BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini berfokus pada analisis data kualitas udara di Kota Bekasi, dengan memeriksa empat parameter utama pencemaran udara yang meliputi: Karbon Monoksida (CO), Nitrogen Dioksida (NO_2), Partikulat Materi berukuran ≤2.5 mikrometer (PM2.5), dan Sulfur Dioksida (SO_2). Keempat parameter ini dipilih karena mereka merupakan indikator kritis yang secara langsung mempengaruhi kualitas udara dan kesehatan masyarakat, serta memberikan gambaran terkait tingkat pencemaran yang terjadi di wilayah tersebut.

Pemilihan kota Bekasi sebagai lokasi penelitian didasarkan pada beberapa faktor. Pertama, berdasarkan data yang ada, Bekasi secara konsisten termasuk dalam daftar kota dengan kualitas udara terburuk di Indonesia. Kondisi ini dapat dikaitkan dengan sejumlah faktor penyebab, antara lain tingginya tingkat polusi yang dihasilkan oleh aktivitas industri, padatnya arus lalu lintas kendaraan bermotor, serta proses urbanisasi yang pesat yang tidak diimbangi dengan pengelolaan lingkungan yang efektif. Kondisi ini menjadikan Kota Bekasi sebagai area yang sangat relevan untuk dianalisis, guna memberikan gambaran nyata tentang dampak kualitas udara terhadap kesehatan masyarakat dan lingkungan.

Sebagai tambahan, analisis terhadap kualitas udara di daerah ini sangat penting karena Bekasi merupakan bagian dari wilayah metropolitan Jakarta, yang memiliki populasi besar dan aktivitas ekonomi yang sangat tinggi. Ketergantungan pada kendaraan bermotor dan industri menyebabkan kota ini sering kali menghadapi polusi udara yang melebihi batas aman, yang dapat berpotensi menyebabkan masalah kesehatan jangka panjang bagi penduduknya.

Dalam penelitian ini, sumber data yang digunakan berasal dari platform daring *Air Quality Index* (AQI), yang dikenal menyediakan informasi kualitas udara secara

waktu nyata untuk berbagai wilayah di dunia. Data tersebut dikumpulkan melalui perangkat pemantauan udara *AirVisual Outdoor*, yang bekerja secara otomatis dalam merekam dan mengirimkan data polusi udara dari sejumlah titik pemantauan di wilayah Kota Bekasi. Rentang waktu pengumpulan data meliputi tanggal dari 1 Januari 2021 hingga 4 Mei 2025, dengan format data digital yang memungkinkan untuk diproses dan dianalisis secara sistematis dalam rangka mendukung penelitian ini.

3.4 Metode Penelitian

3.1.1 Alur Penelitian

Setelah topik penelitian ditentukan, langkah awal yang dilakukan adalah proses pengumpulan data, di mana dalam konteks ini data diperoleh melalui situs AQI (*Air Quality Index*) yang menyediakan data kualitas udara berdasarkan alat pemantauan AirVisual Outdoor. Tahap pemahaman terhadap permasalahan (*business understanding*), yaitu melihat tingginya tingkat pencemaran udara di Kota Bekasi sebagai isu lingkungan yang serius. Kota Bekasi dipilih karena data menunjukkan bahwa wilayah ini secara konsisten memiliki kualitas udara yang buruk dibandingkan kota-kota lain di Indonesia. Berdasarkan permasalahan tersebut, ditetapkanlah tujuan penelitian, yaitu untuk menganalisis tren dan karakteristik polutan udara selama empat tahun terakhir menggunakan data sekunder yang bersumber dari situs *Air Quality Index* (AQI).

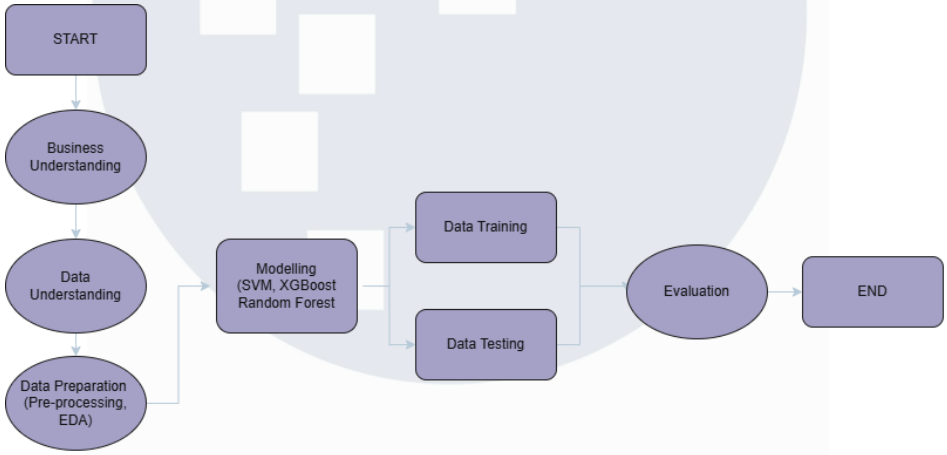
Selanjutnya, dilakukan tahap *data understanding* untuk mengenali struktur dan isi data yang diperoleh. Data mencakup parameter utama kualitas udara, yaitu Karbon Monoksida (CO), Nitrogen Dioksida (NO_2), Partikulat Materi (PM2.5), dan Sulfur Dioksida (SO_2), yang diambil dari alat pemantau *AirVisual Outdoor*. Tahapan ini mencakup identifikasi jenis data, satuan ukur, rentang waktu, serta frekuensi pencatatan data.

Setelah data dikenali, masuk ke tahap *data preparation*, yang mencakup proses pembersihan data dari nilai kosong (*missing values*), *outlier*, serta pemilihan atribut yang relevan. Tahapan ini juga melibatkan proses *exploratory data analysis* (EDA), yaitu eksplorasi data secara visual dan statistik untuk menemukan pola, tren, dan anomali.

Setelah melalui tahap pembersihan dan analisis eksploratif, data kemudian dipisahkan ke dalam dua kelompok utama: data pelatihan (*training*) dan data pengujian (*testing*). Pemisahan ini dilakukan guna membangun model prediktif sekaligus menguji tingkat keakuratannya. Dalam proses pemodelan, penelitian ini mengaplikasikan beberapa algoritma pembelajaran mesin, di antaranya adalah Support Vector Machine (SVM), XGBoost serta Random Forest, yang masing-masing digunakan untuk membandingkan performa dalam memprediksi kualitas udara. Model dibangun menggunakan data *training*, kemudian diuji menggunakan data *testing* untuk mengetahui performanya dalam mengklasifikasikan atau memprediksi kualitas udara berdasarkan parameter-parameter yang telah ditentukan sebelumnya.

Langkah terakhir dalam alur penelitian ini adalah proses evaluasi model yang telah dibangun. Evaluasi dilakukan dengan menggunakan metrik evaluasi tertentu, seperti akurasi, *precision*, *recall*, atau *Root Mean Square Error* (RMSE), tergantung pada jenis model dan pendekatan yang digunakan. Jika model belum memberikan hasil yang optimal, dilakukan penyesuaian parameter atau pemilihan algoritma lain agar hasil prediksi menjadi lebih akurat. Setelah model mencapai performa terbaik, hasil akhir divisualisasikan dalam bentuk grafik atau plot untuk memberikan gambaran yang lebih jelas mengenai pola pencemaran udara di Kota Bekasi selama periode yang diteliti.

Data yang diunduh merupakan kumpulan nilai polusi udara harian dari berbagai parameter seperti CO, NO_2, PM2.5, dan SO_2 selama periode 2021 hingga 2025. Pada penelitian yang dilakukan, terdapat beberapa tahap metodologi yang dilakukan, diantaranya adalah *Business Understanding, Data Understanding, Data Preparation (Pre-processing, EDA), Modelling, Data Training* dan *Data Testing*, *Evaluation*.



Tabel 3. 1 Flowchart Tahap Alur Penelitian

Sumber: [18]

3.1.2 Metode Data Mining

Penelitian ini menerapkan metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*), yaitu sebuah metodologi yang banyak digunakan dalam proses pengolahan data untuk menyelesaikan masalah di berbagai industri, termasuk dalam bidang lingkungan dan prediksi data. Metode ini dipilih karena memiliki struktur yang sistematis dan fleksibel, sehingga cocok digunakan dalam penelitian berbasis data time-series seperti data kualitas udara. CRISP-DM terdiri dari enam fase utama: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment*. Masing-masing tahapan ini saling berkaitan dan berperan penting dalam menghasilkan

model prediksi yang akurat dan dapat diterapkan untuk tujuan analisis jangka panjang.

1. Business Understanding

   Tahapan ini difokuskan untuk memperoleh pemahaman yang mendalam secara menyeluruh latar belakang dan tujuan penelitian. Tujuan utama penelitian adalah untuk mengklasifikasikan kualitas udara ke dalam beberapa kategori seperti “Baik”, “Sedang”, “Tidak sehat untuk Kelompok Sensitif”, “Tidak Sehat”, “Sangat Tidak Sehat” , hingga “Berbahaya”, dengan menggunakan parameter polusi utama seperti karbon monoksida (CO), nitrogen dioksida (NO_2), sulfur dioksida (SO_2), dan partikel halus berukuran 2.5 mikrometer (PM2.5). Dengan hasil klasifikasi tersebut, penelitian juga bertujuan untuk membandingkan performa tiga algoritma klasifikasi, yakni *Support Vector Machine (SVM*), *XGBoost* dan *Random Forest Classifier*, guna menentukan metode mana yang paling akurat dan efisien dalam memodelkan kualitas udara di wilayah Bekasi.guna mengetahui algoritma mana yang paling efektif dalam memodelkan data kualitas udara di Bekasi.

2. Data Understanding

   Tahapan ini mencakup proses pengumpulan dan eksplorasi awal terhadap data. Data diperoleh dari situs *Air Quality Index (AQI)* yang menyediakan informasi pemantauan kualitas udara berdasarkan perangkat *AirVisual Outdoor*. Data yang dikumpulkan mencakup tanggal 1 Januari 2021 hingga 4 Mei 2025 dengan parameter CO, NO_2, PM2.5, dan SO_2. Pada tahap ini dilakukan pemahaman struktur data,

eksplorasi visual awal, serta identifikasi adanya *missing value*, *outlier*, atau distribusi data yang tidak seimbang antar kelas kategori kualitas udara. Tahapan ini penting untuk memahami konteks data sebelum dilakukan pemodelan lebih lanjut.

Tabel 3. 2 Sample Data Kualitas Udara Kota Bekasi Tahun 2021 – 2025

date	CO	NO2	O3	PM10	PM2.5	SO2	count
1/1/2021	43	11	58	29	35	65	65
1/2/2021	58	11	86	38	64	80	86
1/3/2021	64	11	93	25	62	86	93
1/4/2021	50	24	67	24	31	77	77
1/5/2021	59	23	89	24	35	77	89
1/6/2021	73	23	81	29	66	85	85
1/7/2021	36	24	52	22	55	72	72
1/8/2021	38	24	68	26	51	71	71
1/9/2021	60	51	77	34	42	80	80
1/10/2021	24	12	39	16	38	59	59

3. Data Preparation

   Data yang telah dikumpulkan kemudian dipersiapkan agar dapat digunakan dalam proses pemodelan. Kegiatan yang dilakukan pada tahap ini meliputi pembersihan data (*data cleaning*), normalisasi nilai-nilai parameter, penggabungan atribut yang relevan, serta pemilihan fitur (*feature selection*) Setelah data disiapkan, data akan dibagi menjadi data pelatihan dan data pengujian untuk mendukung proses validasi model.

4. Modeling

   Setelah data siap, tahap selanjutnya adalah proses penerapan algoritma klasifikasi terhadap data yang telah dipersiapkan.

Dalam penelitian ini digunakan tiga model utama yaitu: *Support Vector Machine (SVM), XGBoost dan Random Forest Classifier*. SVM digunakan karena kemampuannya dalam menemukan *hyperplane* terbaik dalam memisahkan kelas, terutama untuk data non-linear [34]. XGBoost dipilih karena kemampuannya dalam menangani data yang kompleks, mengoptimalkan akurasi melalui teknik boosting, serta memiliki performa komputasi yang efisien. Selain itu, XGBoost terbukti unggul dalam meminimalkan error dan *overfitting*, sehingga menghasilkan model yang lebih stabil dan presisi tinggi pada data klasifikasi berbasis AQI. Random Forest dipilih dalam penelitian ini karena dikenal mampu mengatasi masalah *overfitting* serta memberikan hasil prediksi yang konsisten dan andal dengan memanfaatkan *ensemble* dari banyak pohon keputusan. Untuk membangun model, proses pelatihan dilakukan terlebih dahulu menggunakan *data training*, sedangkan evaluasi performa klasifikasi dilakukan pada *data testing*. Selain itu, untuk memastikan bahwa model memiliki tingkat akurasi dan kestabilan yang baik terhadap data yang bervariasi, digunakan pula pendekatan validasi silang (*cross-validation*) sebagai bagian dari strategi pengujian yang lebih menyeluruh.

5. Evaluation

   Pada tahap ini, model yang telah dilatih akan dievaluasi performanya dengan menggunakan metrik seperti akurasi, *precision*, *recall*, dan *F1-score*. Tujuannya adalah untuk membandingkan efektivitas ketiga algoritma dalam

mengklasifikasikan data kualitas udara secara objektif. Tahapan ini menjadi dasar dalam menentukan model mana yang paling tepat untuk digunakan dalam konteks data Kota Bekasi.

6. Deployment

   Pada tahap ini, hasil klasifikasi ditampilkan dalam bentuk visualisasi seperti grafik, *confusion matrix*, dan *dashboard interaktif* agar mudah dipahami. Informasi ini diharapkan dapat membantu pihak terkait seperti pemerintah daerah dalam mengambil kebijakan berbasis data terkait pengendalian polusi udara, serta memberikan informasi yang berguna bagi masyarakat umum mengenai kondisi kualitas udara di wilayah tempat tinggal mereka.

3.4 Teknik Pengumpulan Data

Penelitian ini menggunakan pendekatan metode kuantitatif, dengan sumber data sekunder yang diperoleh melalui situs *Air Quality Index* (AQI). Data yang digunakan berupa data kualitas udara dari Kota Bekasi, yang mencakup parameter penting seperti Karbon Monoksida (CO), Nitrogen Dioksida (NO_2), Partikulat Materi (PM2.5), dan Sulfur Dioksida (SO_2). Data tersebut diambil dari pengukuran yang dilakukan oleh perangkat *AirVisual Outdoor* dan telah tersedia dalam format digital yang siap untuk diunduh. Rentang waktu pengambilan data mencakup periode 1 Januari 2021 hingga 5 Mei 2025, dan data ini telah melalui proses kurasi oleh penyedia sehingga dapat langsung dimanfaatkan sebagai bahan analisis kuantitatif.

Metode kuantitatif ini diterapkan dengan memanfaatkan data historis dari AQI untuk membangun model klasifikasi tingkat kualitas udara. Setiap parameter akan dikonversi menjadi bentuk numerik yang dapat dianalisis, dan nilai-nilai tersebut kemudian akan digunakan untuk memprediksi kategori

kualitas udara seperti "Baik", "Sedang", "Tidak Sehat untuk Kelompok Sensitif", “Tidak Sehat”, “Sangat Tidak Sehat”, dan "Berbahaya". Pemilihan kota Bekasi sebagai lokasi penelitian didasarkan pada tingkat polusi udara yang tinggi serta konsistensi datanya di platform AQI selama beberapa tahun terakhir. Data yang telah dikumpulkan selanjutnya akan dianalisis menggunakan beberapa algoritma klasifikasi untuk menentukan metode yang paling akurat.

3.4 Teknik Analisis Data

Berdasarkan metodologi penelitian yang telah ditentukan dan akan digunakan, yaitu dengan *framework* CRISP-DM serta penerapan algoritma *Support Vector Machine (SVM), XGBOOST* dan *Random Forest Classifier*, maka diperlukan sejumlah tools pendukung untuk menjalankan seluruh proses klasifikasi dan evaluasi. *Tools* utama yang digunakan dalam penelitian ini adalah *Python* dengan bantuan *Jupyter Notebook*, yang akan digunakan mulai dari tahap eksplorasi dan pemrosesan data sekunder yang diperoleh dari situs *Air Quality Index* (AQI) hingga ke tahap pembangunan model klasifikasi dan visualisasi hasil evaluasi.

Data sekunder yang digunakan adalah data kualitas udara Kota Bekasi, yang mencakup parameter seperti PM2.5, NO_2, SO_2, dan CO dalam kurun waktu 1 Januari 2021 hingga 4 Mei 2025. Data ini telah tersedia dalam bentuk numerik dan dapat langsung diolah untuk kebutuhan klasifikasi kualitas udara ke dalam kategori seperti "Baik", "Sedang", "Tidak Sehat untuk Kelompok Sensitif", dan lainnya. Proses awal mencakup pembersihan data (*data cleaning*), penanganan *missing value* (jika ada), dan normalisasi data agar dapat diproses secara optimal oleh algoritma klasifikasi.

Selanjutnya, ketiga algoritma yaitu *SVM*, *XGBOOST* dan *Random Forest Classifier* akan dibangun dan diuji pada *dataset* yang sama. Model akan dilatih dengan *data training*, lalu pengujian dilakukan menggunakan *data testing*.

untuk memperoleh performa prediktif yang obyektif. Proses ini akan diulang dalam bentuk *k-fold cross validation* untuk memastikan stabilitas model. Setiap algoritma akan dianalisis kinerjanya menggunakan beberapa metrik, antara lain *accuracy*, *precision*, *recall*, dan F1-score, guna mengetahui seberapa baik model dalam mengklasifikasikan kualitas udara di Kota Bekasi.

Hasil dari ketiga algoritma akan dibandingkan untuk melihat model mana yang memberikan kinerja terbaik dalam memetakan kualitas udara berdasarkan data historis AQI. Visualisasi perbandingan metrik evaluasi juga akan disajikan dalam bentuk *confusion matrix*, *classification report*, dan grafik bar atau *line* untuk membantu interpretasi hasil secara intuitif. Dengan proses ini, diharapkan dapat diperoleh algoritma yang paling optimal dan dapat dijadikan dasar bagi pengambilan kebijakan lingkungan yang lebih baik di wilayah Kota Bekasi.

3.5 Teknik Pengujian

Dalam penelitian ini, dilakukan berbagai pengujian untuk mengetahui seberapa baik algoritma klasifikasi bekerja dalam mengelompokkan kualitas udara berdasarkan data parameter polusi seperti PM2.5, PM10, CO, NO_2, dan SO_2. Teknik pengujian digunakan untuk mengevaluasi performa model secara kuantitatif, sehingga hasil yang diperoleh tidak hanya berdasarkan asumsi, tetapi terukur secara objektif. Berikut adalah beberapa teknik evaluasi yang digunakan:

Beberapa teknik pengujian digunakan dalam penelitian ini, antara lain:

1. *Confusion Matrix*

   Confusion matrix digunakan untuk melihat secara langsung kinerja model dalam melakukan klasifikasi. Matriks ini memperlihatkan jumlah prediksi yang benar dan salah dari masing-masing kelas. Dengan menggunakan *confusion matrix*, dapat dilihat seberapa banyak data yang berhasil diklasifikasikan dengan benar (*True Positive* dan *True Negative*), serta jumlah kesalahan klasifikasi (*False Positive* dan *False Negative*) untuk

setiap kategori udara seperti Baik, Sedang, Tidak Sehat, dan Sangat Tidak Sehat.

2. Akurasi

   Akurasi adalah ukuran dasar yang menunjukkan seberapa banyak prediksi model yang benar dari seluruh data yang diuji. Nilai akurasi yang tinggi menandakan bahwa model mampu melakukan klasifikasi secara umum dengan baik. Namun, akurasi saja tidak cukup, apalagi jika data tidak seimbang.

3. *Precision*, *Recall*, dan F1-Score

   Ketiga metrik ini digunakan untuk memberikan penilaian yang lebih mendalam terhadap performa model, terutama dalam konteks klasifikasi dengan beberapa kelas. Precision mengukur seberapa tepat model dalam memprediksi suatu kelas (berapa banyak yang benar dari yang diprediksi). *Recall* menunjukkan seberapa banyak kasus sebenarnya yang berhasil ditemukan oleh model. Sedangkan F1-Score merupakan rata-rata harmonis dari *precision* dan *recall*, yang berguna ketika ingin menyeimbangkan antara ketepatan dan kelengkapan prediksi.

4. *K-Fold Cross Validation*

   Untuk memastikan bahwa hasil evaluasi tidak bergantung pada satu set data uji saja, digunakan teknik *K-Fold Cross Validation*. Dalam penelitian ini, digunakan *5-fold cross validation*, di mana data dibagi menjadi lima bagian, dan setiap bagian akan bergantian menjadi data uji, sementara empat sisanya digunakan untuk melatih model. Teknik ini memberikan hasil evaluasi yang lebih stabil dan tidak bergantung pada satu subset data tertentu.

Melalui gabungan semua metode evaluasi ini, hasil dari ketiga algoritma *Support Vector Machine* (SVM), *XGBoost*, *Random Forest* dapat dibandingkan secara objektif. Pendekatan ini memberikan gambaran menyeluruh mengenai

kelebihan dan kelemahan masing-masing algoritma dalam tugas klasifikasi kualitas udara, dan mendukung kesimpulan yang kuat di akhir penelitian.