 Klasifikasi Akhir dengan ntree=500.....	45
Tabel 4.25 Klasifikasi Akhir dengan ntree=750.....	46
Tabel 4.26 Klasifikasi Akhir dengan ntree=1000.....	47
Tabel 4.27 Perbandingan Akurasi Berdasarkan ntree.....	48

DAFTAR GAMBAR

Gambar 4.1 Pohon Keputusan Root Node.....	34
Gambar 4.2 Pohon Keputusan Node 1.1	35
Gambar 4.3 Cabang Pohon Keputusan Teratas.....	36
Gambar 4.4 Confusion Matrix ntree=5.....	37
Gambar 4.5 Confusion Matrix ntree=10.....	39
Gambar 4.6 Confusion Matrix ntree=25.....	40
Gambar 4.7 Confusion Matrix ntree=50.....	41
Gambar 4.8 Confusion Matrix ntree=100.....	42
Gambar 4.9 Confusion Matrix ntree=150.....	43
Gambar 4.10 Confusion Matrix ntree=200.....	44
Gambar 4.11 Confusion Matrix ntree=500	45
Gambar 4.12 Confusion Matrix ntree=750.....	46
Gambar 4.13 Confusion Matrix ntree=1000.....	47

BAB I

PENDAHULUAN

1.1 Latar Belakang

Metode *random forest* merupakan salah satu metode statistik yang biasa digunakan untuk regresi dan klasifikasi (Primajaya & Sari, 2018). Menurut Breiman (Kulkarni & Sinha, 2014), *Random Forest* bekerja dengan baik pada database yang besar, karena dapat menangani ribuan variabel masukan tanpa menghapus variabel apapun, dapat memberikan estimasi variabel-variabel yang penting, dapat menghasilkan estimasi internal yang tidak bias terhadap kesalahan generalisasi seiring dengan bertambahnya jumlah hutan sebagai kelompok-kelompok pohon keputusan, memiliki teknik-teknik yang efektif dalam mengestimasi data yang hilang dan mempertahankan keakuratannya saat sebagian besar data hilang, serta memiliki metode-metode untuk menyeimbangkan *class error* pada kelas-kelas yang memiliki jumlah kelas yang tidak seimbang.

Data hilang yang tidak ditangani dapat menyebabkan hasil analisis dan kesimpulan yang salah (Chandra et al., 2023). Regresi *K-Nearest Neighbor* (K-NN) adalah salah satu metode imputasi ganda untuk penanganan nilai hilang dengan memanfaatkan informasi tambahan yang diberikan oleh masing-masing prediktor untuk mempertahankan struktur asli data (Huang et al., 2017). Beberapa penelitian terdahulu menunjukkan bahwa penerapan metode regresi K-NN dalam menangani data hilang dapat meningkatkan kinerja metode klasifikasi (Huang et al., 2017).

Ketidaksimbangan kelas merupakan dataset kelas data satu dengan dataset kelas yang lain tidak berimbang. Kelas yang didalamnya mempunyai data paling banyak

disebut dengan kelas mayoritas, sedangkan kelas yang didalamnya mempunyai data yang lebih sedikit disebut dengan kelas minoritas (Hapsari, R, K. Indriyani, 2022). Ketidakseimbangan kelas yang tidak diatasi dapat menyebabkan prediksi yang tidak akurat mengenai kelas minoritas (Chandra et al., 2023). Masalah ketidakseimbangan kelas dapat ditangani dengan menggunakan pendekatan tingkat data dan tingkat algoritma seperti oversampling (Upadhyay & Kaur, 2021). *The Synthetic Minority Oversampling Technique* (SMOTE) merupakan suatu metode yang populer untuk mengatasi kelemahan yang ada pada oversampling. Untuk mengatasi overfitting, SMOTE mempertimbangkan tetangga terdekat saat membuat data sintesis, sehingga hasilnya batas keputusan yang lebih baik dapat melihat contoh dari kelas minoritas yang saling berdekatan (Novitasari et al., 2023).

Polusi udara menjadi salah satu tantangan lingkungan yang dihadapi oleh negara-negara di seluruh dunia termasuk India. India merupakan salah satu negara yang mempunyai posisi sebagai suatu negara dengan tingkat polusi paling parah di dunia. Berdasarkan data yang didapat dari CNN Indonesia yang berisi daftar kota dengan polusi udara terburuk di dunia pada tahun 2018, 7 kota di India masuk pada posisi 10 pada kategori wilayah kualitas udara terburuk di dunia (Trianisa et al., 2020). Oleh karena itu, penelitian ini dapat dijadikan rujukan bagi negara Indonesia agar dapat terhindar dari polusi udara yang memburuk.

Penelitian ini mengklasifikasi kualitas udara dan mengintegrasikan penanganan nilai yang hilang serta kelas yang tidak seimbang pada dataset AQI India. Pengklasifikasian dilakukan dengan menggunakan metode *Random Forest*, penanganan nilai yang hilang ditangani dengan menggunakan regresi K-NN yang

biasanya digunakan untuk data yang besar dan dapat meningkatkan kinerja metode klasifikasi, sedangkan ketidakseimbangan kelas akan menggunakan metode SMOTE dengan melakukan *oversampling* untuk mensintesis sampel baru dan menghindari *overfitting* yang dapat meningkatkan jumlah kelas minoritas dan membantu meningkatkan pengklasifikasian. Data yang memiliki nilai hilang dapat menyebabkan hilangnya informasi pada dataset, dan data yang memiliki kelas tidak seimbang dapat menyebabkan informasi pada kelas mayoritas menjadi mudah diperoleh (Chandra et al., 2023)

Penelitian terdahulu dilakukan oleh Nugroho *et al* (2023) yang melakukan penelitian tentang Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan *Open Government Data* Menggunakan Algoritma *Random Forest* bertujuan untuk mencari hasil terbaik. Hasil dari klasifikasi berdasarkan data Indeks Pencemaran Udara (ISPU) memiliki performa yang baik, dengan kategori Sedang memiliki nilai precision 90,9%, recall 89,28% f1-score 90,09%, dan kategori Tidak Sehat memiliki nilai precision 89,09%, recall 90,74%, dan f1-score 89,9%

Selanjutnya, penelitian yang dilakukan oleh Alita dan Rahman (2020) penelitian yang dilakukan dengan menggunakan metode *Random Forest Classifier* untuk mendeteksi sarkasme pada proses analisis sentimen. Hasil penelitian menunjukkan bahwa pendeteksian sarkasme pada analisis sentimen tersebut dapat meningkatkan rata-rata dengan kenaikan dari akurasi sebesar 16,61%, presisi sebesar 5,45%, *recall* sebesar 9,64%, dan *f1-score* sebesar 11,27%.

Penelitian lain oleh Amaliah *et al* (2022) melakukan penelitian untuk mengklasifikasi varian minuman kopi di kedai kopi konijiwa bantaeng dengan

menggunakan metode *Random Forest*. Mendapat hasil bahwa penggunaan metode *random forest* pada data penjualan minuman di kedai kopi konijiwa bantaeng dengan varian kopi dikatakan akurat berdasarkan nilai akurasi sebesar 94,12% yang termasuk tinggi.

Penelitian-penelitian tersebut menunjukkan bahwa metode *Random Forest* mendapatkan hasil klasifikasi dengan tingkat ketepatan dan nilai akurasi yang tinggi. Penelitian-penelitian tersebut tentu memiliki perbedaan dengan penelitian yang dilakukan peneliti saat ini yaitu perbedaan pada lokasi yang diteliti, hal apa yang diteliti, dan metode-metode yang digunakan untuk mengisi data hilang dan kelas tidak seimbang, selain itu metode *random forest* juga dipengaruhi oleh jumlah pohon yang digunakan (*ntree*) dalam proses prediksinya.

Berdasarkan uraian permasalahan penelitian terdahulu dan permasalahan mengenai kajian tersebut, maka peneliti tertarik untuk melakukan penelitian mengenai klasifikasi kualitas udara dengan menggunakan metode *Random Forest* yang berjudul **“Klasifikasi Kualitas Udara Menggunakan *Random Forest* dengan Penerapan Regresi KNN dan SMOTE untuk mengatasi Data Hilang dan Kelas Tidak Seimbang”**. Pengklasifikasian dilakukan menggunakan beberapa pohon yang berbeda. Hasil klasifikasi yang diperoleh dengan menggunakan metode *Random Forest* pada penelitian ini dibatasi oleh nilai *accuracy*, *precision*, *recall*, dan *f1-score* untuk menentukan ketepatan klasifikasi kualitas udara.

1.2 Perumusan Masalah

Perumusan masalah pada penelitian ini adalah bagaimana tingkat ketepatan pengklasifikasian kualitas udara menggunakan metode *Random Forest* dengan

berbagai jumlah pohon yang berbeda dan yang mana memberikan tingkat ketepatan terbaik?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Data yang digunakan merupakan data kualitas udara 26 kota di India pada tahun 2015 – 2020.
2. Pada partisi data akan dibagi menjadi rentang tahun 2015 – 2019 untuk data latih dengan persentase 82,07% dan tahun 2020 untuk data uji dengan persentase 17,92%.
3. Banyaknya pohon yang digunakan pada metode random forest adalah 5, 10, 25, 50, 100, 150, 200, 500, 750, dan 1000.

1.4 Tujuan

Tujuan dari penelitian ini adalah memperoleh memperoleh tingkat ketepatan klasifikasi kualitas udara menggunakan metode *Random Forest* berdasarkan *accuracy*, *precision*, *recall*, dan *f1- score* untuk berbagai jumlah pohon yang berbeda dan mengetahui jumlah pohon yang memberikan tingkat ketepatan terbaik.

1.5 Manfaat

Manfaat dari penelitian ini adalah dapat memberikan referensi tentang pengklasifikasian untuk menentukan tingkat ketepatan dalam mengklasifikasikan kualitas udara dengan menggunakan metode *Random Forest*.

DAFTAR PUSTAKA

- Agista, P., Gusdini, N., & Maharani, M. (2020). Analisis Kualitas Udara dengan Indeks Standar Pencemar Udara (ISPU) Dan Sebaran Kadar Polutannya Di Provinsi Dki Jakarta. *Sustainable Environmental and Optimizing Industry Journal*, 2(2), 39–57. Tersedia pada <https://doi.org/10.36441/seoi.v2i2.491>. Diakses pada 12 Desember 2023.
- Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). Prediksi Kualitas Udara Menggunakan Algoritma K- Nearest Neighbor. *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 7(2), 496–507.
- Amna, et al. (2022). Data Mining. PT Global Eksekutif Teknologi. Tersedia pada <http://www.globaleksekutifteknologi.co.id>. Diakses pada 18 Maret 2025.
- Chandra, W., Suprihatin, B., & Resti, Y. (2023). Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction. *Symmetry*, 15(4), 1 – 16. Tersedia pada <https://doi.org/10.3390/sym15040887>. Diakses pada 10 Desember 2023.
- Cornelius, K., & Shanthini, B. (2023). Air Quality Data Analysis And Prediction Using Modified Differential Evolution - Random Forest Algorithm. *Journal of Survey in Fisheries Sciences*, 10(1), 3067–3078.
- Gaikaar, D., Patel, U., Vispute, O., Singh, S., & Sanghvi, T. (2023). Prediction of Air Quality Index using Random Forest Algorithm. *International Reseach Journal of Engineering and Technology (IRJET)*, 10(4) 1248–1252.
- Hapsari, R, K. Indriyani, T. (2022). Implementasi Algoritma SMOTE Sebagai Penyelesaian Imbalance Hight Dimensional Datasets. *Prosiding Seminar Nasional Teknik Elektro, Sistem Informasi, Dan Teknik Informatika*. Surabaya.
- Hasna, H., Amalita, N., Permana, D., & Salma, A. (2024). Random Forest Implementation for Air Pollution Standard Index Classification in DKI Jakarta 2022. *UNP Journal of Statistics and Data Science*. 2(2), 226–233.
- Huang, J., Keung, J. W., Sarro, F., Li, Y. F., Yu, Y. T., Chan, W. K., & Sun, H. (2017). Cross-Validation Based K Nearest Neighbor Imputation For Software Quality Datasets: An Empirical Study. *Journal of Systems and Software*, 132. 226–252. Terdapat pada <https://doi.org/10.1016/j.jss.2017.07.012>. Diakses pada 25 November 2023.
- Kulkarni, V. Y., & Sinha, P. K. (2014). Effective Learning and Classification using Random Forest Algoritm. *International Journal of Engineering and*

Innovative Technology (IJEIT), 3(11), 267–273.

- Kullarni, V. Y., & Sinha, P. K. (2013). Random Forest Classifier: A Survey and Future Research Directions. *International Journal of Advanced Computing*, 36(1), 1144–1156.
- Lin, T. H., & Jiang, J. R. (2021). Credit Card Fraud Detection With Autoencoder and Probabilistic Random Forest. *Mathematics*, 9(21), 4–15. Terdapat pada <https://doi.org/10.3390/math9212683>. Diakses pada 2 Desember 2023.
- Mirqotussa'adah, Muslim, M. A., Sugiharti, E., Prasetyo, B., & Alimah, S. (2017). Penerapan *Dizcresization* dan Teknik *Bagging* untuk Meningkatkan Akurasi Klasifikasi Berbasis *Ensemble* pada Algoritma C4.5 dalam Mengdiagnosa Diabetes. *Lontar Komputer*, 8(2), 135–145.
- Nababan, A. A., Jannah, M., Aulina, M., & Andrian, D. (2023). Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (SMOTE) Berdasarkan Indeks Standar Pencemaran Udara (ISPU). *JTIK (Jurnal Teknik Informatika Kaputama)*, 7(1), 214–219. Terdapat pada <https://doi.org/10.59697/jtik.v7i1.66>. Diakses pada 5 Januari 2024.
- Novitasari, F., Haerani, E., Nazir, A., Jasril, J., & Insani, F. (2023). Sistem Klasifikasi Penyakit Jantung Menggunakan Teknik Pendekatan SMOTE Pada Algoritma Modified K-Nearest Neighbor. *Building of Informatics, Technology and Science (BITS)*, 5(1), 274–284. Terdapat pada <https://doi.org/10.47065/bits.v5i1.3610>. Diakses pada 19 Desember 2023.
- Prayudha, J., Pranata, A., & Al Hafiz, A. (2018). Implementasi Metode Fuzzy Logic Untuk Sistem Pengukuran Kualitas Udara Di Kota Medan Berbasis Internet of Things (Iot). *Jurteksi*, 4(2), 141–148. Terdapat pada <https://doi.org/10.33330/jurteksi.v4i2.57>. Diakses pada 21 Desember 2023.
- Primajaya, A., Sari, B. N., Studi, P., Informatika, T., Komputer, F. I., Karawang, U. S., Info, A., & Forest, R. (2018). *Random Forest Algorithm for Prediction of Precipitation*. 1(1), 27–31.
- Rahman, B., Fauzi, F., & Amri, S. (2023). Perbandingan Hasil Klasifikasi Data Iris menggunakan Algoritma K-Nearest Neighbor dan Random Forest. *Journal Of Data Insights*, 1(1), 19–26. Terdapat pada <https://doi.org/10.26714/jodi.v1i1.135>. Diakses pada 5 Januari 2024.
- Shah, D. P., & Patel, D. P. (2021). A Comparison Between National Air Quality Index, India and Composite Air Quality Index for Ahmedabad, India. *Environmental Challenges*, 5(8), 1–25. Terdapat pada <https://doi.org/10.1016/j.envc.2021.100356>. Diakses pada 23 Desember 2023.

- Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, 4(3), 121–127. Terdapat pada <https://doi.org/10.35580/variansium31>. Diakses pada 13 Desember 2023.
- Susanti, Martha, S., & Sulistianingsih, E. (2018). K Nearest Neighbor Dalam Imputasi Missing Data. *Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster)*, 07(1), 9–14. Terdapat pada <http://archive.ics.uci.edu/ml/datas/Iris>. Diakses pada 5 Januari 2024.
- Susanto, J. P. (2005). Kualitas Udara Beberapa Kota di Asia (Monitoring Kandungan SO₂ Udara Ambien dengan Passive Sampler). *J.Tek.Ling*, 6(1), 324–329.
- Trianisa, K., Purnomo, E. P., & Kasiwi, A. N. (2020). Pengaruh Industri Batubara Terhadap Polusi Udara dalam Keseimbangan World Air Quality Index in India. *Jurnal Sains Teknologi & Lingkungan*, 6(2), 156–168.
- Upadhyay, K., & Kaur, P. (09-10 April 2021). *A Review on Data level Approaches to address the Class Imbalance Problem*. International Conference on Recent Challenges in Engineering Science and Technology, India.
- Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. A. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors (Switzerland)*, 16(1), 1–18. Terdapat pada <https://doi.org/10.3390/s16010086>. Diakses pada 15 Desember 2023.