BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Beberapa penelitian sebelumnya telah menggunakan algoritma SVM, XGBoost, dan Random Forest dalam melakukan klasifikasi untuk mengidentifikasi kualitas udara. Oleh karena itu, Tabel 2.1 memuat ringkasan dari sejumlah penelitian terdahulu yang digunakan sebagai referensi dan dasar pertimbangan dalam pelaksanaan penelitian ini.

Tabel 2. 1 Penelitian Terdahulu

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
1	*SENTIMENT ANALYSIS ON E-SPORTS FOR EDUCATION CURRICULUM USING NAIVE BAYES AND SUPPORT VECTOR MACHINE.* (2020)	Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information) 13/2 (2020), 109-122.	Rian Ardianto, Tri Rivanie, Yuris Alkhalifi, Fitra Septia Nugraha dan Windu Gata	Naïve Bayes, Support Vector Machine, Synthetic Minority Over-Sampling Techinique (SMOTE),	Hasil pengujian menunjukkan bahwa perbandingan akurasi pada Naïve Bayes dengan SMOTE adalah 70.32% sedangkan SVM dengan SMOTE mennghasilkan akurasi sebesar 66.92%.	Penelitian ini hanya menggunakan data dari Twitter sehingga masih memiliki keterbatasan pada jumlah data dan penggunaan metode yang belum mendalam sehingga hasil analisis sentimen bisa kurang akurat.	Jurnal ini menjelaskan gambaran umum mengenai pandangan publik terhadap e-sports dalam pendidikan serta membandingkan kinerja Naive Bayes dan SVM sebagai metode analisis sentimen.
2	*Application of machine learning models and landsat 8 data for estimating seasonal pm*	*Environmental Analysis, Health and Toxicology* 39 (2024): e2024011	Bashir Olasunkanmi Ayinde, Muhammed Rabiu Musa , Abdul-Afeez	XGBoost, K-Nearest Neighbour (KNN) and Naive Bayes (NB)	Hasil penelitian menunjukkan bahwa model machine learning, khususnya	Penelitian masih ketergantungan terhadap kualitas citra Landsat 8 yang bisa terpengaruh	Penelitian ini berkontribusi dengan menunjukkan bahwa

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
	2.5 concentrations (2024)		Olakunle Ayinde		Random Forest, mampu memperkirakan PM2.5 musiman dengan akurasi yang cukup tinggi, di mana nilai R^2 mencapai lebih dari 0,75 pada beberapa musim, dan penggunaan data Landsat 8 serta faktor lingkungan membuat hasilnya lebih akurat.	oleh tutupan awan sehingga estimasi PM2.5 tidak selalu konsisten di setiap musim.	kombinasi data satelit dan model machine learning dapat digunakan secara efektif untuk memperkirakan konsentrasi PM2.5 musiman di wilayah yang minim data pemantauan langsung.
3	*Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms. (2024)*	*International Journal of Intelligent Systems and Applications in Engineering*, *10* (2), 269–274.	Abdullah Batuhan Yilmaz, Yavuz Selim Taspinar dan Murat Koklu	Support Vector Machine (SVM) dan Naïve Bayes	Hasil penelititian ini menunjukkan bahwa nilai akurasi pada algoritma Naïve Bayes adalah 92.4% sedangkan algoritma SVM adalah 90.9%.	Penelitian ini masih terbatas pada fitur statis dari aplikasi, sehingga kurang efektif dalam mendeteksi *malware* yang menggunakan teknik penyamaran.	Penelitian ini berkontribusi dengan membuktikan bahwa Naive Bayes dan SVM dapat digunakan untuk mengklasifikasikan aplikasi Android berbahaya secara efisien, sekaligus menjadi

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
							referensi awal dalam penerapan *machine learning* untuk keamanan *mobile*.
4	*A Comparison Between Support Vector Machine (SVM) and Convolutional Neural Network (CNN) Models For Hyperspectral Image Classification (2019)*	*IOP Conference Series: Earth and Environmental Science* (Vol. 357, No. 1, p. 012035). IOP Publishing.	Hayder Hasa, Helmi Z.M. Shafri, Mohammed Habshi	Support Vector Machine dan Convulutional Neural Network	Penelitian ini menunjukkan bahwa akurasi pada CNN sebesar 94.01%, sedangakna kurasi pada algoritma SVM sebesar 98.84%. Maka dapat disimpulkan bahwa algoritma SVM lebih unggul daripada CNN	Tidak membahas kebutuhan CNN yang jauh lebih tinggi dibanding SVM, sehingga kurang cocok untuk penerapan di sistem dengan sumber daya terbatas.	Penelitian ini berkontribusi dengan memberikan perbandingan yang jelas antara akurasi SVM dan CNN dalam klasifikasi citra hiperspektral, sehingga memudahkan orang memilih metode yang paling sesuai.

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
5	*Heart Disease Prediction System using hybrid model of Multilayer perception and XGBoost algorithms. (2024)*	BIO Web of Conferences vol. 97, p. 00047. EDP Sciences, 2024.	Israa Nadheer	Multi-layer perception , Neural Network, dan XGBoost	Pada penelitian ini, peneliti berpendapat bahwa model gabungan MLP-NN dan XGBoost, menunjukkan kinerja yang bagus dalam tugas klasifikasi penyakit jantung, dengan pencapaian akurasi 96,67%, sensitivitas 95,92%, presisi 97,92%, serta F1 score mencapai 96,91%.	Model gabungan MLP dan XGBoost cukup kompleks dan membutuhkan waktu pelatihan lebih lama, sehingga kurang efisien untuk sistem dengan keterbatasan komputasi.	Penelitian ini memberikan solusi prediksi yang lebih akurat dengan menggabungkan dua algoritma kuat, yang dapat membantu deteksi dini penyakit jantung secara lebih efektif.
6	*Sentiment Analysis for Zoning System Admission Policy Using Support Vector Machine and Naive Bayes Methods.* (2023)	*Journal of Physics: Conference Series* (Vol. 1776, No. 1, p. 012058). IOP Publishing.	Reynaldy Aries Ariyanto, N Chamidah	Support Vector Machine dan Naïve Bayes	Penelitian ini menunjukkan bahwa hasil dari akurasi penggunaan SVM adalah 92.93% dan Naïve Bayes adalah 79.86%. Hal ini menunjukkan bahwa algoritma SVM memberikan hasil analisis sentimen yang lebih akurat dibanding Naive Bayes dalam	Keterbatasan utama dalam penelitian ini adalah tidak adanya validasi dari data lapangan secara langsung, sehingga hasil analisis sentimen bergantung sepenuhnya pada opini publik di media sosial yang belum tentu mewakili keseluruhan populasi. seimbang.	Penelitian ini memberikan kontribusi dengan menyediakan gambaran awal tentang opini publik terhadap sistem zonasi dan membandingkan efektivitas dua metode analisis sentimen yang sering digunakan.

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
					menilai respons publik terhadap kebijakan sistem zonasi.		
7	*Comparing Random Forest and Support Vector Machine for breast cancer classification*. (2020)	*TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(2), 815-821.*	Chelvian Aroef, Yuda Rivan, Zuherman Rustam	Random Forest dan Support Vector Machine	Pada penelitian ini, menunjukkan bahwa hasil dari akurasi algoritma Support Vector Machine adalah 95% sedangkan akurasi pada algoritma Random Forest adalah 90%	Meskipun penelitian ini memberikan hasil akurasi yang baik, namun hanya menggunakan data yang sudah bersih dan terstruktur, sehingga belum tentu seakurat itu jika diterapkan pada data nyata di lapangan.	Penelitian ini membantu menunjukkan bahwa Random Forest memiliki kinerja yang lebih stabil dibanding SVM dalam mengklasifikasikan kanker payudara, sehingga bisa jadi acuan untuk pengembangan sistem deteksi dini.
8	*Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing* (2022)	*Vol. 11 No. 1 (2022) : Komputika: Jurnal Sistem Komputer*	Nabila Bianca Putri dan Arie Wahyu Wijayanto	Naive Bayes, Random Forest, dan Support Vector Machine	Nilai akurasi tertinggi adalah algoritma metode Random Forest yaitu sebesar 90.77%, lalu diikuti oleh oleh metode Support Vector Machine sebesar 86,25% dan Naive Bayes	Penelitian ini masih terbatas pada data yang bersifat statis, sehingga kurang efektif dalam menangani phishing website yang terus berubah dan berkembang.	Penelitian ini memberikan gambaran yang jelas tentang perbandingan kinerja beberapa algoritma klasifikasi dalam mendeteksi website phishing, yang dapat membantu

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
					sebesar 82.31%.		dalam memilih metode yang paling tepat dan efisien. Temuan ini bermanfaat sebagai referensi awal dalam pengembangan sistem keamanan berbasis klasifikasi data, khususnya bagi lembaga yang ingin meningkatkan deteksi otomatis terhadap situs berbahaya.
9	*Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease.* (2021)	*ICIC Express Letters*, *15*(1), 9-18.	Dimas Aryo Anggoro dan Dian Novitaningrum	Support Vector Machine dan Artifical Neural Network	Penelitian ini membandingkan performa dua algoritma klasifikasi, yaitu SVM dan ANN, dalam memprediksi risiko diabetes mellitus. Hasil pengujian menunjukkan bahwa ANN memiliki tingkat akurasi lebih tinggi sebesar 85,20%, dibandingkan SVM yang	Penelitian ini terbatas pada jumlah data yang relatif kecil, sehingga hasilnya mungkin belum sepenuhnya mewakili kondisi nyata secara luas.	Penelitian ini memberikan perbandingan yang bermanfaat antara SVM dan ANN dalam prediksi diabetes, sehingga dapat membantu memilih metode yang lebih tepat untuk sistem deteksi awal.

No	Judul Penelitian	Nama Jurnal	Nama Peneliti	Metode	Hasil Penelitian	Kelemahan	Kontribusi
					memperoleh akurasi 83,54%, setelah dilakukan tahap normalisasi data.		
10	*Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services.* (2019)	*Journal of Physics: Conference Series (Vol. 1320, No. 1, p. 012016). IOP Publishing.*	Rosita Kusumawati, Putu Agus Aditya Pramana, A D'arofah	Naïve Bayes Classifier dan Support Vector Machine	Penelitian ini menunjukkan dengan bahwa akurasi pada algoritma Naïve Bayes sebesar 75% sedangkan akurasi pada algoritma SVM sebesar 83.34%.	Penelitian ini terbatas pada analisis data dari Twitter saja, sehingga belum mencerminkan keseluruhan kepuasan pengguna Tokopedia dari platform lain.	Penelitian ini memberikan gambaran awal tentang performa Naive Bayes dan SVM dalam mengklasifikasikan opini pengguna, yang dapat digunakan untuk meningkatkan kualitas layanan Tokopedia. Temuan ini membantu pengembang sistem dalam memilih metode yang paling efektif untuk klasifikasi teks pada platform digital.

Berdasarkan penelitian terdahulu pada tabel 2.1, terdapat artikel jurnal yang dijadikan penulis untuk melakukan pada klasifikasi kualitas udara dengan

menerapkan algoritma Naïve Bayes dan Support Naïve Bayes *(*SVM) untuk melakukan analisis pada data. Seperti artikel jurnal dengan judul "*SENTIMENT ANALYSIS ON E-SPORTS FOR EDUCATION CURRICULUM USING NAIVE BAYES AND SUPPORT VECTOR MACHINE*" yang ditulis oleh Rian Ardianto, Tri Rivanie, Yuris Alkhalifi, Fitra Septia Nugraha dan Windu Gata akan dijadikan sebagai referensi untuk metode Support Vector Machine [8]. Artikel jurnal "*Heart Disease Prediction System using hybrid model of Multilayer perception and XGBoost algorithms*." yang ditulis oleh Israa Nadheer akan dijadikan sebagai referensi untuk algoritma XGBoost [9]. Selanjutnya, artikel jurnal "*Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms*" yang ditulis oleh Abdullah Batuhan Yilmaz, Yavuz Selim Taspinar dan Murat Koklu akan dijadikan sebagai referensi untuk algoritma Naïve Bayes dan Support Vector Machine [5]. Selanjutnya, pada artikel jurnal dengan judul "*Comparing Random Forest and Support Vector Machine for breast cancer classification."* menjadi sebuah referensi untuk penggunaan akurasi pada algoritma Random Forest dan Support Vector Machine[10]. Pada artikel jurnal dengan judul "*Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing"* yang ditulis oleh Nabila Bianca Putri dan Arie Wahyu Wijayanto akan menjadi sebuah referensi untuk penggunaan akurasi pada algoritma Random Forest [8].

2.2 Teori Penelitian

2.2.1 Udara

Udara merupakan campuran dari berbagai komposisi yang meliputi atmosfer Bumi dan mengelilingi planet ini, dengan berbagai unsur kimia, seperti nitrogen, oksigen, gas, energi, ions, zat organik. Komposisi udara memiliki variasi yang signifikan dalam berbagai kondisi geografis. Di dataran tinggi, komposisi udara berbeda dari di dataran rendah. Daerah yang terletak di sekitar khatulistiwa memiliki komposisi udara yang

berbeda dibandingkan dengan daerah kutub. Komposisi udara juga bisa berbeda antara daerah yang memiliki banyak vegetasi dengan daerah yang didominasi oleh industri. Selain itu, perbedaan dapat ditemukan antara daerah rural dan daerah urban. Secara keseluruhan, komposisi udara kering dan bersih terdiri dari berbagai gas seperti nitrogen, oksigen, argon, karbondioksida, neon, helium, metana, kripton, nitrous oksida, hidrogen, xenon, dan ozon [11].

Atmosfer di sekitar bumi, yang kita kenal sebagai udara, memiliki peran yang krusial dalam mendukung kehidupan di planet ini. Dalam atmosfer, terdapat unsur- unsur esensial seperti oksigen (O2) yang diperlukan untuk bernafas bagi mahluk hidup, karbon dioksida (CO2) yang memiliki peran dalam proses fotosintesis oleh klorofil daun, serta ozon (O3) yang berfungsi melindungi dari sinar ultraviolet matahari. Selain itu, komposisi udara juga dipengaruhi oleh faktor seperti suhu udara, tekanan udara, dan kondisi lingkungan sekitar. Dalam kondisi udara yang bersih dan kering, komposisi umumnya terdiri dari nitrogen (N2) sekitar 78,09%, oksigen (O2) sekitar 20,94%, argon (Ar) sekitar 0,93%, dan karbon dioksida (CO2) sekitar 0,032%. [12]. Komponen yang paling rentan mengalami perubahan konsentrasi adalah uap air dan CO2. Kegiatan-kegiatan tertentu, seperti proses pembusukan bahan organik, tindakan pembakaran, atau ketika sejumlah besar orang berkumpul dalam ruangan yang terbatas, dapat berpotensi meningkatkan konsentrasi CO2 dalam atmosfer.

Dalam Peraturan Gubernur Daerah Istimewa Yogyakarta Nomor 8 Tahun 2010 yang mengatur tentang Program Langit Biru untuk periode 2009–2013 mendefinisikan udara ambien sebagai udara yang berada tepat di permukaan bumi, terutama pada lapisan troposfer di wilayah Republik Indonesia. Udara ini memiliki peranan yang sangat penting

dalam menopang berbagai aspek kehidupan, termasuk kesehatan manusia, keberlangsungan organisme lain, serta kelangsungan ekosistem dan berbagai komponen lingkungan yang saling terkait. Udara ambien menjadi media utama yang menyediakan oksigen dan unsur penting lainnya yang diperlukan oleh makhluk hidup untuk beraktivitas. Namun, keberadaan dan aktivitas makhluk hidup, baik manusia maupun organisme lainnya, dapat mempengaruhi dan mengubah komposisi alami udara tersebut. Perubahan ini bisa terjadi akibat berbagai faktor seperti polusi dari kendaraan bermotor, aktivitas industri, pembakaran sampah, hingga kegiatan domestik yang menghasilkan zat-zat pencemar. Bila perubahan komposisi udara ini melebihi batas ambang yang telah ditentukan, maka udara tidak lagi mampu menjalankan fungsi dasarnya secara optimal [13]. Kondisi tersebut dikategorikan sebagai pencemaran udara yang dapat menimbulkan dampak negatif terhadap kesehatan manusia serta keseimbangan lingkungan secara keseluruhan. Oleh karena itu, upaya menjaga kualitas udara ambien menjadi sangat penting dengan melaksanakan berbagai langkah pencegahan dan penanggulangan pencemaran udara. Langkah-langkah ini meliputi pengawasan ketat terhadap sumber-sumber polusi, penerapan teknologi ramah lingkungan, serta edukasi kepada masyarakat agar lebih sadar akan pentingnya menjaga kebersihan udara. Selain tindakan pencegahan, program pemulihan kualitas udara juga menjadi komponen kunci dalam strategi perlindungan lingkungan hidup yang diamanatkan oleh peraturan tersebut. Pemulihan ini melibatkan rehabilitasi area yang tercemar dan pengembalian kondisi udara ke tingkat yang aman dan sehat bagi kehidupan di wilayah tersebut, sehingga kebijakan ini tidak hanya berfokus pada penanggulangan dampak polusi, tetapi juga pada

pengelolaan lingkungan secara menyeluruh demi terciptanya udara yang bersih dan sehat bagi semua makhluk hidup [14].

A. Pencemaran Udara

Pencemaran udara adalah penambahan zat-zat fisik atau kimia ke dalam lingkungan udara yang biasanya bersih hingga mencapai tingkat tertentu yang dapat dideteksi oleh manusia atau diukur, serta dapat memberikan dampak negatif pada manusia, hewan, tumbuhan, dan bahan-bahan lain. Dalam konteks lain, pencemaran udara dapat dijelaskan sebagai kehadiran kontaminan dalam atmosfer yang berasal dari aktivitas manusia. Selain itu, pencemaran udara juga dapat diartikan sebagai perubahan dalam komposisi atmosfer karena adanya penambahan zat-zat kontaminan baik yang bersifat alami maupun buatan [14].

Peningkatan konsentrasi zat-zat pencemar di atas tingkat yang dianggap aman akan menghasilkan dampak negatif yang berpotensi membahayakan lingkungan. Dampak tersebut mencakup kerusakan terhadap manusia, tanaman, hewan, dan benda-benda materi, serta berpotensi mempengaruhi kualitas air hujan dengan menghasilkan fenomena hujan asam. Akibatnya, dampak ini akan memengaruhi ekosistem flora dan fauna dalam rantai makanan.

- Sumber Pencemaran Udara

Sumber pencemaran udara dapat diklasifikasikan menjadi dua kategori, yaitu pencemaran yang disebabkan oleh faktor alam, seperti letusan gunung berapi, aktivitas manusia,

seperti emisi dari transportasi, pabrik, dan lainnya. Pencemaran udara berpotensi terjadi di berbagai tempat, termasuk dalam lingkungan perumahan, institusi pendidikan, dan tempat kerja. Pencemaran dalam ruangan seperti ini sering disebut sebagai pencemaran *indoor*. Di sisi lain, pencemaran udara di luar ruangan berasal dari sumber seperti emisi kendaraan bermotor, industri, perkapalan, dan juga dari proses alamiah yang dilakukan oleh makhluk hidup. Sumber pencemar udara dapat dibagi menjadi dua kategori, yaitu sumber tetap, yang melibatkan pembangkit listrik, industri, dan rumah tangga, dan sumber bergerak, yang mencakup aktivitas lalu lintas kendaraan bermotor di darat dan transportasi laut [15].

- Jenis Pencemaran Udara

Prabowo & Muslim (2018) [15], mengidentifikasi beberapa zat pencemar udara yang sering ditemukan di kota-kota. Dilihat dari sifat fisiknya, zat-zat pencemar ini dapat berwujud:

1. Karbon Monoksida (CO)

Sumber utama karbon monoksida di kota adalah emisi dari kendaraan bermotor. Data menunjukkan bahwa sekitar 60%-70% pencemaran udara di Indonesia disebabkan akibat emisi dari kendaraan bermotor berbahan bakar solar, terutama kendaraan seperti kendaraan bermotor [15]. Pembentukan CO terjadi sebagai akibat dari perbandingan antara udara dan bahan bakar selama proses pembakaran dalam mesin diesel. Peningkatan kadar karbon monoksida di perkotaan dapat

memiliki dampak negatif, penurunan berat badan janin, peningkatan angka kematian bayi, serta gangguan pada fungsi otak. Oleh sebab itu, strategi pengurangan kadar karbon monoksida perlu difokuskan pada pengendalian emisi, salah satunya melalui penggunaan katalis yang mampu mengubah karbon monoksida menjadi karbon dioksida.serta penggunaan bahan bakar yang lebih ramah lingkungan untuk kendaraan bermotor.

2. Nitrogen Oksida (NO2)

Transportasi laut di Jepang pada periode 2000 telah berkontribusi sebesar 38% dari total emisi NOx, setara dengan 25.000 ton per tahun [15]. Pembentukan gas NOx melibatkan tiga faktor utama, yaitu suhu (T), waktu reaksi (t), dan konsentrasi oksigen (O2), yang dapat dijelaskan dengan rumus NOx = f (T, t, O2). Ada tiga teori yang menjelaskan proses terbentuknya NOx, yaitu:

Thermal NOx (Mekanisme Zeldovich yang Diperpanjang): Proses ini terjadi ketika gas nitrogen mengalami oksidasi pada suhu tinggi di dalam ruang bakar, khususnya pada suhu di atas 1800 K. Proses thermal NOx didominasi oleh emisi NO (NOx → NO + NO2).

[1] Prompt NOx: Proses pembentukan NOx segera terjadi pada wilayah pembakaran

[2] Fuel NOx: Mengacu pada emisi nitrogen oksida yang

bersumber dari kandungan nitrogen dalam bahan bakar selama proses pembakaran berlangsung

Proses thermal NOx bertanggung jawab atas sekitar 90% emisi nitrogen oksida (NOx). Perlu ditekankan bahwa Heavy Fuel Oil (HFO), yang sering digunakan sebagai bahan bakar kapal, memiliki peranan penting dalam peningkatan emisi tersebut, berkontribusi sekitar 20-30% dari emisi NOx. Nitrogen oksida yang terdapat dalam udara yang dihirup oleh manusia dapat mengakibatkan kerusakan paru-paru [15].

3. Sulfur Oksida (SOx)

Emisi SOx terjadi sebagai akibat dari kandungan sulfur dalam bahan bakar, dan sebagian sulfur juga berasal dari pelumas. Struktur sulfur ini terdapat dalam ikatan aromatic dan alkil. Proses pembakaran menghasilkan sulfur dioksida serta sulfur trioksida melalui reaksi seperti di bawah ini:

[1] $S + O2 \rightarrow SO2$

[2] $SO2 + ½\ O2 \rightarrow SO3$

Sulfur trioksida menyumbang sekitar 1–5% dari keseluruhan emisi SO_x, menjadikannya komponen yang sangat sedikit, dan gas ini tidak berwarna. namun memiliki aroma menyengat, dan diketahui dapat memicu serangan asma pada individu yang sensitif. Di atmosfer, gas ini bisa bereaksi dan

menghasilkan senyawa asam yang berpotensi membahayakan lingkungan. Berdasarkan data dari Organisasi Kesehatan Dunia (WHO), selama rentang waktu tahun 1997 hingga 2003, tingkat konsentrasi sulfur dioksida di udara telah mencapai atau bahkan melampaui batas aman yang telah ditetapkan. [15].

4. Particulate Matter (PM)

Dalam emisi gas, partikel debu terdiri dari beragam komponen, yang berbentuk zat padat dan zat cair yang menetap pada partikel debu. Pembentukan debu terjadi setelah unsur hidrokarbon yang mengalami pemisahan dan juga unsur oksidasi yang mengalami pemisahan juga. Debu tersebut mengandung berbagai jenis debu sendiri serta beberapa metal oksida. Selama proses ekspansi di atmosfer, kandungan metal dan debu tersebut berkomponen membentuk partikulat [22]. Beberapa unsur dalam partikulat termasuk karbon, SOF (Soluble Organic Fraction), debu, SO4 (sulfat), dan H2O (air). Beberapa partikulat mungkin keluar dari cerobong pabrik sebagai asap hitam yang tebal, tetapi yang paling berpotensi berbahaya adalah butiran-butiran halus yang dapat menembus hingga ke dalam bagian terdalam paru-paru manusia. Dalam beberapa kota besar di seluruh dunia, perubahan partikel menjadi sulfat di atmosfer banyak disebabkan oleh proses oksidasi yang dipicu oleh molekul sulfur.

2.2.2 AQI

AQI merupakan alat untuk mengukur kualitas udara secara global. IQAir mencakup berbagai parameter, seperti *Particulate Matter* (PM), nitrogen oksida (NO_x), karbon monoksida (CO), sulfur oksida (SO_x), yang digunakan untuk menilai tingkat pencemaran udara dalam suatu wilayah [16]. Data IQAir digunakan untuk memantau kualitas udara secara global yang membantu untuk meningkatkan kualitas udara pada suatu negara melalui informasi yang diberikan secara publik.

Berikut adalah indeks-indeks yang mengukur tingkat kuallitas udara yang digunakan untuk menilai kategori kualitas udara serta pengaruhnya terhadap kesehatan.

Tabel 2. 2 Indeks Air Quality Index (AQI)

Sumber: [16]

Indeks	Kategori	Dampak
0 - 50	Baik	Tidak ada resiko kesehatan bagi mahluk hidup.
51 - 100	Sedang	Tidak berpengaruh mahluk hidup namun menimbulkan sedikit resiko kesehatan pada individu yang sensitif
101 - 150	Tidak sehat untuk Kelompok sensitif	Individu dan masyarakat umum, memiliki potensi untuk mengalami gangguan pernapasan dan iritasi.

151 – 200	Tidak sehat	Adanya risiko pada jantung dan paru-paru semakin mungkin terjadi di antara masyarakat umum, terutama pada individu yang rentan.
201 – 300	Sangat tidak sehat	Masyarakat umum harus berhati-hati saat keluar ruangan, serta menggunakan masker.
301 – 500+	Berbahaya	Masyarakat umum dan kelompok sensitif sangat memungkinkan untuk mengalami iritasi.

2.3 Framework dan Algoritma Penelitian

2.3.1 CRISP-DM

Gambar 2. 1 Diagram CRISP-DM

Sumber: [17]

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan salah satu metode *data mining*, yang dapat memberikan proses standar yang umum dalam solusi pemecahan masalah perusahaan dengan *data mining* yang tepat [17] *Framework* CRISP-DM terdiri dari 6 fase yaitu:

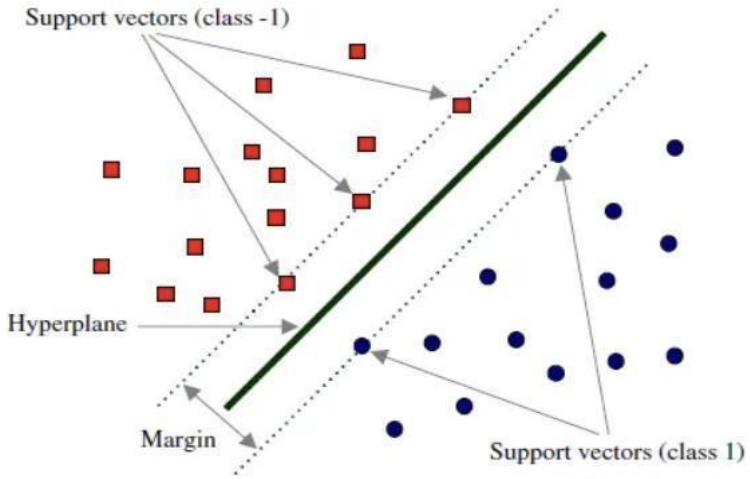
1. *Business Understanding*, adalah fase pertama dimana bisnis perusahaan memfokuskan pada memahami tujuan serta kebutuhan yang dibutuhkan oleh perusahaan tersebut, menyiapkan strategi awal

untuk mencapai tujuan tersebut.

2. *Data Understanding*, adalah fase kedua dimana perusahaan akan melakukan untuk mengindentifikasi, mengumpulkan serta menganalisis data agar dapat mencapai tujuan dari proyek.
3. *Data Preparation*, adalah fase ketiga dimana data mentah yang sudah dikumpulkan akan diolah dan dilakukan pembersihan sehingga menghasilkan data yang bagus serta layak digunakan.
4. *Modelling*, adalah fase keempat dimana perusahaan melakukan pembuatan model kemudian diaplikasikan terhadap data yang sudah dibersihkan dan disiapkan.
5. *Evaluation,* adalah fase kelima dimana saat perusahaan melakukan evaluasi pada data yang sudah bersih dan melakukan analisis tambahan untuk memverivikasi model tersebut dapat mencapai tujuan penelitian.
6. *Deployment,* adalah proses dimana model akan dimanfaatkan untuk menghasilkan laporan berdasarkan dari hasil informasi yang didapat dari evaluasi pada proses *data mining*.

2.3.2 Support Vector Machine

Support Vector Machine (SVM) merupakan teknik klasifikasi dalam machine learning yang memanfaatkan fungsi linier guna mengoptimalkan pemisahan antar kelas. Proses pelatihannya dilakukan melalui algoritma yang mengacu pada prinsip optimasi, serta mengintegrasikan bias pembelajaran berdasarkan teori pembelajaran statistik [18]. Tujuan utama dari penerapan SVM adalah untuk menentukan hyperplane yang paling optimal, yaitu bidang pemisah yang mampu memaksimalkan margin antar kelas sehingga meningkatkan akurasi dalam proses klasifikasi. [24].



Gambar 2. 2 Hyperplane SVM

Sumber: [24]

Untuk mendapatkan garis *hyperplane* yang optimal dan yang dapat memisahkan kelas dalam data, langkah pertama adalah menghitung *margin hyperplane* dan mencari titik maksimumnya. Garis *hyperplane* dihitung dengan menggunakan persamaan berikut:

$$(w * x) + b = 0$$

Rumus 2. 1 Rumus Hyperplane

Pada data x yang termasuk dalam kelas -1, hubungannya dapat dinyatakan seperti berikut:

$$(w * x + b) \leq 1, y_i = 1$$

Rumus 2. 2 Rumus Hyperplane (2)

Sedangkan pada data Xi yang termasuk dalam kelas +1, formulanya dapat dijelaskan sebagai berikut:

- “x" adalah titik data input dari Support Vector Machine.
- "y" adalah pemisahan antara garis hyperplane dan titik data terdekat untuk vektor bobot w dan nilai bias b yang diberikan.
- "b" adalah nilai bias yang merupakan tolak ukur dari hyperplane yang sedang dicari.

2.3.3 XGBoost (Xtreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) merupakan salah satu algoritma machine learning yang menggunakan struktur pohon keputusan. Algoritma ini dirancang untuk meningkatkan ketepatan prediksi dan mempercepat proses komputasi [19]. Dalam metode ini, pendekatan boosting digunakan, yaitu dengan membentuk model secara bertahap. Beberapa pohon keputusan yang lemah digabungkan agar menjadi model yang lebih kuat. Pohon yang dibuat selanjutnya bertugas untuk memperbaiki kesalahan dari pohon sebelumnya, sehingga kinerja model menjadi lebih baik secara keseluruhan [33]. ini bertujuan untuk memperkirakan kemungkinan kejadian di masa depan. Metode ini didasarkan pada asumsi sederhana yang mengatakan bahwa atribut-atribut adalah independen secara bersyarat jika kita mengetahui *nilai output-nya*. Berikut merupakan rumus 2.3 yaitu rumus XGBoosting.

$$\sum_{k=0}^{n} L(y_i, p_i) = \frac{1}{2}(y_i, p_i)^2$$

Rumus 2. 3 Rumus XGBoosting

Penjelasan mengenai rumus tersebut:

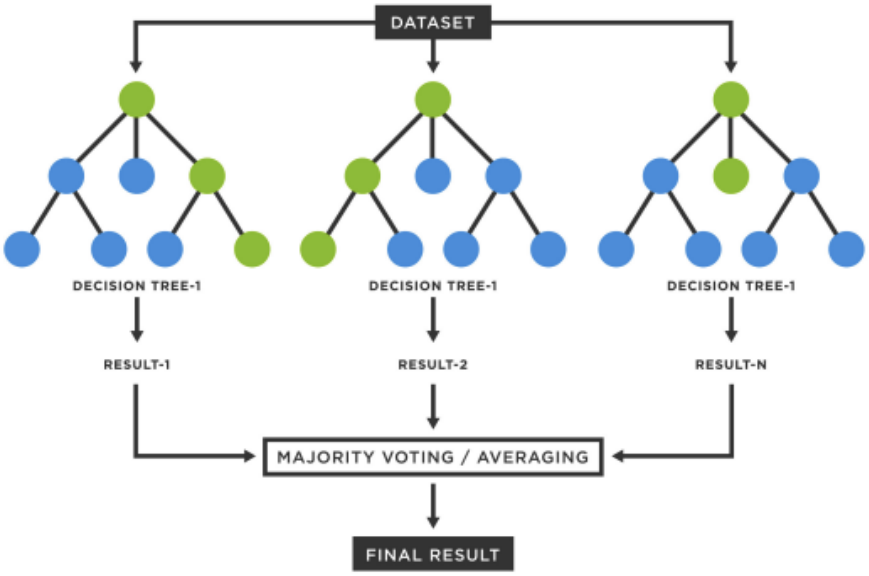
- n = Jumlah data.
- y_i = nilai sebenarnya.
- p_i = nilai prediksi dari model
- $L(y_i, p_i)$ = nilai kerugian (loss) untuk data ke-iii

Dalam proses pelatihannya, XGBoost menggunakan fungsi objektif yang mengintegrasikan fungsi kerugian seperti kuadrat selisih antara nilai aktual dan prediksi dengan komponen regularisasi untuk mengendalikan kompleksitas model [34]. Hal ini menjadikan XGBoost tidak hanya unggul dalam hal performa prediksi, tetapi juga efisien dalam komputasi serta mampu mengurangi risiko overfitting.

2.3.4 Random Forest

Random Forest adalah salah satu metode ensemble learning yang memanfaatkan pendekatan *bagging* untuk membangun sejumlah *decision tree* berdasarkan sampel data yang diambil secara acak (*bootstrapped*). Dalam konteks *ensemble learning*, pendekatan ini melibatkan pelatihan beberapa model pembelajar (*learners*) untuk menyelesaikan satu permasalahan yang sama secara kolektif [29]. Teknik *bagging* secara khusus menciptakan berbagai model serupa dari subset data yang bervariasi, yang kemudian digabungkan untuk menghasilkan prediksi yang lebih andal dan akurat dibandingkan dengan model tunggal [20]. Keunggulan utama dari algoritma ini terletak pada kemampuannya dalam mengelola dataset berukuran besar, mengenali fitur yang relevan secara efisien, menangani data yang tidak lengkap, serta menurunkan risiko *overfitting*. Hal ini dicapai melalui penggabungan sejumlah *decision tree*

yang cenderung *overfit* jika berdiri sendiri, tetapi ketika digabungkan, justru menghasilkan model prediksi yang lebih stabil. Penentuan kelas akhir dilakukan berdasarkan hasil pemungutan suara (*voting*) mayoritas dari setiap pohon keputusan yang dilatih dalam proses ensemble tersebut.



Gambar 2. 3 Cara kerja Random Forest

2.4 Tools dan Software Penelitian

2.4.1 Jupyter Notebook

Jupyter Notebook merupakan *tool open-source* berbasis web yang digunakan secara luas dalam proses analisis data dan eksperimen kode berbasis Python. Alat ini dipilih karena mendukung proses pengembangan secara interaktif, menyediakan fitur bagi pengguna agar dapat menulis serta menjalankan kode secara bertahap, serta menampilkan output langsung dalam satu dokumen. Dalam penelitian ini, Jupyter Notebook digunakan untuk mengimplementasikan algoritma klasifikasi seperti SVM, XGBoost dan Random Forest, sekaligus memvisualisasikan hasil evaluasi seperti *confusion matrix* dan grafik distribusi kelas. Kemampuannya untuk

menggabungkan kode, teks penjelasan, dan visualisasi menjadikannya alat yang sangat efisien dalam eksperimen *data science*.

Gambar 2. 4 Logo Jupyter

2.4.2 Python

Python dipilih sebagai bahasa pemrograman utama karena kemampuannya yang luas dalam bidang analisis data dan pembelajaran mesin. Python memiliki sintaksis yang sederhana dan berbagai library powerful yang dirancang khusus untuk *data science* [14]. Dalam penelitian ini, Python digunakan untuk membaca *dataset* AQI, membersihkan data, membangun model *machine learning*, serta mengevaluasi performanya [22]. Beberapa library yang digunakan meliputi:NumPy digunakan untuk operasi numerik dan manipulasi array.

- Pandas digunakan untuk analisis data dalam format tabel.
- Matplotlib digunakan untuk mempresentasikan data dalam bentuk yang mudah dipahami melalui diagram, plot, dan grafik yang dapat disesuaikan sesuai kebutuhan analisis.
- Seaborn memudahkan pembuatan grafik statistik yang kompleks dengan tampilan yang lebih menarik

serta menyediakan fungsi bawaan untuk analisis data yang terintegrasi secara harmonis dengan Matplotlib, sehingga memperkaya hasil visualisasi data.

- Scikit-learn, digunakan untuk machine learning.
- TensorFlow sebagai *framework* pembelajaran mesin dan deep learning.

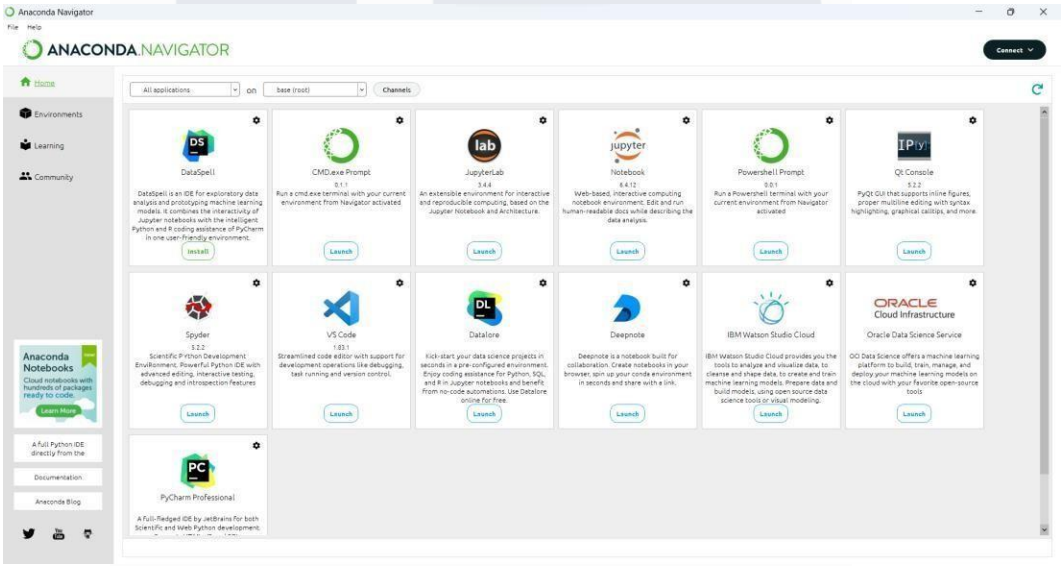
Oleh karena itu, Python dipandang sebagai bahasa pemrograman yang mampu mengakomodasi berbagai kebutuhan di bidang teknologi data, seperti pengolahan *Big Data*, penerapan *Data Science*, eksplorasi data melalui *Data Mining*, serta pengembangan model dalam *Machine Learning* hingga *Deep Learning*. [27]. Maka dari itu, Python dipilih karena kemampuannya dalam menyederhanakan seluruh proses, dari *preprocessing* hingga visualisasi, secara terintegrasi.

Gambar 2. 5 Logo Python

2.4.3 Anaconda

Anaconda digunakan karena menyediakan lingkungan manajemen paket dan distribusi Python yang terintegrasi secara lengkap. Alasan utama pemilihan Anaconda adalah karena paket ini sudah menyertakan sebagian besar *library* yang dibutuhkan, sehingga mempermudah instalasi dan pengelolaan lingkungan kerja. Selain itu, *Anaconda* menyediakan antarmuka pengguna seperti *Navigator*, yang memungkinkan pengguna

mengakses tools seperti *Jupyter Notebook* dan *Spyder* tanpa perintah terminal. Dalam penelitian ini, Anaconda digunakan sebagai platform utama untuk menjalankan *Jupyter Notebook* dan mengelola dependensi proyek secara efisien. Misalnya, instalasi semua pustaka *machine learning* dan visualisasi dalam penelitian ini dilakukan melalui *Anaconda Navigator* untuk efisiensi.



Gambar 2. 6 Anaconda Navigator