https://doi.org/10.1007/s40747-021-00435-5

ARTIKEL ASLI

Algoritma klasifikasi SVM berbasis kernel yang dapat diskalakan pada data kualitas udara yang tidak seimbang untuk perawatan kesehatan yang mahir

Shwet Ketu[1] **- Pramod Kumar Mishra**[1]

Diterima: 9 Desember 2020 / Diterima: 9 Juni 2021

Abstrak
Dalam dekade terakhir, kita telah melihat perubahan drastis pada tingkat polusi udara, yang telah menjadi masalah lingkungan yang kritis. Hal ini harus ditangani dengan hati-hati untuk menghasilkan solusi perawatan kesehatan yang baik. Mengurangi dampak polusi udara terhadap kesehatan manusia hanya mungkin dilakukan jika data diklasifikasikan dengan benar. Dalam berbagai masalah klasifikasi, kita menghadapi masalah ketidakseimbangan kelas. Belajar dari data yang tidak seimbang selalu menjadi tugas yang menantang bagi para peneliti, dan dari waktu ke waktu, solusi yang memungkinkan telah dikembangkan oleh para peneliti. Dalam makalah ini, kami berfokus untuk menangani distribusi kelas yang tidak seimbang dengan cara agar algoritma klasifikasi tidak akan mengganggu kinerjanya. Algoritma yang diusulkan didasarkan pada konsep metode penskalaan kernel yang disesuaikan (AKS) untuk menangani dataset yang tidak seimbang dengan banyak kelas. Pemilihan fungsi kernel telah dievaluasi dengan bantuan kriteria pembobotan dan uji chi-square. Semua evaluasi eksperimental telah dilakukan pada dataset Central Pollution Control Board (CPCB) India yang berbasis sensor. Algoritma yang diusulkan dengan akurasi tertinggi sebesar 99,66% memenangkan perlombaan di antara semua algoritma klasifikasi yaitu Adaboost (59,72%), Multi-Layer Per-ceptron (95,71%), GaussianNB (80,87%), dan SVM (96,92). Hasil dari algoritma yang diusulkan juga lebih baik daripada metode literatur yang ada. Dari hasil ini juga terlihat jelas bahwa algoritma yang kami usulkan efisien untuk menangani masalah ketidakseimbangan kelas dengan kinerja yang lebih baik. Dengan demikian, klasifikasi kualitas udara yang akurat melalui algoritma yang kami usulkan akan berguna untuk meningkatkan kebijakan pencegahan yang ada dan juga akan membantu meningkatkan kemampuan tanggap darurat yang efektif dalam situasi polusi terburuk.

Kata kunci Kualitas udara - Klasifikasi - Perawatan kesehatan yang mahir - SVM berbasis kernel yang dapat diskalakan - Data yang tidak seimbang

Pendahuluan

Dalam paradigma pembelajaran mesin, klasifikasi objek baru berdasarkan contoh yang serupa adalah salah satu tugas penting. Tugas klasifikasi menjadi lebih rumit ketika salah satu kelas berisi lebih sedikit contoh daripada kelas lainnya [1]. Masalah ketidakseimbangan kelas tidak lain adalah distribusi data yang tidak merata di antara berbagai kelas. Dalam masalah ketidakseimbangan kelas, sebagian besar sampel data termasuk dalam kelas tertentu, dan sampel data lainnya termasuk dalam kelas lainnya. Sehubungan dengan masalah ketidakseimbangan kelas biner, satu kelas berisi

✉ Shwet Ketu shwetiiita@gmail.com
Pramod Kumar Mishra
mishra@bhu.ac.in

[1] Departemen Ilmu Komputer, Institut Sains, Banaras Hindu University, Varanasi, India

Published online: 29 June 2021

jumlah sampel data maksimum, dan kelas lainnya hanya berisi sedikit sampel data [2]. Kelas yang berisi jumlah sampel maksimum dikatakan sebagai kelas mayoritas, dan kelas dengan jumlah sampel minimal dikatakan sebagai kelas minoritas [3, 4].

Dalam bidang pembelajaran mesin, merupakan salah satu tugas yang menantang bagi algoritma klasifikasi untuk belajar dari data yang tidak seimbang. Kita menghadapi masalah ketidakseimbangan data di hampir semua domain, atau dapat dikatakan bahwa ini adalah masalah yang cukup umum di semua bidang. Bidang-bidang yang menghadapi masalah ini adalah domain medis [5, 6], domain pemasaran, klasifikasi gambar [7], pertanian, domain big data [8-10], IoT [11-13], dan seterusnya [14-16]. Ketidakseimbangan kelas adalah salah satu masalah penting dalam paradigma pembelajaran mesin. Jika algoritma klasifikasi bias terhadap kelas mayoritas, maka akurasi algoritma klasifikasi akan sangat menurun. Dengan demikian, jika sampel baru akan datang untuk klasifikasi, maka sampel tersebut akan diklasifikasikan ke dalam kelas mayoritas karena pengklasifikasi memiliki akurasi prediksi yang lebih rendah terhadap kelas minoritas.

kelas. Situasi ini sangat tidak pantas dan sangat memprihatinkan [17].

Saat ini, telah terjadi perubahan drastis pada tingkat polusi udara [18]. Tingkat polusi di kota-kota metropolitan semakin meningkat, dan ini bukanlah pertanda yang baik bagi kita. Untuk membuat lingkungan menjadi lebih sehat dan nyaman, tingkat polusi udara haruslah seminimal mungkin. Ada berbagai faktor penyebab yang membuat udara menjadi tercemar [19-22]. Beberapa di antaranya ada yang secara langsung dan ada pula yang secara tidak langsung turut mencemari udara. Polutan tersebut berasal dari berbagai macam sumber seperti dari industri, jasa transportasi, lalu lintas harian, pembangkit listrik tenaga panas, berbagai macam peralatan rumah tangga, sampah dari industri, rumah sakit, rumah, dan lain sebagainya. Tingginya tingkat polusi udara dapat membahayakan manusia, hewan, dan juga tumbuhan [23]. Secara berurutan, banyak kasus baru yang berkaitan dengan sistem pernapasan telah terlihat, yang merupakan dampak dari kualitas udara yang buruk pada manusia. Hal ini juga mempengaruhi kualitas tanaman dan produksi tanaman secara keseluruhan. Oleh karena itu, untuk mengurangi dampak polusi udara, kita harus mengklasifikasikan tingkat polusi dengan benar secara real-time. Dari waktu ke waktu, banyak peneliti telah menyumbangkan pendekatan mereka, yang akurat sampai batas tertentu [24-28]. Namun karena sifat data yang tidak seimbang, model-model ini tidak memberikan prediksi yang benar dari kelas-kelas tersebut [29-32].

Membangun pengklasifikasi menggunakan set data yang tidak seimbang adalah salah satu tugas yang sulit. Dalam tugas klasifikasi dataset yang tidak seimbang, kelas minoritas selalu menderita dari kelas mayoritas karena model klasifikasi bias dengan kelas mayoritas [33, 34]. Akibatnya, jika ada sampel baru yang datang untuk klasifikasi, maka sampel tersebut akan diklasifikasikan dalam kelas mayoritas. Kebutuhan yang mendesak dan minat yang sangat besar ini memotivasi para peneliti untuk menangani masalah ketidakseimbangan kelas. Dari waktu ke waktu, banyak peneliti telah memberikan solusi yang bernilai untuk menangani masalah ketidakseimbangan kelas ini. Pendekatan-pendekatan ini bermanfaat dan mampu memecahkan masalah sampai batas tertentu dengan meningkatkan kinerja para pengklasifikasi. Sebagian besar solusi diusulkan untuk masalah ketidakseimbangan kelas biner, yang tidak cocok untuk masalah ketidakseimbangan multi-kelas. Keterbatasan ini memotivasi kami untuk menangani masalah ketidakseimbangan multi-kelas dan juga mendorong kami untuk memberikan kontribusi yang dapat menyelesaikan masalah ketidakseimbangan multi-kelas. Kontribusi yang telah kami kerjakan adalah:

- Solusi ini dirancang sedemikian rupa, yang cocok untuk masalah ketidakseimbangan kelas biner dan multi-kelas.
- Solusi ini didasarkan pada modifikasi algoritmik daripada resampling data pada fase pemrosesan.
- Dalam solusi kami, fungsi pemilihan kernel yang baru telah diusulkan.

Dalam makalah ini, algoritma klasifikasi SVM (Support Vector Machine) berbasis kernel yang dapat diskalakan telah diusulkan, yang mampu menangani masalah ketidakseimbangan data multi-kelas. Pertama-tama, perkiraan hyperplane diperoleh dengan menggunakan algoritma SVM standar. Setelah itu, faktor pembobotan dan fungsi parameter untuk setiap vektor pendukung pada setiap iterasi dihitung. Nilai-nilai parameter ini dihitung menggunakan uji Chi-square. Setelah itu, fungsi kernel baru atau fungsi transformasi kernel dihitung. Dengan bantuan fungsi transformasi kernel ini, batas-batas kelas yang tidak rata telah diperluas, dan kemencengan data telah dikompensasi. Dengan demikian, perkiraan hyperplane dapat dikoreksi oleh algoritma yang diusulkan, dan juga dapat menyelesaikan masalah penurunan kinerja. Dalam penelitian ini, kami juga telah membahas dampak polusi udara terhadap kesehatan manusia.

Sisa dari makalah ini telah disusun sebagai berikut. Penelitian terkait telah digambarkan dalam "Penelitian terkait". Diskusi singkat tentang dataset, cara kerja algoritma yang diusulkan dengan landasan matematika, dan sepuluh metrik evaluasi kinerja telah diilustrasikan secara singkat dalam "Bahan dan metode". Hasil dari metode standar, metode literatur yang ada, dan algoritma klasifikasi yang diusulkan telah dipaparkan dalam "Hasil". Dalam "Diskusi", pembahasan komprehensif mengenai hasil klasifikasi dan pengaruh kualitas udara yang buruk terhadap kesehatan telah dibahas. Kesimpulan dengan ruang lingkup masa depan telah ditarik dalam "Kesimpulan".

Pekerjaan terkait

Membangun pengklasifikasi menggunakan dataset yang tidak seimbang adalah salah satu tugas yang sulit. Dalam tugas klasifikasi dataset yang tidak seimbang, kelas minoritas selalu menderita dari kelas mayoritas karena model klasifikasi bias dengan kelas mayoritas [33, 34]. Akibatnya, jika ada data baru yang masuk untuk klasifikasi, maka data tersebut akan diklasifikasikan dalam kelas mayoritas. Dari waktu ke waktu, banyak strategi yang telah dibuat untuk mengatasi masalah ketidakseimbangan kelas. Strategi yang diusulkan ini bekerja pada tingkat algoritma atau pada tingkat data.

Pendekatan tingkat data didasarkan pada teknik resampling. Banyak algoritma klasifikasi seperti SVM, naïve Bayes, C4.5, AdaBoost, dan sebagainya menggunakan teknik resampling untuk menangani masalah ketidakseimbangan data. Tugas resampling terdiri dari dua sub-tugas, yaitu under-sampling dan over-sampling [35, 36]. Teknik under-sampling adalah proses menyaring sampel yang tidak relevan dari kumpulan data, dan dalam teknik oversampling, kami menghasilkan data sintetis yang baru. Dua metode under-sampling yang efektif telah diusulkan oleh Liu dkk. [37], yaitu, BalanceCascade dan EasyEnsemble. Dalam BalanceCascade

Tabel 1 Algoritme klasifikasi untuk menangani masalah ketidakseimbangan data

Penulis	Pendekatan	Tujuan	Algoritma	Hasil	Cakupan
Liu dkk. [37]	Pendekatan tingkat data	Usulan dua metode pengambilan sampel di bawah metode Balance-Cascade dan Merakit dengan Mudah	MudahMerakit dan MenyeimbangkanKaskade	Menangani ketidakseimbangan data	Digunakan untuk pendekatan tingkat data dan untuk menyelesaikan ketidakseimbangan data. masalah-masalah yang terjadi
Wang dkk. [38]	Pendekatan tingkat data	Pendekatan pengambilan sampel berlebih yang adaptif telah diusulkan	Pendekatan kepadatan data	Berurusan dengan ketidakseimbangan data	Digunakan untuk menyelesaikan masalah ketidakseimbangan data
Geo dkk. [39]	Pendekatan tingkat data	Pengambilan sampel berlebih kelas biner telah dilakukan diusulkan	Menggunakan metode probabilistik	Berurusan dengan ketidakseimbangan data	Digunakan untuk menyelesaikan masalah ketidakseimbangan data
Batuwita dan Palade [44]	Tingkat algoritmik	Data tidak seimbang dalam adanya kebisingan	SVM berbasis fuzzy	Menghapus data yang tidak seimbang ance	Pengoptimalan pengklasifikasi
Cano dkk. [45]	Tingkat algoritmik	Data yang diusulkan tidak seimbang pengklasifikasi yang sudah ada sebelumnya	Berat gravitasi-berbasis	Menghapus data yang tidak seimbang ance	Pengoptimalan pengklasifikasi
Wu dan Chang [46, 47]	Tingkat algoritmik	Usulan batas kelas berbasis batas yang diusulkan penyelarasan	SVM yang ditingkatkan	Menghapus ketidakseimbangan data	Pengoptimalan pengklasifikasi
Oh dkk. [48]	Tingkat algoritmik	Usulan teknik pemilihan sampel aktif untuk data masalah ketidakseimbangan	Pemilihan sampel aktif	Mengatasi masalah ketidakseimban gan data dengan meningkatkan kinerja mance	Meningkatkan akurasi pengklasifikasi
Liu dkk. [49]	Tingkat algoritmik	Mengusulkan teknik pemilihan sampel	SVM	Meningkatkan kinerja dari pengklasifikasi	Meningkatkan akurasi pengklasifikasi
Fu dan lee [51]	Tingkat algoritmik	Mengusulkan algoritma pembelajaran aktif berbasis kepastian	Pembelajaran mesin	Mengatasi ketidakseimbangan data dan meningkatkan kinerja	Pendekatan pembelajaran aktif

teknik ini, sampel yang diklasifikasikan dengan benar pada setiap langkah akan dihapus dan tidak berpartisipasi dalam tugas klasifikasi lebih lanjut. Dalam metode EasyEnsemble, kelas mayoritas dibagi menjadi beberapa subset. Subset-subset ini digunakan sebagai masukan untuk pelajar. SMOTE adalah singkatan dari Synthetic Minority Over-Sampling Technique. Ini adalah salah satu teknik cerdas yang didasarkan pada pendekatan oversampling [36]. Oversampling dalam SMOTE dilakukan dengan menghasilkan sampel sintaksis untuk kelas minoritas. Metode adaptive oversampling telah diusulkan oleh Wang dkk. [38], yang didasarkan pada pendekatan kepadatan data. Pendekatan oversampling kelas biner telah diusulkan oleh Geo dkk. [39], yang didasarkan pada fungsi kepadatan probabilitas. Gu dkk. [40] telah membahas pendekatan data mining pada dataset yang tidak seimbang.

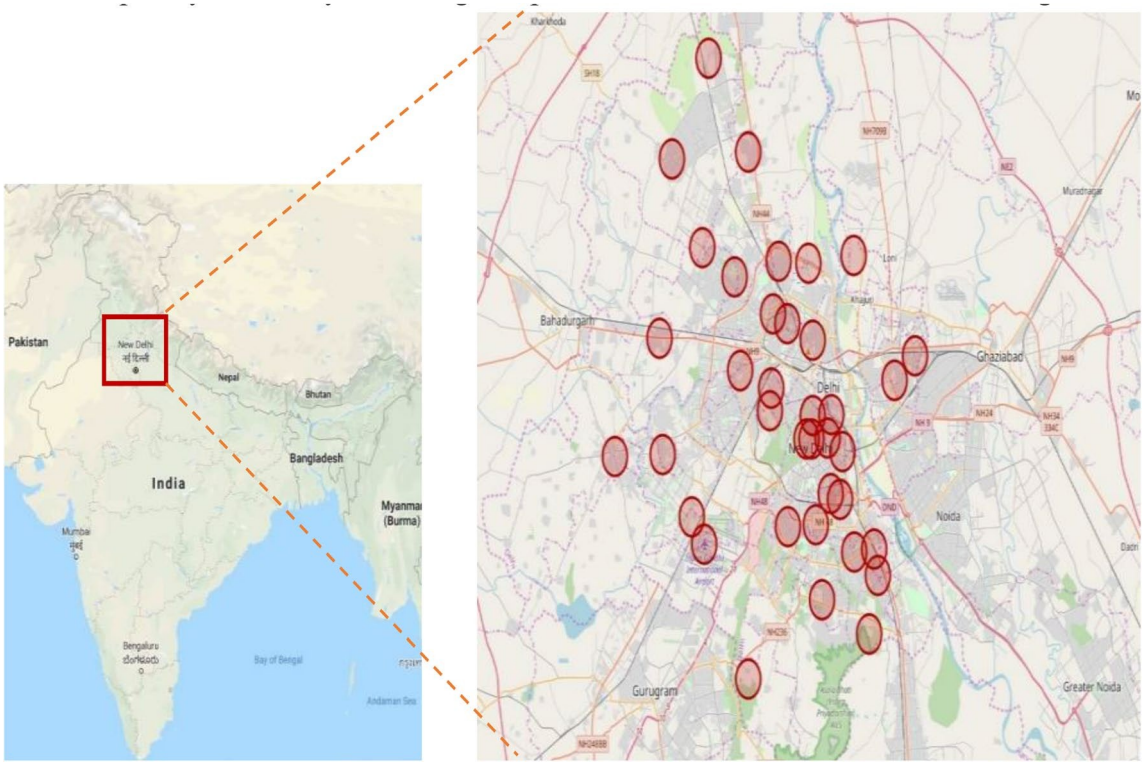
Pendekatan tingkat algoritmik dirancang untuk membiaskan proses pembelajaran untuk mengurangi partisipasi kelas mayoritas dan meningkatkan kinerja pengklasifikasi. Solusi untuk pendekatan tingkat algoritmik terutama terdiri dari modifikasi dalam algoritme, pembelajaran yang peka terhadap biaya, pembelajaran ansambel, dan pembelajaran aktif.

Pendekatan pembelajaran yang peka terhadap biaya

didasarkan pada konsep kebijakan penugasan biaya asimetris dengan meminimalkan
biaya sampel yang salah klasifikasi. Minimalisasi biaya dalam pendekatan yang sensitif terhadap biaya adalah proses menghukum kelas yang salah diklasifikasikan dengan penalti. Tetapi memberikan penalti yang diinginkan pada setiap tingkat kelas adalah tugas yang sulit [41, 42]. Tiga algoritma penguat yang peka terhadap biaya untuk klasifikasi dataset yang tidak seimbang dalam kerangka kerja AdaBoost telah diperkenalkan oleh Sun dkk. [41]. SVM yang sensitif terhadap biaya (mesin vektor suport) telah diusulkan oleh Wang [43] untuk menangani masalah ketidakseimbangan data.

Untuk mengatasi masalah ketidakseimbangan data, beberapa peneliti telah melakukan modifikasi pada tingkat algoritma. Modifikasi pada level algoritma dapat dilakukan pada level pengklasifikasi dengan melakukan optimasi pada pengklasifikasi. SVM berbasis fuzzy diusulkan oleh Batuwita dan Palade [44] untuk menangani data yang tidak seimbang dengan adanya noise dan outlier. Cano dkk. [45] telah mengusulkan klasifikasi data yang tidak seimbang berdasarkan gravitasi berbobot. Penyelarasan batas kelas berbasis batas yang disesuaikan dengan peningkatan kinerja SVM telah diusulkan oleh Wu dan Chang [46, 47].

Pendekatan pembelajaran ensembel dirancang untuk meningkatkan akurasi algoritme klasifikasi. Dalam pendekatan ini, beberapa pengklasifikasi digunakan untuk melatih model, dan keluaran keputusan dari pengklasifikasi ini digabungkan ke dalam satu kelas. Hasil akhir ini digunakan untuk pengambilan keputusan [3]. Bag-
ging dan boosting adalah algoritme pembelajaran mesin yang penting



Gbr. 1 Pusat pengumpulan data kualitas udara di Wilayah Delhi

dalam paradigma pembelajaran ensembel [3]. Teknik pemilihan sampel aktif digunakan oleh Oh dkk. [48] untuk mengatasi masalah ketidakseimbangan data. Teknik pengambilan sampel (baik under-sampling maupun over-sampling) telah diintegrasikan dengan SVM untuk meningkatkan kinerja pengklasifikasi oleh Liu dkk. [49].

Pendekatan pembelajaran aktif adalah salah satu kasus luar biasa dari paradigma pembelajaran mesin yang telah digunakan untuk melabeli titik sampel data baru dengan bantuan output yang diinginkan dengan mendapatkan kueri secara interaktif dengan pengguna [50]. CBAL, yang merupakan algoritma pembelajaran aktif berbasis kepastian, diusulkan oleh Fu dan lee pada tahun 2013 [51] untuk memecahkan masalah ketidakseimbangan data. Berdasarkan berbagai literatur yang ada, algoritma klasifikasi yang digunakan untuk menangani masalah ketidakseimbangan data telah ditunjukkan secara singkat pada Tabel 1.

Bahan dan metode

Pada bagian ini, kami akan membahas tentang materi dan metode yang digunakan dalam analisis eksperimental. Bagian ini terdiri dari tiga subbagian, yaitu ilustrasi dataset, algoritma yang diusulkan, dan ukuran statistik. Pada sub-bagian pertama, dataset CPCB berbasis sensor di Delhi telah dibahas. Pada sub-bagian kedua, algoritma terukur yang diusulkan

Algoritma klasifikasi SVM berbasis kernel telah dibahas dengan landasan matematisnya. Pada sub-bagian ketiga, diskusi singkat tentang metrik evaluasi kinerja telah disajikan.

Data

Untuk penelitian ini, kami telah mengambil data CPCB berbasis sensor dari kota Delhi, yang merupakan kota paling tercemar di India. Alasan di balik pengambilan data tolok ukur ini adalah karena pemantauan kualitas udara secara terus menerus dengan lebih dari 200 stasiun pangkalan di sekitar 20 negara bagian dikelola oleh CPCB (Dewan Pengendalian Polusi Pusat). Semua data dari stasiun-stasiun ini dapat diakses secara terbuka dari situs web CPCB. Sejauh menyangkut Delhi, ada 37 stasiun pangkalan yang memantau data secara terus menerus (24 * 7).

Seperti yang kita ketahui, India berada di urutan kedua dalam hal jumlah populasi setelah Cina [52, 53]. Pertumbuhan populasi yang sangat besar merupakan salah satu alasan utama meningkatnya tingkat polusi. Delhi merupakan ibu kota dan pusat industri India; oleh karena itu, kepadatan penduduk Delhi lebih tinggi dibandingkan dengan kota-kota lainnya. Akibatnya, polusi yang disebabkan oleh limbah industri dan kendaraan menjadi alasan utama meningkatnya tingkat polusi di Delhi [54, 55]. Tingginya pembuangan berbagai gas, yaitu NO2, NH3, NO, CO2, O3,

dan CO, dengan faktor tambahan seperti arah angin, kecepatan angin, suhu, dan kelembaban relatif membuat udara Delhi sangat tercemar dan beracun. Partikel-partikel beracun dan partikel-partikel berbahaya lainnya tidak dapat larut di udara. Dengan demikian, tinggal di lingkungan yang tercemar seperti itu dapat menyebabkan beberapa gangguan kesehatan yang parah. Bahkan kematian juga mungkin terjadi pada kasus-kasus yang lebih parah. Jadi, kita harus mengambil langkah-langkah pencegahan untuk meningkatkan kualitas hidup yang sangat baik dengan mengurangi tingkat polusi untuk kesejahteraan manusia.

Untuk analisis eksperimental, dataset dari Dewan Pengendalian Pencemaran Pusat India (CPCB) di ibukota Delhi telah diambil [56]. Dataset tersebut telah diekstraksi dari berbagai perangkat berbasis sensor. Perangkat berbasis sensor ini telah ditempatkan di berbagai lokasi di Delhi dan telah ditunjukkan pada Gambar 1. Gambar tersebut telah diplot dengan bantuan garis bujur dan garis lintang dari berbagai titik pengumpulan data yang berada di wilayah Delhi. 37 pusat pengumpulan data di Delhi telah diplot dengan lingkaran merah pada Gbr. 1. Kami telah mengambil data dari tanggal 1 Januari 2019 hingga 1 Oktober 2020. Data telah direkam dua puluh empat kali sehari, yang berarti, setiap jam. Kumpulan data kualitas udara CPCB diperkaya dengan berbagai fitur yang dapat dipertanggungjawabkan yang dapat memainkan peran penting dalam tugas klasifikasi kualitas udara. Fitur-fitur yang bertanggung jawab ini adalah PM10 (Konsentrasi Partikel yang Dapat Terhirup), SO_2 (Sulfur Dioksida), PM5 (Materi Partikulat Halus), O_3 (ozon), NOx (Nitrogen Oksida), NO_2 (Nitrogen Dioksida), NO (Nitrogen Monoksida), NH_3 (Amonia), CO (Karbon Monoksida), AQI (Indeks Kualitas Udara), WD (Arah Angin), C H_{66} (Benzena), WS (Kecepatan Angin), RH (Kelembaban Relatif), SR (Radiasi Matahari), BP (Tekanan Bar) dan AT (Suhu Absolut). Dataset yang diambil untuk tugas klasifikasi berisi 16 kolom dan 332.880 baris atau 16 kolom dan 8760 baris di setiap stasiun bumi (37 stasiun bumi dipertimbangkan).

Dalam pekerjaan penelitian ini, tugas klasifikasi telah dilakukan pada dataset kualitas udara BPKB, yang berisi berbagai atribut. Hanya atribut-atribut yang telah dipertimbangkan, yang bertanggung jawab atas tingginya tingkat polusi udara. Atribut-atribut tersebut adalah konsentrasi partikel yang dapat terhirup (PM10), sulfur dioksida (SO_2), partikulat halus (PM2.5), ozon (O_3), nitrogen oksida (NOx), nitrogen dioksida (NO_2), nitrogen monoksida (NO), amonia (NH3), karbon monoksida (CO), Indeks Kualitas Udara (AQI), dan seterusnya. Pada Tabel 2, berbagai fitur dari dataset yang berpartisipasi dalam tugas klasifikasi telah disajikan. Berbagai fitur telah dieksplorasi dengan bantuan beberapa parameter, yaitu

Tabel 2 Fitur-fitur substansial dari dataset sekilas

	Singkatan Variabel	Sifat data	Unit	Periode pengumpulan	Jenis	Sumber
Materi Partikulat10	PM10	Waktu	ug/m	01 Januari 2019 hingga 01	Polutan	BPKB
Materi Partikulat	SO2	Nyata	ug/m^3	Oktober 2020	Polutan	CPCB
Belerang	PM2.5	Real-Time	$_3$	01 Januari 2019 hingga 01	Polutan	CPCB
Dioksida2.5 Ozon	O3	Real-Time	ug/m^3	Oktober 2020	Polutan	CPCB
Nitrogen Oksida	NOx	Real-Time	ug/m^3	01 Januari 2019 hingga 01	Polutan	CPCB
Nitrogen Dioksida	NO2	Real-Time	Ppb	Oktober 2020	Polutan	CPCB
Nitrogen	NO	Real-Time	ug/m^3	01 Januari 2019 hingga 01	Polutan	CPCB
Monoksida	NH3	Real-Time	ug/m^3	Oktober 2020	Polutan	CPCB
Amonia	CO	Real-Time	ug/m^3	01 Januari 2019 hingga 01	Polutan	CPCB
Indeks Kualitas	AQI	Real-Time	ug/m^3	Oktober 2020	Polutan	CPCB

nama variabel dengan singkatannya, sifat data, unit pengukur variabel, periode pengumpulan data, jenis variabel, dan terakhir ekstraksi data. sumber.

Tabel 3 menyajikan berbagai fitur dataset dengan bantuan beberapa parameter, seperti nama variabel dengan nilai rata-rata, unit pengukuran, standar

Variabel

derivasi, dan rentang variabel aktual dan yang ditentukan. Fitur-fitur yang dapat dipertanggungjawabkan ini telah digunakan dalam tugas klasifikasi.

Tabel 4 menunjukkan data yang berasal dari prapemrosesan dan diambil untuk analisis eksperimental. Ini
Data yang telah diproses berisi enam kelas, 270.596 sampel, dan sepuluh atribut di setiap sampel. Distribusi kelas-bijaksana dari dataset ini adalah 13.452, 47.910, 93.167, 55.045, 30.421, dan 30.601 untuk kelas satu hingga kelas enam. Rasio ketidakseimbangan kelas di antara kelas-kelas tersebut adalah 6,92.

Tabel 5 merupakan deskripsi Indeks Kualitas Udara (AQI), yang berisi kisaran AQI, AQI yang sesuai pelabelan, dan tingkat kelas. Pelabelan telah dilakukan menjadi enam
bagian sesuai dengan kisaran dari 0 hingga lebih dari 400 [56].
Tautan dataset CPCB dengan rentang AQI juga ditetapkan di sini.

Metodologi yang diusulkan

Tujuan utama dari algoritma yang diusulkan adalah untuk menangani masalah ketidakseimbangan data secara efisien. Algoritma yang diusulkan didasarkan pada konsep metode penskalaan kernel yang disesuaikan (AKS) [57] untuk menangani dataset yang tidak seimbang multi-kelas. Dalam makalah ini, kami telah mengusulkan klasifikasi SVM yang telah diintegrasikan dengan metode penskalaan kernel yang disesuaikan. Pada bagian ini, diskusi rinci tentang algoritma yang diusulkan telah disajikan.

Algoritma mesin vektor pendukung dasar (SVM)

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang banyak digunakan dan terkenal untuk klasifikasi data. Algoritma SVM telah diusulkan oleh Vapnik dkk. [58] pada tahun 1995. Tujuan utama dari perancangan algoritma ini adalah untuk memetakan data input ke dalam ruang dimensi tinggi dengan bantuan fungsi kernel sehingga kelas-kelasnya dapat secara linear
dapat dipisahkan [58-60]. Dalam kasus masalah kelas biner, margin maksimum yang dapat memisahkan hyperplane disajikan:

$$w.x + b = 0 \quad (1)$$

Berdasarkan pasangan optimal (w_0, b_0), fungsi keputusan untuk SVM diwakili oleh:

$$f(x) = \sum_{j \in SV} \beta_{jyj} \langle x_j, x \rangle + b \quad (2)$$

$$K(x_j, x_i) = \langle x_j, x_i \rangle \quad (3)$$

Pemilihan fungsi kernel

Pada bagian ini, fungsi kernel telah dipilih dari SVM standar untuk menghitung posisi batas. Pada awalnya, dataset P dibagi menjadi beberapa sampel yaitu P^1 , P^2 , P^3 , ... , P^j dan setelah itu, fungsi pembentukan kernel diterapkan yang didefinisikan dalam persamaan di bawah ini.

$$f(x) = \begin{cases} e^{-z_1 h(x)^2}, & jika x \in P1 \\ e^{-z h(x) 2^2}, & jika x \in P \\ \vdots \\ e^{-zCh(x)^2}, & jika x \in PC \end{cases} \quad (4)$$

di mana, $h(x) = \sum_{j \in SV} \beta_{jyj} x_{xj} + b$ (di mana, β_j adalah support vector), P^j adalah sampel ke-j dari training set, nilai parameter z_j dihitung dari uji chi-square (3^2), yang dijelaskan pada Bagian 2.2.2 dan $j = 1, 2, \ldots, C$.

Pengujian Chi-square

Uji Chi-square (3^2) adalah salah satu uji statistik penting yang diterapkan pada set fitur kategorikal untuk menentukan asosiasi berbasis distribusi frekuensi di antara kelompok fitur kategorikal. Dengan kata lain, kita dapat mengatakan bahwa uji ini digunakan untuk mengevaluasi korelasi di antara kelompok-kelompok tersebut. Pentingnya menghitung uji chi-square adalah untuk menentukan hubungan antara sampel dari setiap kategori dan parameter z_j . Formulasi matematis untuk mengevaluasi uji chi-square (3^2) adalah:

$$3^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \quad (5)$$

di mana, f_e dan f_0 dilambangkan sebagai frekuensi yang diharapkan dan
frekuensi yang diamati, masing-masing.

Menghitung faktor pembobotan

Faktor pembobotan adalah salah satu masalah penting dan sulit saat menangani masalah ketidakseimbangan

kelas. Ini adalah sangat sulit karena menetapkan bobot yang sesuai untuk

di mana, β_j adalah vektor pendukung, x_j adalah sampel data dan
$j = 1, 2, \ldots, C$.

Gambar 2 menunjukkan hyperplane dengan margin pemisah maksimum dan vektor pendukung dalam paradigma algoritma SVM.

Untuk ruang fitur dimensi yang lebih tinggi, nilai $\langle x.x_j \rangle$ digantikan oleh fungsi kernel $K\langle x_j.x_i \rangle$ yaitu:

mengatasi masalah ketidakseimbangan kelas menjadi rumit. Cara sederhana untuk mengatasi masalah tersebut adalah dengan memberikan bobot yang lebih kecil kepada kelas mayoritas dan bobot yang lebih besar kepada kelas minoritas dengan memenuhi kondisi bobot $z_i \in (0,1)$.

Perumusan metode pengaturan faktor pembobotan telah digunakan dalam algoritma yang diusulkan untuk menangani

Tabel 3 Deskripsi variabel aktual

Variabel	Rata-rata	Satuan	Std. Dev	Rentang yang ditentukan		Kisaran	
				Min	Max	Min	Max
PM10	208.869	ug/m^3	154.392	0.00	100	0.14	1000
SO2	106.398	ug/m^3	99.803	0.00	80	0.7	989.58
PM2.5	30.339	ug/m^3	55.716	0.00	60	0.01	499.1
O3	51.994	ug/m^3	60.044	0.00	18	0.01	500
NOx	43.873	ppb	33.533	0.00	200	0.01	485.85
NO2	35.515	ug/m^3	20.61	0.00	200	0.01	494.11
TIDAK	14.821	ug/m^3	11.381	0.00	200	0.01	194.9
NH3	1.362	ug/m^3	1.082	0.00	200	0.01	40.25
CO	41.407	ug/m^3	59.011	0.00	4	0.01	997
AQI	217.321	ug/m^3	152.63	0.00	100	8.85	1000

Tabel 4 Deskripsi dataset yang telah diproses sebelumnya

Dataset	CPCB (Central Pol-Dewan Pengendalian Lusi India)
Panjang sampel	270,596
Jumlah Atribut	10
Jumlah Kelas	6
Sampel di setiap kelas	
Kelas 1	13,452
Kelas 2	47,910
Kelas 3	93,167
Kelas 4	55,045
Kelas 5	30,421
Kelas 6	30,601
Rasio Ketidakseimbangan	6.92

Tabel 5 Deskripsi kualitas udara

AQI Jangk	Pelabelan yang Ditentukan milik	Kelas
0-50	Bagus.	Kelas 1
50-100	Memuaskan	Kelas 2
100-200	Sedang	Kelas 3
200-300	Misk in	Kelas 4
300-400	Sangat Buruk	Kelas 5

masalah ketidakseimbangan multi-kelas. Dengan kata lain, kita dapat mengatakan bahwa metode yang digunakan

Menghitung parameter $_{zj}$

Misalkan P adalah kumpulan data, yang mencakup N jumlah sampel dengan C kategori. Nilai parameter $_{zj}$ dihitung dengan menggunakan Persamaan 2 dan 3. Nilai chi-square (3^2) dalam distribusi optimal adalah,

$$3^2 = \sum_{j=1}^{C} \frac{(nj - N/C)^2}{N/C} \tag{7}$$

di mana $_{nj}$ = jumlah sampel pada kategori ke-j dan $j = 1, 2, \ldots, C$

Misalkan, $_{Xj} = \frac{(nj - N/C)^2}{N/C}$

Kalau begitu,

$$3^2 = \sum_{j=1}^{C} Xj \tag{8}$$

Jadi, parameter $_{zj}$ dapat didefinisikan sebagai

$$z_j = w_j \times \frac{Xj}{3^2} \tag{9}$$

Dari Persamaan (8), masukkan nilai 3^2

$$z_j = w_j \times \frac{Xj}{\sum_{j=1}^{C} Xj} \tag{10}$$

untuk mengkompensasi distribusi data yang tidak merata didefinisikan sebagai:

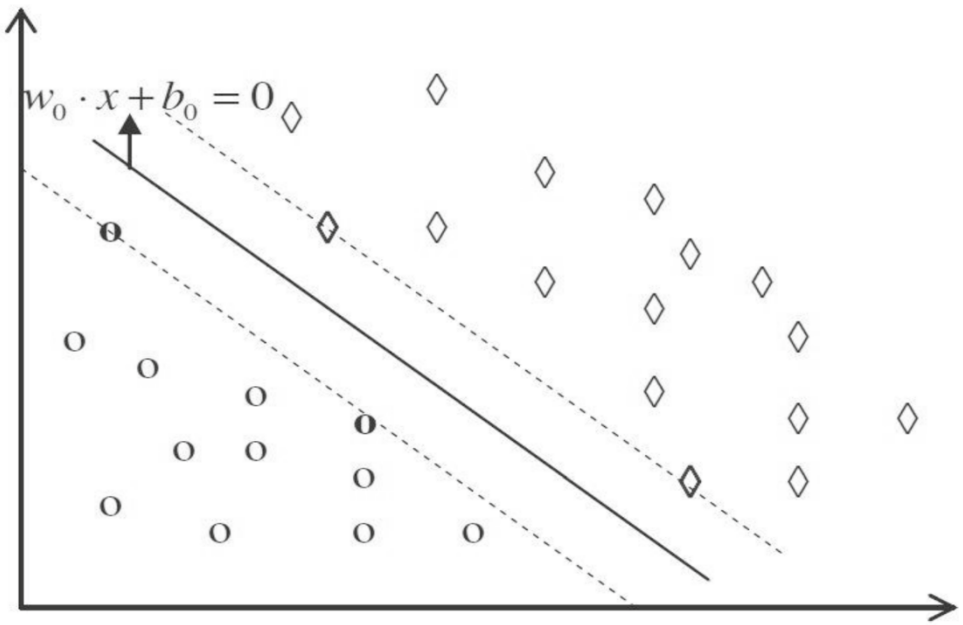
$N\hat{n}j$

$$_{wj} = \frac{}{\sum_{j=1}^{C} N\hat{n}j} \quad (6)$$

di mana, N dan C masing-masing menunjukkan ukuran sampel pelatihan dan ukuran kategori. nj menunjukkan ukuran sampel dari setiap kucing
dengan $j = 1, 2, \ldots, C$.

Deskripsi algoritma yang diusulkan

Diagram alir algoritme yang diusulkan telah ditunjukkan pada Gbr. 3. Pertama-tama, pembersihan kualitas udara CPCB harus dilakukan.
dilakukan, dan setelah itu, data yang diusulkan ini disajikan ke algoritma klasifikasi untuk mendapatkan partisi awal. Pada langkah kedua, kami menghitung nilai faktor pembobotan $_{wj}$ dan parameter $_{zj}$ untuk setiap vektor sup port pada setiap iterasi. Nilai dari parameter ini



Gbr. 2 Hyperplane dengan Vektor Pendukung dalam Paradigma Algoritma SVM

dihitung dengan menggunakan uji Chi-square. Pada langkah berikutnya, fungsi transformasi kernel dihitung, dan akhirnya, model klasifikasi dilatih ulang menggunakan matriks kernel baru yang telah dihitung, yaitu $_{Kmt}$.

Algoritma untuk model klasifikasi yang diusulkan terdiri dari 11 langkah dan langkah-langkah ini dijelaskan dalam Algoritma 1.

Algorithm 1. Procedure of the Proposed Algorithm

Step 1: START

Step 2: Initialization of SVM classifier with the training set X_{train} and kernel matrix $K = K_m$

Step 3: Based on training sample $x \in X_{train}$ the distance $h(x)$ is obtained with the initial partition of data $\{P^j, j = 1,2, \ldots, C\}$. ($C$ = No. of categories)

Step 4: $t \leftarrow 1$

Step 5: while $(t <= T)$ {

Step 6: Obtain the values of the parameters $z_j = w_j \times \frac{X_j}{\sum_{j=1}^{C} X_j}$ and $w_j = \frac{N/n_j}{\sum_{j=1}^{C} N/n_j}$

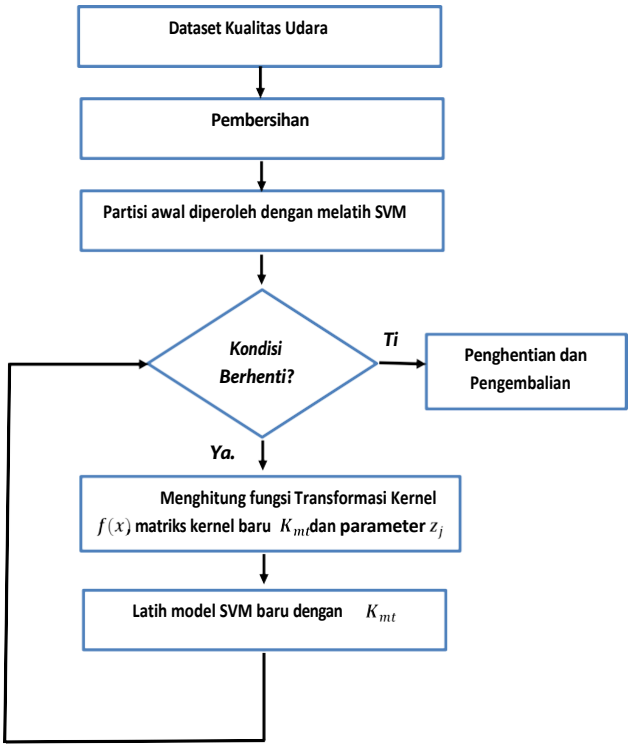
Step 7: Obtain the value of $f_{t-1}(x)$ for the training sample $x \in X_{train}$ by using Eqn. (17).

Step 8: The new kernel matrix K_{mt} is obtained by using the old kernel matrix (K_m) and $f_{t-1}(x)$

Step 9: Again, train the original SVM classifier with the training set X_{train} and kernel matrix K_{mt}.

Step 10: $i = i + 1$ }

Step 11: END



Gbr. 3 Diagram alir algoritma yang diusulkan

Analisis statistik

Pada bagian ini, berbagai ukuran statistik yang digunakan untuk mengevaluasi performa algoritme telah dibahas. Analisis statistik adalah salah satu tugas penting yang membantu kita memilih algoritme terbaik berdasarkan kinerjanya.
Dalam makalah ini, beberapa ukuran statistik untuk mengevaluasi
Algoritma yang diusulkan dan algoritma yang ada telah dipilih
sen untuk menemukan algoritme terbaik di antara mereka.
Statistik
Ukuran-ukuran yang telah dipertimbangkan adalah akurasi, presisi, recall, f1-score, dan TNR, NPV, FNR, FPR, FDR, FOR [61-64]. Dengan bantuan sepuluh ukuran evaluasi ini, kita dapat menentukan algoritma yang tepat yang dapat melakukan tugas klasifikasi dengan lebih efektif dan efisien.

Akurasi

Akurasi sehubungan dengan tugas klasifikasi adalah persentase contoh yang diklasifikasikan dengan benar. Dengan kata lain, kita dapat mengatakan bahwa akurasi

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \quad (11)$$

di mana TP, FP adalah jumlah positif sejati dan positif palsu masing-masing, dan FN, TN mewakili jumlah negatif palsu dan negatif sejati, masing-masing.

Presisi

Presisi, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur jumlah kelas positif yang diprediksi yang benar-benar termasuk dalam kelas positif. Dengan kata lain, kita dapat mengatakan bahwa presisi adalah rasio dari kelas positif yang benar terhadap jumlah total kelas yang benar-benar positif dan kelas positif palsu. Formulasi presisi telah dijelaskan dalam persamaan di bawah ini [61, 63].

$$Presisi = \frac{(TP)}{(TP + FP)} \quad (12)$$

di mana, $TPdanFP$ masing-masing adalah jumlah true positive dan false positive.

Ingat

Recall, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur jumlah kelas positif yang diprediksi keluar dari semua contoh positif dalam kumpulan data. Dengan kata lain, kita dapat mengatakan bahwa recall adalah rasio dari kelas positif yang benar dibandingkan dengan jumlah total kelas yang benar-benar positif dan negatif. Formulasi recall telah dijelaskan dalam persamaan di bawah ini [63].

$$Recall = \frac{(TP)}{(TP + FN)} \quad (13)$$

adalah rasio persentase dari

di mana, *TP dan FN* masing-masing adalah jumlah true positive dan false negative.

Skor F1

Skor F1 juga dikenal sebagai F Measure atau F Score. F1-kelas yang diprediksi dengan benar dari seluruh kelas pengujian. Formulasi akurasi telah dijelaskan dalam persamaan di bawah ini [61].

score, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur keseimbangan antara recall dan presisi. Dengan kata lain, kita dapat mengatakan bahwa F1-score adalah hasil kali antara recall dan precision dari penjumlahan recall dan precision. Formulasi f1-score telah dijelaskan dalam persamaan di bawah ini [62].

$$f1 = \frac{2 \times (presisi \times recall)}{ketepatan + penarikan\ kembali} \quad (14)$$

Tingkat negatif sejati (TNR)

TNR, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur spesifisitas atau tingkat negatif yang sebenarnya. Dengan kata lain, kita dapat mengatakan bahwa TNR adalah rasio dari kelas negatif yang benar terhadap jumlah total kelas yang benar-benar negatif dan positif. Formulasi TNR telah dijelaskan dalam persamaan di bawah ini [61, 64].

$$TNR = \frac{(TN)}{(TN + FP)} \quad (15)$$

di mana, TN dan FP adalah jumlah negatif dan positif yang benar dan salah
positif, masing-masing.

Nilai prediksi negatif (NPV)

NPV, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur rasio nilai prediksi negatif. Dengan kata lain, kita dapat mengatakan bahwa NPV adalah rasio kelas positif yang benar terhadap jumlah total kelas yang benar-benar positif dan negatif. Formulasi NPV telah dijelaskan di bagian
di bawah persamaan [61, 64].

$$NPV = \frac{(TN)}{(TN + FN)} \quad (16)$$

di mana, TN dan FN masing-masing adalah jumlah negatif benar dan salah.

Tingkat negatif palsu (FNR)

FNR, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur tingkat kesalahan. Dengan kata lain, kita dapat mengatakan bahwa FNR adalah rasio kelas negatif palsu terhadap jumlah total kelas yang benar-benar positif dan negatif palsu. Formulasi FNR telah dijelaskan dalam persamaan di bawah ini [61, 64].

$$FNR = \frac{(FN)}{(FN + TP)} \quad (17)$$

di mana, TP dan FN masing-masing adalah jumlah true positive dan false negative.

Tingkat positif palsu (FPR)

FPR, sehubungan dengan tugas klasifikasi, digunakan untuk mengukur tingkat kegagalan. Dengan kata lain, kita dapat mengatakan bahwa FPR adalah rasio dari kelas

di mana, FP dan TN masing-masing adalah jumlah false-positif dan true negatif.

Tingkat penemuan salah (FDR)

FDR adalah rasio dari kelas positif palsu terhadap jumlah total kelas yang benar-benar positif dan positif palsu. Formulasi FDR telah dijelaskan dalam persamaan di bawah ini [61,
64].

$$FDR = \frac{(FP)}{(FP + TP)} \quad (19)$$

di mana, FP dan dTP masing-masing adalah jumlah negatif palsu dan benar.

Tingkat kelalaian yang salah (FOR)

FOR adalah rasio dari kelas false-negative terhadap jumlah total kelas yang benar-benar negatif dan false-negative. Formulasi FOR telah dijelaskan dalam persamaan di bawah ini [61, 64].

$$FOR = \frac{(FN)}{(FN + TN)} \quad (20)$$

di mana, TN dan dFN adalah jumlah negasi yang benar dan salah.
tif, masing-masing.

Hasil

Pada bagian ini, kita akan membahas hasil klasifikasi berdasarkan algoritma klasifikasi, yaitu Algoritma Ada Boost (ADB) [65-67], Algoritma Multilayer Perceptron (MLP) [68-70], Algoritma Gaussian NB (GNB) [71-73], Algoritma Support Vector Machine (SVM) standar [58-60], Algoritma Ada
metode literatur dan usulan berbasis kernel yang dapat diskalakan
Algoritma SVM.

positif palsu terhadap jumlah total kelas yang benar-benar negatif dan positif palsu. Perhitungan FPR telah dijelaskan dalam persamaan di bawah ini [61, 64].

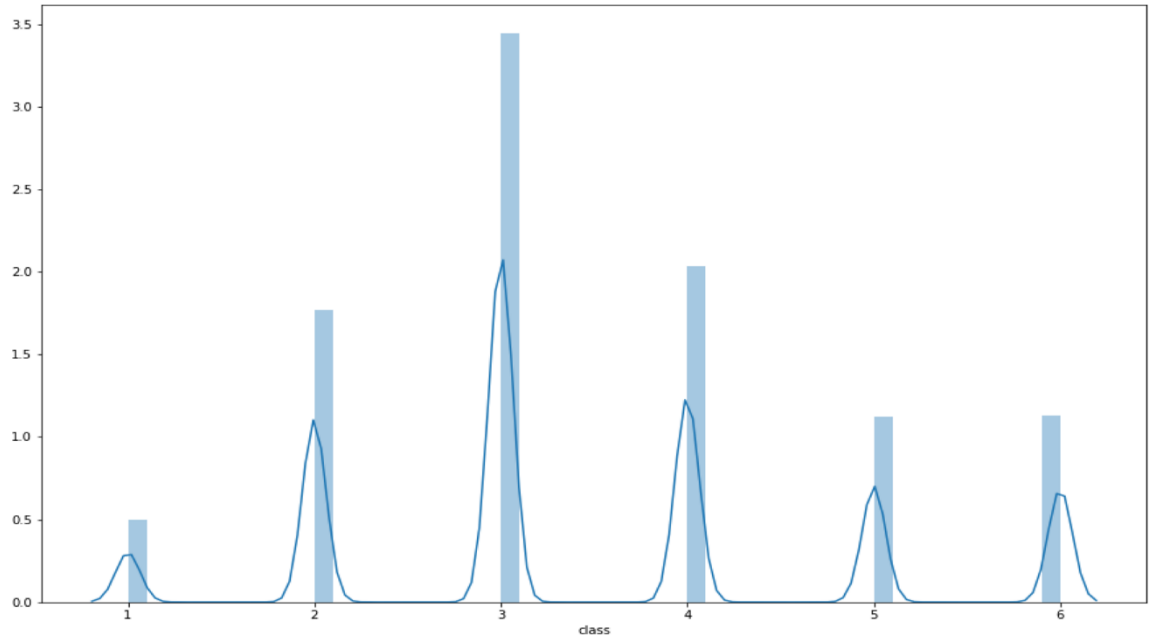
Perbandingan model

Mengidentifikasi model klasifikasi terbaik yang mampu menangani masalah ketidakseimbangan kelas adalah salah satu tugas yang kompleks. Dataset kualitas udara CPCB telah diambil untuk analisis eksperimental. Pada

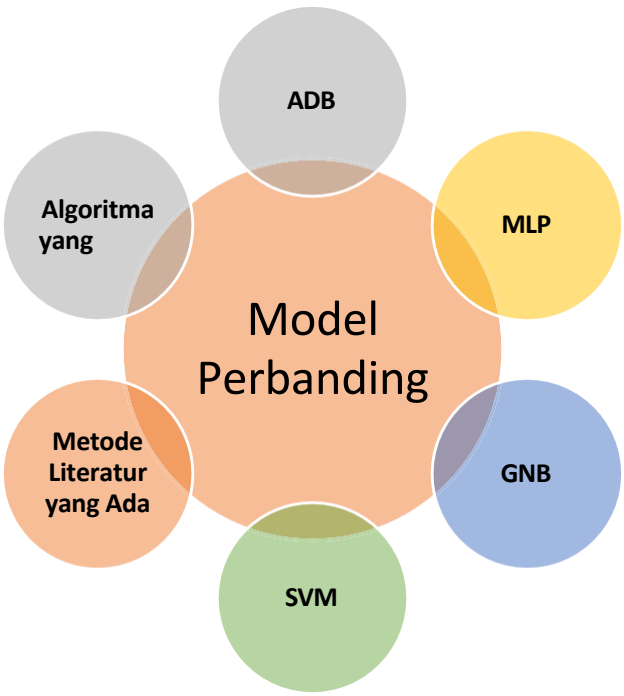
$$FPR = \frac{(FP)}{(FP + TN)} \tag{18}$$

Gbr. 4, *sumbu* x menunjukkan berbagai kelas, dan *sumbu y* menunjukkan jumlah sampel data dalam beberapa kelas. Dari Gbr. 4, jelas bahwa kumpulan data kami berisi distribusi kelas yang tidak merata, atau dapat dikatakan tidak seimbang. Oleh karena itu, menjadi lebih sulit untuk menangani situasi seperti ini dengan model klasifikasi tradisional.

Distribusi kelas dari dataset berdasarkan ukuran sampel adalah: kelas pertama terdiri dari 13.452 sampel, kelas kedua berisi 47.910 sampel, kelas ketiga memiliki 93.167



Gbr. 4 Distribusi berdasarkan kelas dari dataset BPKB



Gbr. 5 Model klasifikasi untuk evaluasi eksperimental

model klasifikasi yang dapat mengatasi ketidakseimbangan kelas

sampel, kelas keempat memiliki 55.045 sampel, kelas kelima berisi 30.421 sampel, dan kelas terakhir berisi 30.601 sampel. Dataset ini juga memiliki rasio ketidakseimbangan kelas sebesar 6,92. Tujuan utama dari penelitian ini adalah untuk menemukan cara terbaik

masalah. Dari waktu ke waktu, banyak peneliti telah memberikan solusi yang berharga untuk menangani masalah ketidakseimbangan kelas ini. Sebagian besar solusi diusulkan untuk masalah ketidakseimbangan kelas biner dan tidak cocok untuk masalah ketidakseimbangan multi-kelas. Keterbatasan ini memotivasi kami untuk memodifikasi algoritma yang dapat secara efisien menangani masalah ketidakseimbangan kelas multi-kelas dan kelas biner tanpa mengorbankan kinerja algoritma. Klasifikasi ini juga akan sangat membantu dalam membuat solusi yang memungkinkan untuk menuju layanan kesehatan yang mahir.

Untuk evaluasi eksperimental, empat algoritme klasifikasi tradisional yang sudah mapan dan metode literatur yang ada dengan algoritme yang kami usulkan telah diambil. Algoritma yang kami usulkan telah dibandingkan dengan algoritma lain untuk menentukan kesesuaian, ketepatan, dan efisiensi. Sepuluh langkah validasi kinerja telah mengukur kinerja semua algoritma klasifikasi. Kebijakan validasi silang sepuluh kali lipat telah digunakan.

Gambar 5 menunjukkan gambaran umum dari algoritma klasifikasi yang telah digunakan dalam tugas klasifikasi. Empat algoritma klasifikasi dan metode literatur yang ada telah dibandingkan dengan algoritma yang kami usulkan untuk menentukan kinerja pengklasifikasi yang diusulkan. Algoritma yang telah digunakan dalam tugas klasifikasi adalah ADB (Algoritma Ada Boost), MLP (Algoritma Multilayer Perceptron), GNB (Algoritma Gaussian NB), SVM (Algoritma Support Vector Machine) standar, metode literatur yang sudah ada, dan algoritma SVM berbasis kernel yang diusulkan.

Tabel 6 Evaluasi kinerja algoritma klasifikasi I

Hasil klasifikasi untuk set data indeks kualitas udara (AQI) yang dihasilkan sensor waktu nyata										
Nama pengklasifikasi	Representasi set data indeks kualitas udara (AQI) yang dihasilkan sensor secara real-time									
	Presisi	Ingat	Skor F1	TNR	NPV	FNR	FPR	FDR	UNT UK	Akurasi
Pengklasifikasi Ada Boost	0.48	0.60	0.46	0.59	0.59	0.41	0.08	0.41	0.41	59.72
Pengklasifikasi MLP	0.96	0.97	0.96	0.95	0.95	0.03	0.01	0.03	0.05	95.71
Gaussian NB	0.81	0.81	0.81	0.80	0.80	0.19	0.03	0.19	0.2	80.87
Pengklasifikasi SVM	0.97	0.97	0.97	0.96	0.96	0.03	0.01	0.03	0.04	96.92
Algoritma yang diusulkan	1.00	1.00	1.00	0.99	0.99	0.002	0.001	0.002	0.01	99.66

*TNR- Tingkat Negatif Sejati, NPV- Nilai Prediksi Negatif, FNR- Tingkat Negatif Palsu, FPR- Tingkat Positif Palsu, FDR- Tingkat Penemuan Palsu, FOR- Tingkat Kelalaian Palsu

Tabel 7 Evaluasi kinerja metode literatur yang ada vs algoritma klasifikasi yang diusulkan

Model yang digunakan	Akurasi (%)
Metode literatur yang ada	
Peningkatan yang sensitif terhadap biaya[41]	90.52
SVM yang peka terhadap biaya [43]	95.01
SVM berbasis fuzzy [44]	97.19
SVM yang ditingkatkan [46]	97.51
Impoved SVM [49]	96.90
Model yang Diusulkan	
SVM berbasis kernel yang dapat diskalakan	**99.66**

Evaluasi kinerja algoritme klasifikasi

Evaluasi kinerja algoritma klasifikasi telah dibagi menjadi dua bagian. Pada bagian pertama, dataset kualitas udara CPCB dari seluruh wilayah Delhi telah diambil, yang berasal dari 37 stasiun pangkalan yang tersebar. Semua data telah disajikan sebagai satu file untuk melakukan tugas klasifikasi. Hasil klasifikasi dari semua algoritma telah dievaluasi dalam bentuk presisi, recall, skor F1, TNR, NPV, FNR, FPR, FDR, FOR, dan akurasi. Evaluasi dari algoritma klasifikasi telah dilakukan berdasarkan akurasi klasifikasi. Seperti yang kita ketahui, data set kami mengandung distribusi kelas yang tidak seimbang yang dapat mempengaruhi kinerja algoritma klasifikasi. Semua model standar berkinerja baik kecuali Ada Boost Classifier (ADB). Pengklasifikasi ADB mencapai akurasi terendah yaitu 59,72 di antara semua pengklasifikasi. Pengklasifikasi SVM standar, pengklasifikasi MLP, dan Gaussian NB berkinerja cukup baik dalam distribusi kelas yang tidak seimbang. Tetapi jika kita membandingkannya dengan pengklasifikasi SVM yang kami usulkan, maka pengklasifikasi ini kalah dalam pertempuran. Algoritma yang kami usulkan memenangkan pertarungan dengan akurasi tertinggi
99,66 di antara semua model lainnya. Analisis terperinci dari hasil klasifikasi telah ditunjukkan pada Tabel 6.

Pada bagian kedua dari evaluasi kinerja, kami telah mengambil dataset CPCB, yang berasal dari 37 tempat di Delhi. Algoritme yang diusulkan mencapai akurasi tertinggi

sebesar 99,66% di antara metode literatur yang ada. Algoritma ini juga efisien untuk menangani masalah ketidakseimbangan kelas tanpa mengorbankan kinerja. Evaluasi kinerja metode literatur yang ada Vs algoritma klasifikasi yang diusulkan telah disajikan pada Tabel 7.

Pada bagian kedua dari evaluasi kinerja, kami telah mengambil data individu dari setiap stasiun basis CPCB, yang diplot di 37 tempat di Delhi. Ke-37 file data tersebut telah digunakan sebagai kumpulan data masukan untuk melakukan tugas klasifikasi dengan bantuan berbagai algoritma klasifikasi. Rincian tentang akronim yang digunakan pada Tabel 8 telah didefinisikan pada Lampiran 1. Algoritma yang kami usulkan telah bekerja dengan sangat baik dalam analisis yang ketat ini untuk semua dataset yang berada di antara A1 hingga A37. Algoritma yang kami usulkan mencapai akurasi rata-rata tertinggi sebesar 99,72 (rata-rata dari A1 hingga A37) di antara semua algoritma. Algoritma ini juga efisien untuk menangani masalah ketidakseimbangan kelas tanpa mengorbankan kinerja. Analisis terperinci dari hasil telah ditunjukkan pada Tabel 8.

Diskusi

Ada banyak faktor terkait yang mungkin memainkan peran penting dalam mempengaruhi kualitas udara. Beberapa faktor secara langsung dan beberapa faktor secara tidak langsung ikut mencemari udara. Polutan yang larut dalam udara berbahaya bagi kesehatan manusia. Kondisi difusi yang buruk adalah salah satu faktor penting yang memainkan peran penting dalam meningkatkan tingkat polutan. Dorongan parsial udara dari ruang konsentrasi tinggi ke ruang konsentrasi rendah dikenal sebagai difusi. Sebelum melakukan tugas klasifikasi, preprocessing dataset dilakukan. Preprocessing adalah proses membuang nilai yang hilang dan objek yang tidak biasa dari set data. Dataset terdiri dari banyak fitur yang dapat dipertanggungjawabkan seperti PM10 (Konsentrasi Partikel yang Dapat Dihirup), SO2 (Sulfur Dioksida), PM2.5 (Partikel Halus), O3 (ozon), NOx (Nitrogen Oksida), NO2 (Nitrogen Dioksida), NO (Nitrogen Monoksida), NH3 (Amonia), CO (Karbon Monoksida), AQI (Air Quality Index), AQI (Air Quality Index).

Indeks), WD (Arah Angin), C6H6 (Benzena), WS (Angin

Tabel 8 Evaluasi kinerja algoritma klasifikasi II

Data yang dikumpulkan dari	Pengklasifikasi				
	ADB	MLP	GNB	SVM	Algoritma yang diusulkan
A1	67.00	86.40	83.13	94.81	**99.67**
A2	53.14	74.77	82.88	94.47	**99.51**
A3	68.92	75.45	81.06	95.12	**99.65**
A4	66.10	81.39	84.80	95.07	**99.67**
A5	84.78	90.24	82.71	94.96	**99.52**
A6	67.44	83.42	86.00	94.29	**99.56**
A7	68.54	90.26	80.92	93.27	**99.59**
A8	67.67	90.98	81.44	95.83	**99.95**
A9	68.12	84.92	85.05	95.68	**99.73**
A10	73.94	91.50	85.13	97.53	**99.86**
A11	60.74	86.13	83.20	95.41	**99.67**
A12	85.69	90.33	82.55	96.49	**99.79**
A13	67.83	87.26	78.62	96.50	**99.81**
A14	97.58	65.50	82.86	95.20	**99.77**
A15	63.94	85.63	83.23	93.68	**99.25**
A16	66.93	85.48	82.03	96.55	**99.41**
A17	63.91	77.76	81.46	95.94	**99.64**
A18	73.02	91.95	81.62	96.54	**99.95**
A19	68.22	89.95	84.67	97.40	**99.45**
A20	70.82	87.94	80.63	94.51	**100**
A21	69.35	85.57	84.79	95.99	**99.78**
A22	75.54	74.08	82.70	95.60	**99.95**
A23	81.24	91.02	82.18	94.81	**99.81**
A24	73.75	90.84	84.18	94.94	**99.72**
A25	69.88	77.36	83.29	96.56	**99.85**
A26	92.20	83.16	79.97	94.83	**99.53**
A27	66.77	79.01	82.08	96.50	**99.82**
A28	65.78	84.26	83.95	95.08	**99.86**
A29	72.48	92.01	83.52	96.87	**99.87**
A30	65.24	87.00	84.39	94.90	**99.5**
A31	64.33	91.36	80.58	94.63	**99.71**
A32	69.91	85.12	80.55	92.09	**99.67**
A33	88.91	92.78	83.11	96.48	**99.78**
A34	72.87	91.05	80.34	95.72	**99.76**
A35	63.84	91.09	83.50	95.12	**99.91**
A36	82.97	85.59	85.44	96.92	**99.79**
A37	79.14	71.17	84.12	96.63	**99.76**
Akurasi total	**71.85**	**85.13**	**82.78**	**95.48**	**99.72**

Kecepatan), RH (Kelembaban Relatif), SR (Radiasi Matahari), BP (Tekanan Bar) dan AT (Suhu Absolut). Korelasi pada data yang telah diproses sebelumnya dihitung untuk mengetahui hubungan antara kelas dan faktor yang bertanggung jawab.

Pada Gbr. 6, hubungan antara faktor kelas dan responden telah ditunjukkan. Dengan bantuan korelasi, kami

dapat dengan mudah menemukan faktor responsif mana yang berkorelasi tinggi dengan kelas.

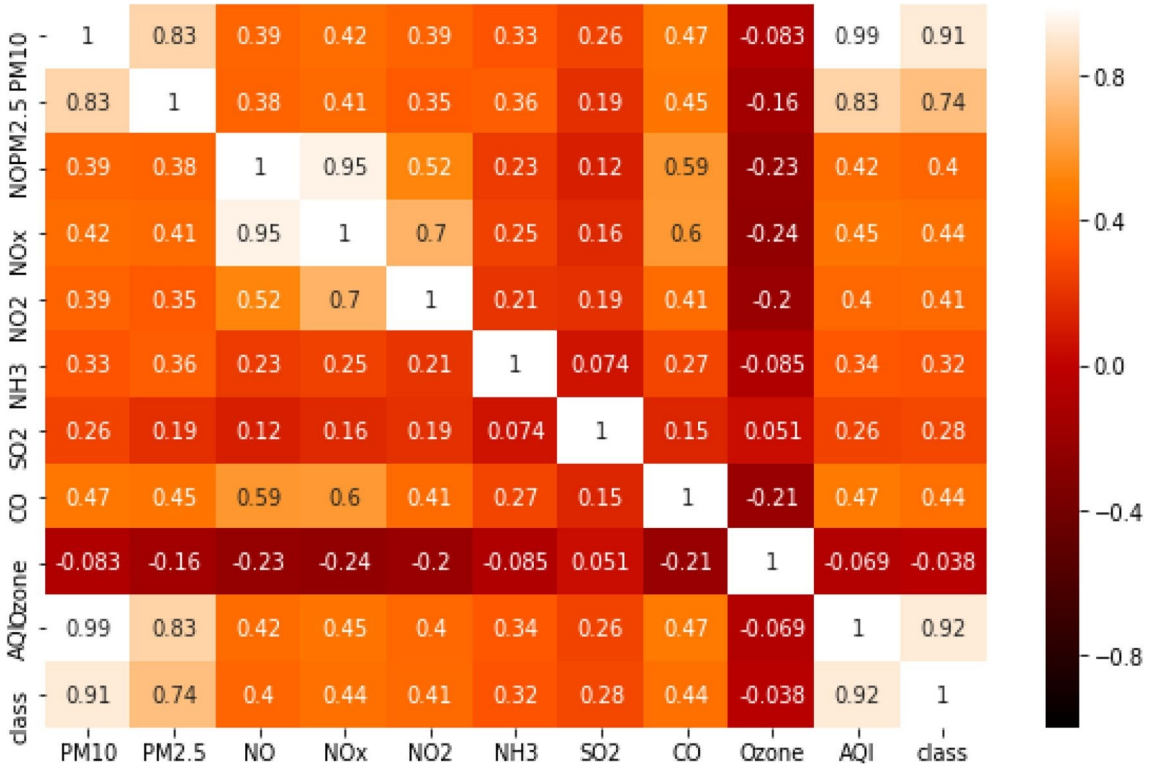
Evaluasi kinerja algoritme klasifikasi

Untuk analisis eksperimental, kumpulan data dari dewan pengendalian polusi pusat India (CPCB) di ibu kota Delhi telah diambil. Data dari 1 Januari 2019 hingga 1 Oktober 2020 telah digunakan untuk tujuan pelatihan dan pengujian. Kebijakan validasi silang sepuluh kali lipat telah digunakan. Validasi silang adalah teknik untuk menilai model dengan mempartisi sampel data yang diberikan ke dalam set pelatihan dan pengujian. Set pelatihan digunakan untuk melatih model sedangkan set pengujian untuk mengevaluasi model. Dalam k-fold cross-validation, sampel data yang diberikan dibagi secara acak menjadi k subsampel dengan ukuran yang sama. Dimana k-1 subsampel digunakan untuk melatih model dan satu subsampel digunakan untuk tujuan validasi. Teknik validasi silang ini diulang hingga k kali (k- fold) dan setiap subsampel digunakan tepat satu kali untuk tujuan validasi. Estimasi tunggal dihasilkan dengan merata-ratakan semua hasil yang berada di bawah k-lipatan. Algoritma yang telah digunakan dalam tugas klasifikasi adalah ADB (Algoritma Ada Boost), MLP (Algoritma Multilayer Perceptron), GNB (Algoritma Gaussian NB), SVM (Algoritma Support Vector Machine) standar, metode literatur yang sudah ada, dan algoritma SVM berbasis kernel yang dapat diskalakan yang diusulkan.

Pada Gbr. 7, hasil eksperimen, (yaitu pengukuran statistik), (yaitu

erbagai algoritma klasifikasi pada dataset CPCB di seluruh wilayah Delhi telah dikirimkan sebelumnya. Dari gambar tersebut, jelas bahwa algoritma yang kami usulkan dengan akurasi tertinggi 99,66 memenangkan perlombaan di antara semua algoritma klasifikasi dan metode literatur yang ada. Hasil dari algoritma yang diusulkan juga lebih baik daripada algoritma SVM tradisional. Jadi, jelas juga dari hasil penelitian ini bahwa algoritma yang kami usulkan efisien untuk menangani masalah ketidakseimbangan kelas tanpa mengorbankan kinerja algoritma.

Pada Gambar 8, hasil klasifikasi berbasis akurasi dari berbagai algoritme klasifikasi pada dataset BPKB, khususnya A1, A10, A20, A30, dan A37 di wilayah Delhi telah diplot menggunakan grafik batang. Dari gambar tersebut, terlihat jelas bahwa algoritme yang kami usulkan mencapai akurasi tertinggi di seluruh area dan memenangkan perlombaan di antara algoritme klasifikasi. Hasil dari algoritma yang diusulkan juga lebih baik daripada algoritma SVM tradisional. Dengan demikian, jelas juga dari hasil bahwa algoritma yang kami usulkan efisien untuk menangani masalah ketidakseimbangan kelas bersama dengan peningkatan kinerja.



Gbr. 6 Koefisien korelasi dari faktor pertanggungjawaban

Efek pada perawatan kesehatan

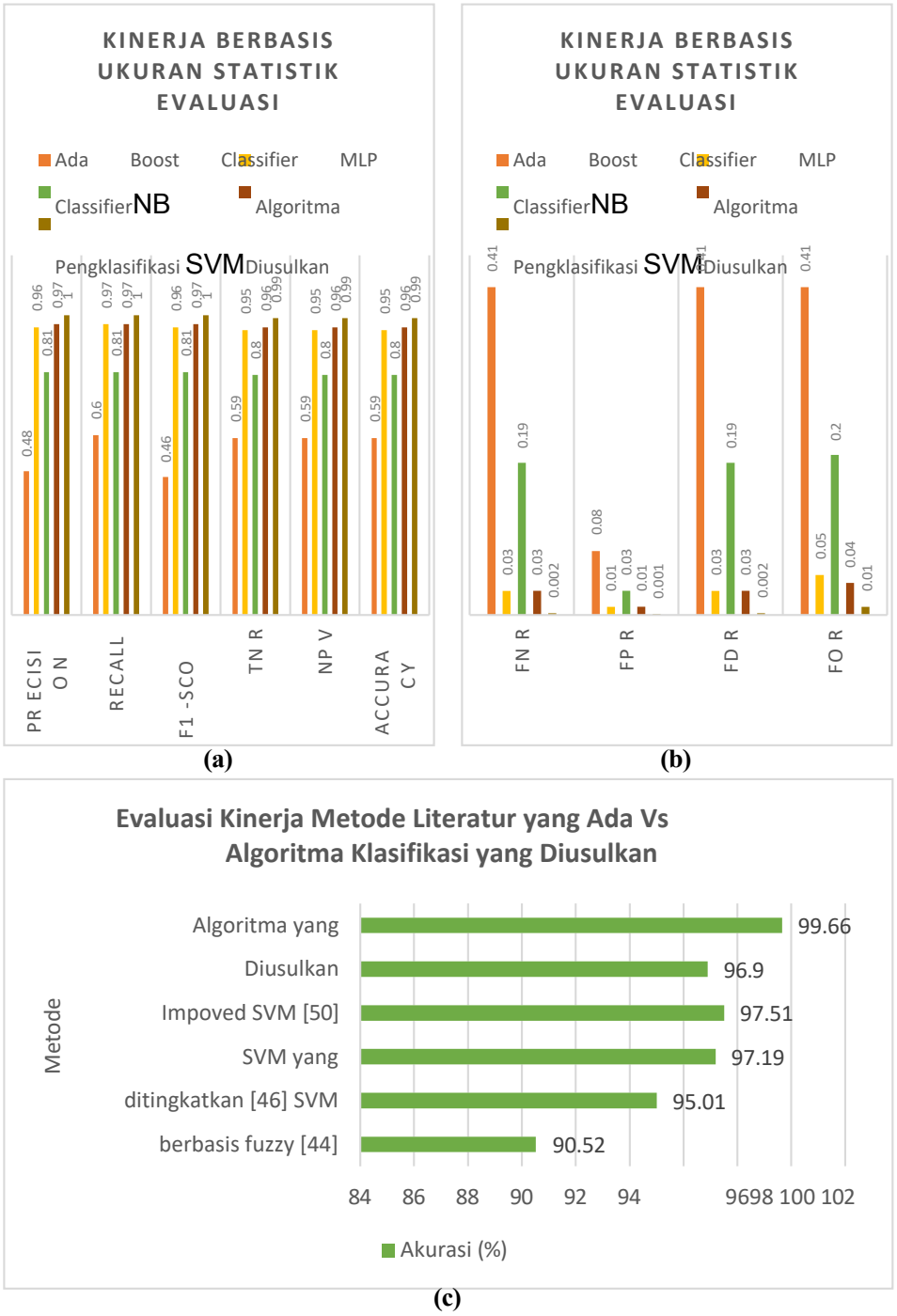
Kualitas udara yang buruk dapat berdampak pada kesehatan dan kualitas hidup individu. Dampak kualitas udara yang buruk dapat menyebabkan masalah dari yang ringan hingga yang berat. Hal ini dapat mempengaruhi sistem kardiovaskular atau peredaran darah, sistem pernapasan, sistem ekskresi (ginjal atau kemih), sistem saraf, sistem endokrin, sistem peredaran darah, sistem pencernaan, sistem limfatik, sistem integumen (kulit), dan sistem mata.

Tabel 9 menunjukkan kisaran AQI dengan pelabelan yang terkait, dan dampak dari berbagai tingkat udara terhadap kesehatan telah ditunjukkan [56]. Tingkat AQI dibagi menjadi enam rentang, mulai dari 0-50 dan berakhir di atas 400.

Konsekuensi dari tingkat AQI yang tinggi terhadap kesehatan individu telah dijelaskan pada Tabel 10. Berbagai dampak dari tingkat AQI yang tinggi dibagi menjadi tiga bagian, yaitu dampak jangka pendek, dampak jangka panjang, dan dampak yang parah. Hal ini dapat menyebabkan masalah parah bagi orang-orang yang menderita penyakit pernapasan. Orang-orang seperti itu membutuhkan perawatan intensif, dan tindakan pencegahan harus dilakukan untuk meminimalkan dampaknya terhadap kesehatan mereka [74-77].

Kesimpulan

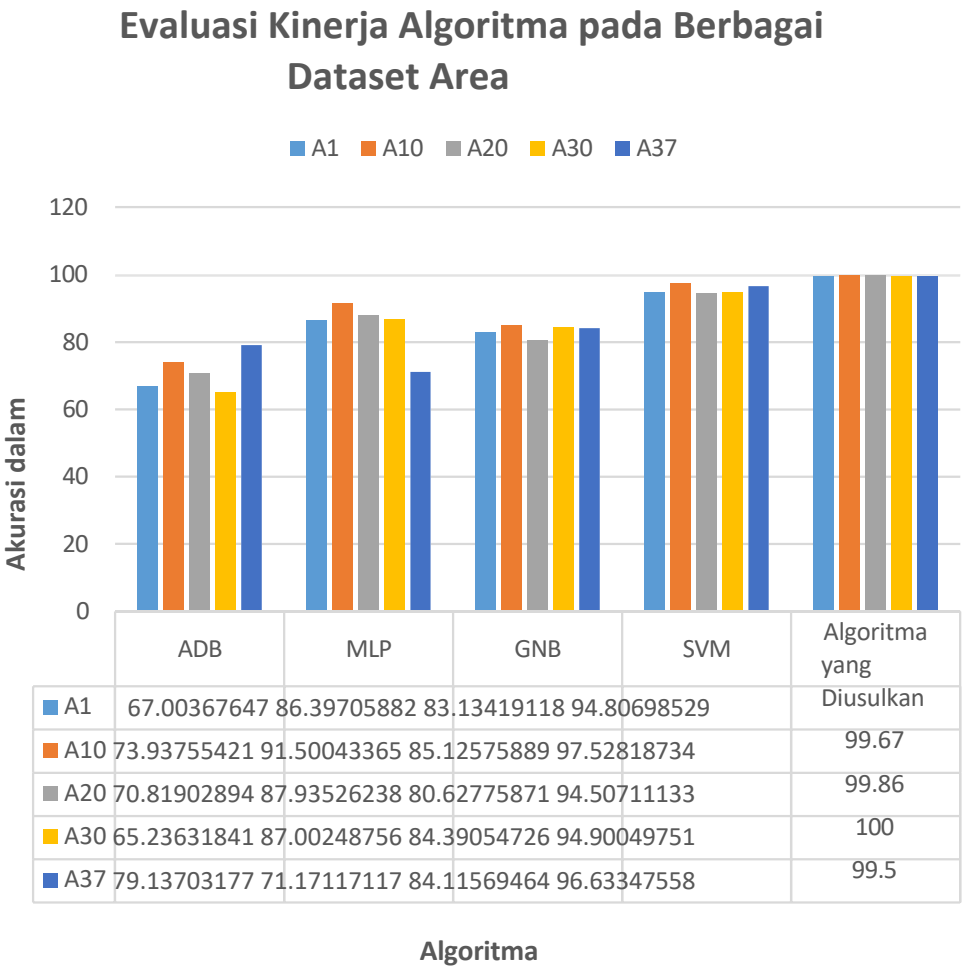
Dalam berbagai masalah klasifikasi, kita menghadapi masalah ketidakseimbangan kelas. Penelitian ini difokuskan untuk menangani distribusi kelas yang tidak seimbang sehingga algoritma klasifikasi tidak akan mengganggu kinerjanya. Algoritma yang diusulkan didasarkan pada konsep metode penskalaan kernel yang dapat disesuaikan (AKS) untuk menangani dataset yang tidak seimbang dengan banyak kelas. Algoritma klasifikasi SVM berbasis kernel yang dapat diskalakan telah diusulkan dan disajikan dalam penelitian ini. Dalam algoritma yang diusulkan, pemilihan fungsi kernel telah dievaluasi berdasarkan kriteria pembobotan dan uji chi- square. Dengan menggunakan fungsi transformasi kernel ini, batas-batas kelas yang tidak rata telah diperluas, dan kemiringan data telah dikompensasi. Untuk evaluasi eksperimental, kami telah mengambil hasil klasifikasi berbasis akurasi dari berbagai algoritma klasifikasi pada dataset BPKB Delhi untuk menemukan dan mengevaluasi kinerja algoritma yang kami usulkan dibandingkan dengan algoritma klasifikasi lainnya. Algoritma yang kami usulkan dengan akurasi tertinggi 99,66% memenangkan perlombaan di antara semua algoritma klasifikasi,



Gbr. 7 Hasil dari algoritma klasifikasi. **a** Ukuran Statistik berdasarkan I. **b** Ukuran Statistik berdasarkan II. **c** Metode Literatur yang Ada Vs Algoritma yang Diusulkan

dan hasil dari algoritma yang diusulkan bahkan lebih baik daripada algoritma SVM tradisional. Hasil dari algoritma yang diusulkan juga lebih baik daripada metode literatur yang ada. Dari hasil ini juga terlihat jelas bahwa algoritma yang kami usulkan efisien dalam menangani ketidakseimbangan kelas dan meningkatkan kinerja. Dalam penelitian ini, kami juga telah membahas pengaruh polusi udara terhadap kesehatan manusia, yang hanya mungkin dilakukan jika data diklasifikasikan dengan benar. Dengan demikian, kualitas udara yang akurat

Gbr. 8 Hasil berbasis akurasi dari algoritme klasifikasi II



Tabel 9 Kisaran indeks kualitas udara dengan kemungkinan dampak kesehatan

Rentang AQI	Pelabelan	Dampak kesehatan
0-50	Dampak Baik	Dampak Kecil
50-100	Memuaskan	Ketidaknyamanan untuk orang yang sensitif seperti masalah pernapasan ringan
100-200	Sedang	Dapat menyebabkan masalah pernapasan bagi orang-orang yang menderita penyakit yang berhubungan dengan paru-paru dan jantung.
200-300	Miskin	Mungkin sebagian besar orang yang hidup dalam situasi ini memiliki masalah pernapasan.
300-400	Sangat Buruk	Mungkinmemiliki penyakit pernapasan
400 +		

klasifikasi melalui algoritma yang kami usulkan akan sangat berguna untuk meningkatkan kebijakan pencegahan yang ada dan juga akan membantu meningkatkan kemampuan tanggap darurat yang efektif jika terjadi pencemaran terburuk.

Di masa depan, algoritma ini akan dibandingkan dengan variasi SVM yang ada saat ini. Algoritma yang diusulkan akan diuji pada dataset lain, dan kami juga akan mencoba meningkatkan metode komputasinya.

Tabel 10 Pengaruh tingkat indeks kualitas udara yang tinggi terhadap kesehatan seseorang

Polutan	AQI
Efek pada kesehatan kardiovaskular yang serius	Jangka pendek 1. Penyakit
	2. Penyakit pernapasan serius
	3. Menyebabkan lebih banyak tekanan pada paru-paru dan jantung
	4. Sel sistem pernapasan yang rusak
Jangka panjang lebih cepat	1. Penuaan paru-paru yang
	2. Pengurangan kapasitas paru-paru
	3. Pengurangan fungsi paru-paru
	4. Bronkitis
	5. Asma
	6. Kemungkinan kanker
	7. Emfisema
	8. Masa pakai yang lebih pendek
Masalah kesehatan yang parah untuk	1. Orang yang menderita penyakit jantung
	2. Orang yang menderita gagal jantung kongestif
	3. Orang yang menderita sindrom arteri koroner
	4. Orang yang menderita asma
	5. Orang yang menderita Emfisema
	6. Orang yang menderita PPOK (Penyakit Paru Obstruktif Kronis)
	7. Wanita dengan Kehamilan
	8. Tenaga kerja di luar ruangan
	9. Orang lanjut usia dan anak-anak di bawah usia 14 tahun
	10. Olahragawan yang berolahraga sangat di luar ruangan

Lampiran 1

Lihat Tabel 11.

Tabel 11 Daftar singkatan

S. tidak	Singkatan	Bentuk lengkap
1	A1	Alipur, Delhi - DPCC
2	A2	Anand Vihar, Delhi - DPCC
3	A3	Ashok Vihar, Delhi - DPCC
4	A4	Aya Nagar, Delhi - IMD
5	A5	Bawana, Delhi - DPCC
6	A6	Persimpangan Burari, Delhi - IMD
7	A7	CRRI Mathura Road, Delhi - IMD
8	A8	Lapangan Tembak Dr. Karni Singh, Delhi-DPCC
9	A9	DTU, Delhi - BPKB
10	A10	Dwarka-Sektor 8, Delhi - DPCC
11	Bandara	A11IGI (T3), Delhi - IMD
12	A12	IHBAS, Taman Dilshad, Delhi-CPCB
13	A13	ITO, Delhi - BPKB
14	A14	Jahangirpuri, Delhi - DPCC
15	A15	Stadion Jawaharlal Nehru, Delhi-DPCC
16	A16	Jalan Lodhi, Delhi - IMD
17	A17	Stadion Nasional Mayor Dhyan Chand, Delhi-Delhi DPCC
18	A18	Mandir Marg, Delhi - DPCC
19	A19	Mundka, Delhi - DPCC
20	A20	Najafgarh, Delhi - DPCC
21	A21	Narela, Delhi - DPCC
22	A22	Nehru Nagar, Delhi - DPCC
23	A23	Kampus Utara, DU, Delhi - IMD
24	A24	NSIT Dwarka, Delhi - BPKB
25	A25	Okhla Fase-2, Delhi - DPCC
26	A26	Patparganj, Delhi - DPCC
27	A27	Punjabi Bagh, Delhi - DPCC
28	A28	Pusa, Delhi - DPCC
29	A29	Pusa, Delhi - IMD
30	A30	R K Puram, Delhi - DPCC
31	A31	Rohini, Delhi - DPCC
32	A32	Shadipur, Delhi - BPKB
33	A33	Sirifort, Delhi - BPKB
34	A34	Sonia Vihar, Delhi - DPCC
35	A35	Sri Aurobindo Marg, Delhi - DPCC
36	A36	Vivek Vihar, Delhi - DPCC
37	A37	Wazirpur, Delhi - DPCC

Deklarasi

Konflik kepentingan Atas nama semua penulis, penulis yang bersangkutan menyatakan bahwa tidak ada konflik kepentingan.

Referensi

1. Menardi G, Torelli N (2014) Melatih dan menilai klasifikasi aturan dengan data yang tidak seimbang. Data Min Knowl Disc 28(1):92-122
2. Japkowicz N, Stephen S (2002) Masalah ketidakseimbangan kelas: studi sistematis . Intell Data Anal 6(5):429-449
3. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2011) Sebuah tinjauan tentang ansambel untuk masalah ketidakseimbangan kelas: pendekatan berbasis bagging, boosting, dan hibrida. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(4):463-484
4. Wang S, Yao X (2012) Masalah ketidakseimbangan multi-kelas: analisis dan solusi potensial. IEEE Trans Syst Man Cybern Part B (Cybern) 42(4):1119-1130
5. Ketu S, Mishra PK (2021) Model klasifikasi hibrida untuk deteksi kondisi mata menggunakan sinyal elektroensefalogram. Cognit Neu- rodyn 1-18
6. Ketu S, Mishra PK (2020). Model pembelajaran mendalam hibrida untuk prediksi COVID-19 dan status terkini uji klinis dunia- luas. Comput Mater Contin 66(2)
7. Tali RV, Borra S, Mahmud M (2021) Deteksi dan klasifikasi leukosit pada gambar apusan darah: keadaan mutakhir dan chal- lenges. Int J Ambient Comput Intell (IJACI) 12(2):111-139
8. Ketu S, Agarwal S (2015) Peningkatan kinerja pengelompokan K-Means yang tidak distribusif untuk analisis data besar melalui komputasi dalam memori. Dalam: Konferensi internasional kedelapan tentang komputasi kontemporer (IC3) 2015, IEEE, hal 318-324
9. Ketu S, Prasad BR, Agarwal S (2015) Pengaruh pemilihan ukuran korpus terhadap kinerja k-means terdistribusi berbasis map-reduce untuk pengelompokan data tekstual berukuran besar. Dalam Prosiding konferensi nasional keenam teknologi komputer dan komunikasi 2015, hal 256-260
10. Ketu S, Kumar Mishra P, Agarwal S (2020). Analisis kinerja kerangka kerja komputasi terdistribusi untuk analisis data besar: hadoop vs spark. Comput Sistemas 24(2)
11. Ketu S, Mishra PK (2020) Analisis kinerja algoritma pembelajaran mesin untuk pengenalan aktivitas manusia berbasis IoT. Dalam Kemajuan dalam teknologi listrik dan komputer, hal 579-591, Springer, Singapura
12. Ketu S, Mishra PK (2021) Model peramalan berbasis regresi proses Gaussian yang disempurnakan untuk wabah COVID-19 dan signifikansi IoT untuk pendeteksiannya. Appl Intell 51(3):1492-1512
13. Ketu S, Mishra PK (2021) Komputasi awan, kabut, dan kabut dalam IoT: indikasi peluang yang muncul. IETE Tech Rev, hal 1-12

14. Chawla NV, Japkowicz N, Kotcz A (2004) Edisi khusus tentang pembelajaran dari kumpulan data yang tidak seimbang. ACM SIGKDD Explor Newsl 6(1):1-6
15. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tou- rassi GD (2008) Melatih pengklasifikasi jaringan saraf untuk pengambilan keputusan medis: efek set data yang tidak seimbang pada kinerja pengklasifikasian. Jaringan Saraf 21(2-3):427-436
16. Kubat M, Holte RC, Matwin S (1998) Pembelajaran mesin untuk mendeteksi tumpahan minyak pada citra radar satelit. Mach Learn 30(2-3):195-215
17. Daskalaki S, Kopanas I, Avouris N (2006) Evaluasi pengklasifikasian untuk masalah distribusi kelas yang tidak merata. Appl Artif Intell 20(5):381-417
18. Vitousek PM (1994) Melampaui pemanasan global: ekologi dan perubahan global. Ekologi 75(7):1861-1876
19. Yilmaz O, Kara BY, Yetis U (2017) Desain sistem pengelolaan limbah berbahaya di bawah pertimbangan populasi dan dampak lingkungan. J Environ Manag 203:720-731
20. De Vito S, Piga M, Martinotto L, Di Francia G (2009) Pemantauan polusi perkotaan CO, NO2 dan NOx dengan hidung elektrik yang dikalibrasi di lapangan dengan regularisasi bayesian otomatis. Aktuator Sensor B Chem 143(1):182-191
21. Northey SA, Mudd GM, Werner TT (2018) Kompleksitas yang belum terselesaikan dalam penilaian penipisan dan ketersediaan sumber daya mineral. Nat Resour Res 27(2):241-255
22. Zhang Q, Jiang X, Tong D, Davis SJ, Zhao H, Geng G, Ni R (2017) Dampak kesehatan lintas batas dari polusi udara global yang diangkut polusi dan perdagangan internasional. Nature 543 (7647): 705-709
23. Du X, Kong Q, Ge W, Zhang S, Fu L (2010) Karakterisasi konsentrasi pajanan pribadi partikel halus untuk orang dewasa dan anak-anak yang terpapar pada konsentrasi ambien yang tinggi di Beijing, China. J Environ Sci 22(11):1757-1764
24. Soh PW, Chang JW, Huang JW (2018) Model prediksi kualitas udara berbasis pembelajaran mendalam yang adaptif menggunakan hubungan spasial dan temporal yang paling relevan. IEEE Access 6:38186-38199
25. Yi X, Zhang J, Wang Z, Li T, Zheng Y (2018) Jaringan fusi terdistribusi dalam untuk prediksi kualitas udara. Dalam Prosiding konferensi internasional ACM SIGKDD ke-24 tentang penemuan pengetahuan & penggalian data, hal 965-973
26. Zhang Y, Wang Y, Gao M, Ma Q, Zhao J, Zhang R, Huang L (2019) Pendekatan prediksi kualitas udara berbasis eksplorasi fitur data prediktif. IEEE Access 7: 30732-30743
27. Iskandaryan D, Ramos F, Trilles S (2020) Prediksi kualitas udara di kota pintar menggunakan teknologi pembelajaran mesin berdasarkan data sensor : sebuah tinjauan. Appl Sci 10(7):2401
28. Xue H, Bai Y, Hu H, Xu T, Liang H (2019) Model hibrida baru berdasarkan algoritma TVIW-PSO-GSA dan mesin vektor pendukung untuk masalah klasifikasi. IEEE Access 7: 27789-27801
29. Mishra M (2019) Racun di udara: Menurunnya kualitas udara di India. Paru-paru India Off Org Indian Chest Soc 36(2):160
30. Bishop CM (2006) Pengenalan pola dan pembelajaran mesin. Springer, New York
31. Packtpub (2018) Algoritma Pembelajaran Mesin. Tersedia secara online: https://www.packtpub.com/in/big-data-and-business-intelligence/ machine-learning-algorithms-second-edition. Diakses pada 9 Desember 2019
32. Longadge R, Dongre S (2013) Masalah ketidakseimbangan kelas dalam tinjauan penambangan data . arXiv:1305.1707
33. He H, Garcia EA (2009) Belajar dari data yang tidak seimbang. IEEE Trans Knowl Data Eng 21(9):1263-1284
34. Gao M, Hong X, Chen S, Harris CJ (2011) Pengklasifikasi RBF berbasis SMOTE dan PSO gabungan untuk masalah ketidakseimbangan dua kelas. Neurocomputing 74(17):3456–3466
35. Kubat M, Matwin S (1997) Mengatasi kutukan set pelatihan yang tidak seimbang: seleksi satu sisi. Dalam: Icml, vol 97, hal 179-186

36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: teknik pengambilan sampel berlebih minoritas sintetis. J Artif Intell Res 16:321-357
37. Liu XY, Wu J, Zhou ZH (2009) Eksplorasi undersampling untuk pembelajaran ketidakseimbangan kelas. IEEE Trans Syst Man Cybern Part B (Cybernetics) 39(2):539-550
38. Prati RC (2012) Menggabungkan algoritma pemeringkatan fitur melalui agregasi peringkat. Dalam: Konferensi gabungan internasional 2012 tentang jaringan syaraf tiruan (IJCNN), hal 1-8. IEEE
39. Gao M, Hong X, Chen S, Harris CJ (2012) Estimasi fungsi kepadatan probabilitas berbasis pengambilan sampel berlebih untuk masalah dua kelas yang tidak seimbang. Dalam: Konferensi gabungan internasional 2012 tentang jaringan saraf (IJCNN), hal 1-8, IEEE
40. Gu Q, Cai Z, Zhu L, Huang B (2008) Penambangan data pada set data yang tidak seimbang. Dalam: Konferensi Internasional 2008 tentang teori dan rekayasa komputasi tingkat lanjut puter (hal 1020-1024). IEEE
41. Sun Y, Kamel MS, Wong AK, Wang Y (2007) Peningkatan yang peka terhadap biaya untuk klasifikasi data yang tidak seimbang. Pengenalan Pola 40(12):3358-3378
42. Zhang Y, Wang D (2013) Metode ensemble yang peka terhadap biaya untuk set data yang tidak seimbang dengan kelas. Dalam Abstrak dan analisis terapan, vol 2013, Hindawi
43. Wang BX, Japkowicz N (2010) Meningkatkan mesin vektor pendukung untuk set data yang tidak seimbang. Knowl Inf Syst 25(1):1-20
44. Batuwita R, Palade V (2010) FSVM-CIL: fuzzy support vector machines untuk pembelajaran ketidakseimbangan kelas. IEEE Trans Fuzzy Syst 18(3):558-571
45. Cano A, Zafra A, Ventura S (2013) Klasifikasi gravitasi data berbobot untuk data standar dan data tidak seimbang. IEEE Trans Cybern 43(6):1672-1687
46. Wu G, Chang EY (2003) Penyelarasan batas kelas untuk pembelajaran dataset yang tidak seimbang. Dalam: Lokakarya ICML 2003 tentang pembelajaran dari set data yang tidak seimbang II, Washington, DC, hal 49-56
47. Wu G, Chang EY (2005) KBA: Penyelarasan batas kernel dengan mempertimbangkan distribusi data yang tidak seimbang. IEEE Trans Knowl Data Eng 17(6):786-795
48. Oh S, Lee MS, Zhang BT (2010) Pembelajaran ansambel dengan pemilihan contoh aktif untuk klasifikasi data biomedis yang tidak seimbang. IEEE/ACM Trans Comput Biol Bioinf 8(2):316-325
49. Liu Y, Yu X, Huang JX, An A (2011) Menggabungkan sam- pling terintegrasi dengan ansambel SVM untuk belajar dari set data yang tidak seimbang. Inf Process Manag 47(4):617-631
50. Ertekin S, Huang J, Giles CL (2007) Pembelajaran aktif untuk masalah ketidakseimbangan kelas. Dalam: Prosiding konferensi internasional ACM SIGIR tahunan ke-30 tentang Penelitian dan pengembangan di pencarian informasi, hal 823-824
51. Fu J, Lee S (2013) Pembelajaran aktif berbasis kepastian untuk pengambilan sampel dataset yang tidak seimbang. Komputasi saraf 119: 350-358
52. Kyrkilis G, Chaloulakou A, Kassomenos PA (2007) Pengembangan indeks kualitas udara agregat untuk aglomerasi mediterania perkotaan: kaitannya dengan potensi efek kesehatan. Environ Int 33(5):670-676
53. Chelani AB, Rao CC, Phadke KM, Hasan MZ (2002) Pembentukan indeks kualitas udara di India. Int J Environ Stud 59(3):331-342
54. Fan S, Hazell PB, Thorat S (1999) Hubungan antara pengeluaran pemerintah, pertumbuhan, dan kemiskinan di pedesaan India (Vol 110). Intl Food Policy Res Inst
55. Deswal S, Verma V (2016) Variasi tahunan dan musiman dalam indeks kualitas udara di wilayah ibu kota negara, India. Int J Environ Ecol Eng 10(10):1000-1005
56. CPCB (2020) Dataset: https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data.

57. Maratea A, Petrosino A, Manzo M (2014) F-measure yang disesuaikan dan penskalaan kernel untuk pembelajaran data yang tidak seimbang. Inf Sci 257:331-341
58. Vapnik VN (1995) Sifat pembelajaran statistik. Teori
59. Wang L (Ed.) (2005) Support vector machines: theory and applikations (Vol 177). Springer, New York
60. Foody GM, Mathur A (2004) Menuju pelatihan cerdas untuk klasifikasi citra super-visual: mengarahkan akuisisi data pelatihan untuk klasifikasi SVM . Lingkungan Penginderaan Jauh 93(1-2):107-117
61. Powers, D. M. (2011). Evaluasi: dari presisi, recall dan F-measure hingga ROC, informedness, markedness dan korelasi.
62. Huang H, Xu H, Wang X, Silamu W (2015) Kriteria pelatihan diskriminatif F1-skor maksimum untuk deteksi kesalahan pengucapan otomatis. IEEE/ACM Trans Audio Speech Lang Process 23(4):787-797
63. Buckland M, Gey F (1994) Hubungan antara recall dan presisi. J Am Soc Inf Sci 45(1):12-19
64. Wikipedia (2021) Matriks kebingungan. https://en.wikipedia.org/ wiki/Confusion_matrix
65. Hastie T, Rosset S, Zhu J, Zou H (2009) Adaboost multi-kelas. Stat Interface 2(3):349-360
66. Schapire RE (2013) Menjelaskan adab. Dalam *Inferensi empiris* (hal 37-52). Springer, Berlin
67. Schapire RE, Freund Y (2013) Meningkatkan: fondasi dan algorithms. Kybernetes
68. Pal SK, Mitra S (1992) Multilayer perceptron, himpunan fuzzy, pengklasifikasian
69. Tang J, Deng C, Huang GB (2015) Mesin pembelajaran ekstrem untuk perceptron multilayer. IEEE Trans Neural Netw Learn Syst 27(4):809-821
70. Chen MS, Manry MT (1993) Pemodelan konvensional dari perceptron multilayer menggunakan fungsi basis polinomial. IEEE Trans Neural Netw 4(1):164-166
71. Bustamante C, Garrido L, Soto R (2006) Membandingkan fuzzy naive bayes dan gaussian naive bayes untuk pengambilan keputusan dalam robocup 3d. Dalam: Konferensi Internasional Meksiko tentang Kecerdasan Buatan, Springer, Berlin, pp 237-247
72. Griffis JC, Allendorfer JB, Szaflarski JP (2016) Klasifikasi Gaussian naïve Bayes berbasis Voxel untuk lesi stroke iskemik pada pemindaian MRI tertimbang T1 individu. J Neurosci Methods 257: 97-108
73. Wu J, Coggeshall S (2012) Dasar-dasar analisis prediktif. CRC Press
74. Ruggieri M, Plaia A (2012) AQI agregat: membandingkan standardisasi yang berbeda dan memperkenalkan indeks variabilitas. Sci Total Environ 420:263-272
75. Friedman JM (1996) Efek obat pada janin dan bayi yang sedang menyusu: buku pegangan bagi para profesional kesehatan. Johns Hopkins University Press, Baltimore
76. Cleland JG, Van Ginneken JK (1988) Pendidikan ibu dan kelangsungan hidup anak di negara berkembang: pencarian jalur pengaruh . Soc Sci Med 27(12):1357-1368
77. Anderson JO, Thundiyil JG, Stolbach A (2012) Membersihkan udara: tinjauan efek polusi udara partikulat pada kesehatan manusia. J Med Toksikol 8(2):166-175