BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian adalah kualitas udara di wilayah DKI Jakarta, dengan focus pada penerapan algoritma *naïve bayes* dengan teknik SMOTE untuk klasifikasi kualitas udara berdasarkan data ISPU. Penelitian ini juga bertujuan untuk meningkatkan akurasi hasil klasifikasi serta membandingkan hasil klasifikasi dengan dan tanpa menggunakan teknik SMOTE.

3.2 Lokasi dan Waktu Penelitian

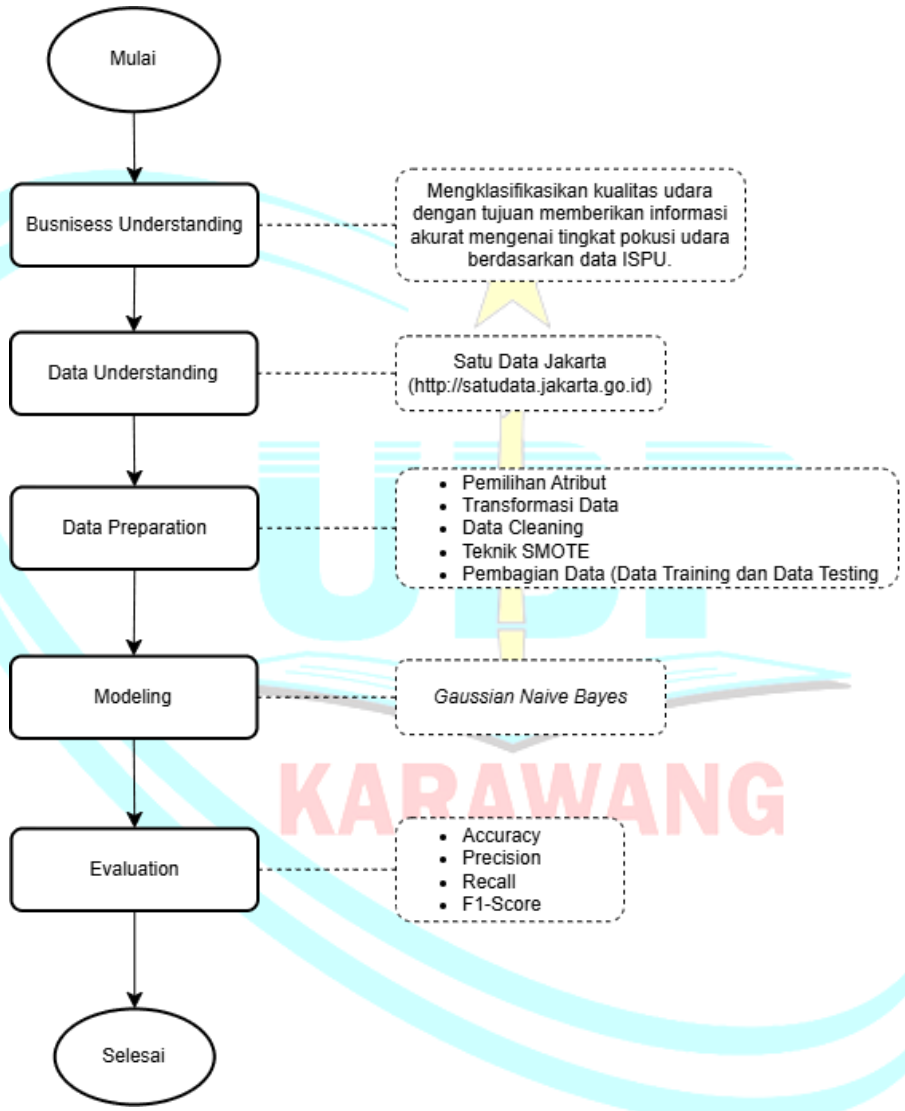
Lokasi penelitian dilakukan di DKI Jakarta. Penelitian dilakukan dengan memproses data yang telah dikumpulkan. Pengumpulan data berlangsung pada :

Tabel 3. 1 Waktu Penelitian 2024 hingga 2025

No	Tahap Penelitian	Bulan 1				Bulan 2				Bulan 3				Bulan 4			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	*Busnisess Understanding*																
2	*Understanding*																
3	*Preparation*																
4	*Modeling*																
5	*Evaluation*																

3.3 Prosedur Penelitian

Untuk memperoleh hasil yang diharapkan dalam penelitian, diperlukan langkah-langkah penelitian yang terstruktur. Berikut prosedur yang dilakukan digambarkan dalam Gambar 3.1



Gambar 3. 1 Prosedur Penelitian

3.3.1 Busnisess Understanding

Tahap pertama adalah memahami atau menganalisis permasalahn yang ingin dipecahkan dalam bisnis. Pada tahap ini, pemahaman dilakukan terhadap pemantauan kualitas udara DKI Jakarta secara menyeluruh, sehingga diperoleh informasi yang akurat.

3.3.2 Data Understanding

Tahap kedua mencakup pengumpulan, analisis dan penjelasan data, serta identifikasi masalah yang berkaitan dengan data. Proses pengumpulan data diperoleh dari website http://satudata.jakarta.go.id tentang data Indeks Standar Pencemaran Udara DKI Jakarta tahun 2022 hingga 2024. Data diambil dari lima stasiun pengukuran, yaitu DKI1 Bundaran HI, DKI2 Kelapa Gading, DKI3 Jagakarsa, DKI4 Lubang Buaya dan DKI5 Kebon Jeruk. Pemilihan periode ini dimaksudkan agar hasil penelitian mempresentasikan kondisi terkini. Data terdiri dari 3.506 baris dan 11 kolom yang mencakup tanggal pengukuran, nilai konsentrasi polutan, nilai ISPU tertinggi, parameter pencemaran dominan, kategori kualitas udara serta lokasi stasiun. Sebagian besar kolom polutan memiki data *object* dan perlu dikonersi menjadi numerik untuk keperluan analisis, selain itu terdapat nilai kosong dibeberapa kolom. Secara umum, data ini memiliki struktur yang cukup lengkap untuk dianalisis lebih lanjut, baik dari segi tren waktu, distribusi polutan, maupun klasifikasi kualitas udara berdasarkan lokasi dan kategori. Namun, dibutuhkan tahap praproses untuk memastikan data siap digunakan dalam analisis.

3.3.3 Data Preparation

Pada tahap ketiga, meliputi proses penyusunan dataset akhir yang akan digunakan sebagai input dalam pemodelan data mining. Proses penyiapan data dilakukan sebagai berikut:

1. Pemilihan atribut, yaitu merupakan proses pemilihan atribut yang paling relevan seperti PM10, PM2.5, SO2, CO, O3, NO2 serta kategori.
2. Transformasi data, yaitu proses mengubah format atau struktur data sesuai dengan pemodelan seperti rekategorisasi dan *encoding categorical feature*.
3. Data cleaning, merupakan proses pembersihan data yang berpotensi menggangu analisis pemodelan, seperti penanganan *missing value* dan *outlier.*
4. *Balancing* data, yaitu proses untuk menyamakan distribusi kelas dalam dataset sehungga jumlah sampel pada setiap kelas menjadi lebih seimbang, seperti teknik SMOTE yang diterapkan untuk menyeimbangkan distribusi dataset pada kelas minoritas sehingga jumlahnya sama dengan dataset kelas mayoritas.

5. Pembagian data, yaitu tahap memisahkan dataset menjadi data latih dan data uji dengan rasio tertentu seperti 80:20.

3.3.4 Modeling

Dalam tahap modeling diterapkan model algoritma yang dipilih dan diterapkan pada dataset. Proses pemodelan data mining yang diterapkan yaitu *naïve bayes gaussian*. Implementasi algoritma *naïve bayes gaussian* dilakukan pada data training untuk membangun model klasifikasi dengan menghitung probabilitas posterior berdasarkan asumsi independensi antara fitur – fitur yang digunakan. Dua pendekatan yang digunakan yaitu menggunakan teknik SMOTE untuk menanganu ketidakseimbangan data dan tanpa SMOTE. Pendekatan ini dilakukan untuk menganalisis perbedaan hasil sebagai bahan perbandingan sehingga dapat menentukan scenario dengan hasil terbaik.

Tabel 3. 2 Alur algoritma naive bayes gaussian + SMOTE

Algoritma	: Klasifikasi kualitas udara dengan *naïve bayes gaussian* + SMOTE
Input	: Training data + SMOTE
Output	: Hasil klasifikasi kualitas udara
	1. Terapkan teknik SMOTE untuk menyeimbangkan jumlah sampel antar kelas. 2. Gunakan data hasil augmentasi sebagai data training baru. 3. Hitung probabilitas prior untuk setiap kelas berdasarkan distribusi data training. 4. Asumsikan bahwa setiap fitur memiliki distribus *gaussian* (normal), hitung parameter distribusi (*mean dan varians*) untuk setiap fitur pada setiap kelas. 5. Hitung probabilitas *likelihood* berdasarkan fungsi distribusi normal. 6. Gunakan *teorema bayes* untuk menghitung probabilitas posterior untuk setiap kelas. 7. Prediksi kelas dengan memilih kelas dengan probabilitas posterior tinggi.

8. Gunakan model yang telah dilatih untuk mengklasifikasikan data testing. 9. Hasil klasifikasi kualitas udara diperoleh berdasarkan prediksi model *naïve bayes gaussian.*

Tabel 3. 3 Alur algoritma naive bayes gaussian

Algoritma : Klasifikasi kualitas udra dengan *naïve bayes gaussian*
Input : Training data Output : Hasil klasifikasi kualitas udara 1. Input data training yang telah dipisahkan sebelumnya digunakan sebagai masukan untuk model. 2. Hitung probabilitas prior untuk setiap kelas berdasarkan distribusi data training. 3. Asumsikan bahwa setiap fitur memiliki distribusi *gaussian* (normal), hitung parameter distribusi *(mean dan variana)* untuk setiap fitur pada setiap kelas. 4. Hitung probabilitas *likelihood* berdasarkan fungsi distribusi normal. 5. Gunakan *teorema bayes* untuk menghitung probabilitas posterior untuk setiap kelas. 6. Prediksi kelas dengan memilih kelas dengan probabilitas posterior tinggi 7. Gunakan model yang telah dilatih untuk mengklasifikasikan data testing. 8. Hasil klasifikasi kualitas udara diperoleh berdasarkan prediksi model *naïve bayes gaussian*

3.3.5 Evaluation

Pada tahap evaluasi dilakukan interpretasi dari hasil pemodelan data mining. Berdasarkan tujuan yang jelas pada tahap *business understanding,* maka pada tahap ini akan dilakukan analisa hasil klasifikasi data kualitas udara, evaluasi dilakukan dengan pengimplementasian *confusion matrix* dengan dua pendekatan,

menggunakan teknik SMOTE serta tanpa menggunakan teknik SMOTE sehingga akan mendapatkan hasil yang telah direncanakan. Hasil evaluasi akan ditampilkan dalam bentuk *accuracy*, *precision*, *recall* dan *F1-score* yang akan memberikan gambaran lebih jelas mengenai performa model yang akan digunakan.