



## BAB III

### METODOLOGI PENELITIAN



Hak cipta milik IBIKKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Institut Bisnis dan Informatika Kwik Kian Gie

#### A. Gambaran Umum Objek Penelitian

Jakarta, sebagai salah satu kota metropolitan terpadat di dunia, menghadapi tantangan serius dalam menjaga kualitas udara yang sehat untuk penduduknya. Pencemaran udara di Kota Jakarta merupakan masalah serius yang berkaitan dengan peningkatan pertumbuhan industri, transportasi yang padat, dan urbanisasi yang cepat. Tingkat polusi udara yang tinggi di kota ini dapat menyebabkan dampak negatif terhadap kesehatan penduduk, termasuk masalah pernapasan dan penyakit terkait polusi udara. Pencemaran udara di Jakarta juga dapat merusak lingkungan, mengurangi kualitas hidup, dan menimbulkan masalah sosial dan ekonomi. Oleh karena itu, pemahaman dan upaya pengelolaan pencemaran udara sangat penting untuk menjaga kualitas udara dan kesejahteraan penduduk di ibu kota Indonesia ini.

#### B. Metodologi Penelitian

##### 1. Teknik Pengumpulan Data

Pada penelitian ini yang digunakan adalah data sekunder yang diambil dari [saudata.jakarta.go.id](http://saudata.jakarta.go.id). Teknik pengumpulan data memiliki arti cara umum dalam mengumpulkan data, sedangkan instrumen pengumpul data merupakan alat untuk mengumpulkan data berdasarkan teknik yang digunakan. Alasan peneliti menggunakan data sekunder adalah untuk menghemat waktu dan sumber daya yang diperlukan untuk mengumpulkan data baru. Dalam mengumpulkan data, penulis menggunakan metode kuantitatif karena penelitian ini menggunakan data numerik dan

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



statistik untuk membandingkan performa dua algoritma dalam klasifikasi kualitas udara di Jakarta.

## 2. Teknik Analisis Data

Dalam penelitian ini, peneliti menggunakan teknik analisis yang dilakukan berdasarkan metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*). CRISP-DM menyediakan proses standar yang tidak berpemilik dan tersedia secara bebas agar dapat menyesuaikan hasil *data mining* ke dalam strategi pemecahan masalah umum bisnis atau unit penelitian. CRISP-DM memiliki 6 siklus yang fasenya, antara lain:

### a. *Business Understanding Phase*

- (1) Pertama, mengidentifikasi permasalahan yang ada.
- (2) Kemudian, terjemahkan pemahaman permasalahan ke dalam perumusan definisi masalah *data mining*.
- (3) Terakhir, siapkan strategi awal untuk mencapai solusi atau tujuan yang ada.

### b. *Data Understanding Phase*

- (1) Pertama mengumpulkan data.
- (2) Eksplorasi data untuk memahami struktur dan karakteristiknya.
- (3) Mengevaluasi kualitas data.
- (4) Terakhir, memilih variabel yang relevan untuk analisis.

### c. *Data Preparation Phase*

- (1) Pembersihan data dari nilai yang hilang atau outliers.
- (2) Integrasi data dari berbagai sumber jika diperlukan.
- (3) Melakukan transformasi variabel untuk memenuhi kebutuhan analisis.



d. *Modeling Phase*



Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)

- (1) Pemilihan model algoritma yang sesuai, seperti SVM dan KNN.
- (2) Melakukan eksperimen dengan parameter model untuk optimisasi.
- (3) Penerapan model pada data latihan dan pelatihan.

e. *Evaluation Phase*

- (1) Mengevaluasi performa model, kualitas, dan efektivitasnya.
- (2) Penentuan apakah model sesuai dengan tujuan penelitian.
- (3) Terakhir, buatlah keputusan mengenai penggunaan hasil *data mining*

f. *Development Phase*

- (1) Melakukan penerapan terhadap model yang sudah dibuat, seperti membuat laporan.
- (2) Persiapan dokumentasi dan presentasi hasil untuk pihak terkait.
- (3) Pemantauan dan pemeliharaan model dalam konteks implementasi.

### 3. Penerapan Algoritma

Pada penelitian ini akan dibahas bagaimana penggunaan komparasi algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) untuk klasifikasi kualitas udara Kota Jakarta dapat digambarkan pada diagram sebagai berikut:

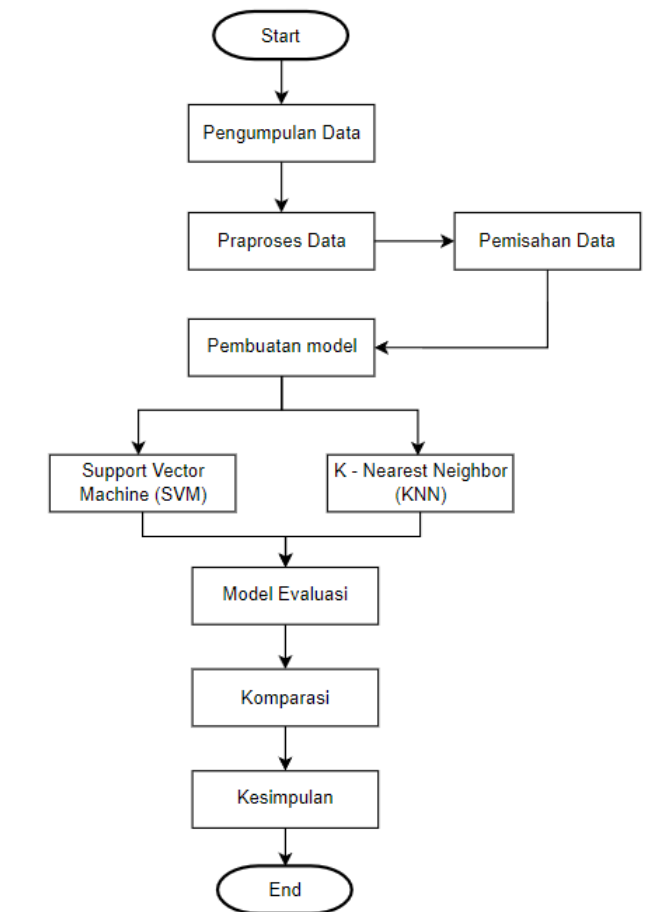
Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



**Gambar 3.1**  
**Workflow Penerapan Algoritma**

Sumber: Olahan Penulis.

Gambar 3.1 merupakan workflow yang digunakan oleh penulis untuk melakukan penelitian ini. Berikut adalah penjelasannya:

**a. Pengumpulan Data**

Pengumpulan data diambil melalui website [satudata.jakarta.go.id](http://satudata.jakarta.go.id).

**b. Praproses Data**

Proses data ini meliputi pembersihan data dan transformasi data menjadi data yang berkualitas untuk dilakukan tahapan data mining.

**© Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)**

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

**c. Pemisahan Data**

Membagi data menjadi dua set, yaitu set pelatihan dan set pengujian. Set pelatihan digunakan untuk melatih model algoritma. Set pengujian digunakan untuk mengevaluasi kinerja model algoritma.

**d. Pembuatan Model**

Membentuk model algoritma data mining menggunakan set pelatihan.

**e. Support Vector Machine (SVM)**

Data diolah menggunakan model *Support Vector Machine* dengan bahasa pemrograman *Python*.

**f. K-Nearest Neighbor (KNN)**

Data diolah menggunakan model *K-Nearest Neighbor* dengan bahasa pemrograman *Python*.

**g. Model Evaluasi**

Melakukan Evaluasi dan mengkomparasi skor hasil pengolahan.

**h. Komparasi**

Melakukan komparasi keakuratan algoritma SVM dan KNN menggunakan *Python* pada klasifikasi kualitas udara Kota Jakarta.

**i. Kesimpulan**

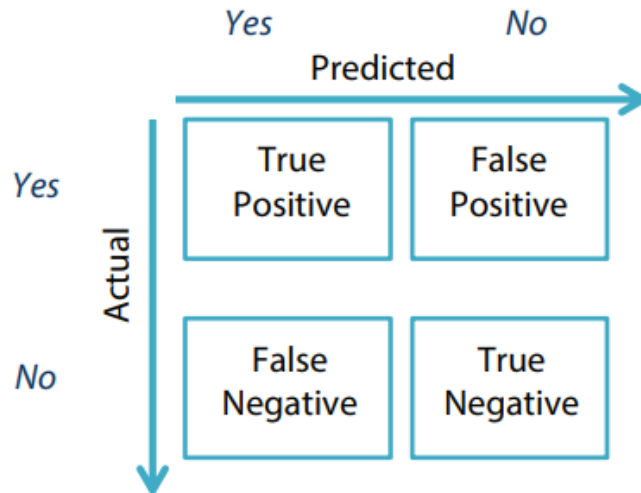
Kesimpulan dari hasil komparasi pada masing-masing algoritma.

**4. Teknik Pengukuran Data**

**a. Menghitung Nilai Accuracy**

Secara umum *Accuracy* dapat dirumuskan sebagai berikut:

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



**Gambar 3.2**  
**Confusion Matrix**

Sumber: Rohit Raja (2022:203).

Cara menghitung nilai *Accuracy* :

(1) *Accuracy*

Akurasi didefinisikan sebagai tingkat kedekatan antara nilai dan prediksi dengan nilai aktual

$$Accuracy : \frac{TP}{TP + TN + FN + FP}$$

b. **Algoritma Support Vector Machine**

Pada algoritma SVM terdapat trik *kernel* yaitu metode untuk menghitung kesamaan ini sebagai fungsi dari atribut aslinya. Dalam penelitian ini, fungsi *kernel* digunakan untuk perhitungan. Fungsi *kernel* yang umum digunakan adalah sebagai berikut:

© Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

### (1) Linear Kernel

$$K(x, y) = x y$$

Keterangan:

K = fungsi *kernel*

x, y = Vektor

### (2) Polynomial Kernel

$$K(x_1, y_1) = (x_1 y_1 + 1)^d$$

Keterangan:

K(x,y) = hasil *kernel* antara dua vektor input x dan y.

d = derajat polinomial

### (3) Radial Basis Function Kernel.

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

Keterangan:

K(x,y) adalah hasil *kernel* antara dua vektor input x dan y, exp adalah fungsi eksponensial, yang merupakan fungsi dasar dari operasi pangkat *e* (bilangan Euler),  $\|x - y\|$  adalah norma Euclidean (jarak Euclidean) antara vektor *x* dan *y* yang dihitung sebagai akar kuadrat dari jumlah kuadrat perbedaan elemen-elemen vektor, dan  $\sigma$  adalah parameter *bandwidth* yang dapat diatur untuk mengontrol seberapa luas distribusi *kernel*.

### (4) Sigmoid Kernel

$$K(x, y) = \tanh(ax^T y + c)$$

Keterangan:

K(x,y) adalah hasil *kernel* antara vektor input x dan y, *tanh* adalah fungsi



tangen hiperbolik, yang menghasilkan nilai antara -1 dan 1, a dan c adalah parameter yang dapat diatur untuk mengontrol bentuk dan sifat dari *kernel*.

**© Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)**

Hak Cipta Dilindungi Undang-Undang

**Institut Bisnis dan Informatika Kwik Kian Gie**

**c. Algoritma K-Nearest Neighbor**

Pada langkah dari algoritma K-Nearest Neighbors (KNN) yaitu hitung jarak antara data baru dengan semua data pada dataset. Jarak yang digunakan dalam penelitian ini adalah jarak *Euclidean* dan jarak *Manhattan*. Berikut adalah formula dari setiap jarak:

**(1) Euclidean distance**

Menurut Parteek Bhatia (2019:157) Jarak Euclidean paling umum digunakan untuk menghitung jarak. Berikut adalah formula untuk *euclidean distance*:

$$\text{distance}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Keterangan:

$\text{distance}(x, y)$  = jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi.

$(x_i - y_i)^2$  = Merupakan kuadrat dari selisih antara komponen ke-i dari vektor x dan y. Ini mengukur perbedaan pada setiap dimensi.

**(2) Manhattan distance**

Menurut Parteek Bhatia (2019:193) Jarak Manhattan didefinisikan sebagai jumlah panjang proyeksi ruas garis antara dua titik pada sumbu koordinat. Berikut adalah formula untuk *manhattan distance*:

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.





1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

$$\text{distance}(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Keterangan:

$\text{distance}(x, y)$  = jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi.

$|x_i - y_i|$ : Menunjukkan nilai absolut dari selisih antara komponen ke-i dari vektor x dan y. Ini mencerminkan panjang proyeksi ruas garis (jarak) antara titik-titik pada sumbu koordinat pada dimensi tertentu.

#### d. Indeks Standar Pencemaran Udara

Menurut Peraturan Menteri Lingkungan Hidup Dan Kehutanan Republik Indonesia Nomor P.14/Menlhk/Setjen/Kum.1/7/2020 Tentang Indeks Standar Pencemar Udara Pasal 1 menyatakan bahwa Indeks Standar Pencemar Udara yang selanjutnya disingkat ISPU adalah angka yang tidak mempunyai satuan yang menggambarkan kondisi mutu udara ambien di lokasi tertentu, yang didasarkan kepada dampak terhadap kesehatan manusia, nilai estetika dan makhluk hidup lainnya. Terdapat penentuan kategori dalam ISPU antara lain:

##### (1) Kategori Angka Rentang ISPU

**Tabel 3.1**  
**Kategori Angka Rentang ISPU**

Kategori	Angka Rentang
Baik	1 - 50
Sedang	51- 100
Tidak Sehat	101 - 200
Sangat Tidak Sehat	201 - 300
Berbahaya	$\geq 301$



Sumber: Peraturan Menteri Lingkungan Hidup Dan  
Kehutanan Republik Indonesia (2020)

## (2) Penjelasan Nilai ISPU

**Tabel 3.2**

**Penjelasan Nilai ISPU**

Kategori	Keterangan
Baik	Tingkat kualitas udara yang sangat baik, tidak memberikan efek negatif terhadap manusia, hewan, tumbuhan.
Sedang	Tingkat kualitas udara masih dapat diterima pada kesehatan manusia, hewan dan tumbuhan.
Tidak sehat	Tingkat kualitas udara yang bersifat merugikan pada manusia, hewan dan tumbuhan.

Sumber: Peraturan Menteri Lingkungan Hidup Dan  
Kehutanan Republik Indonesia (2020)

### C Hak cipta milik IBI KKG (Institut Bisnis dan Informatika Kwik Kian Gie)

Hak Cipta Dilindungi Undang-Undang

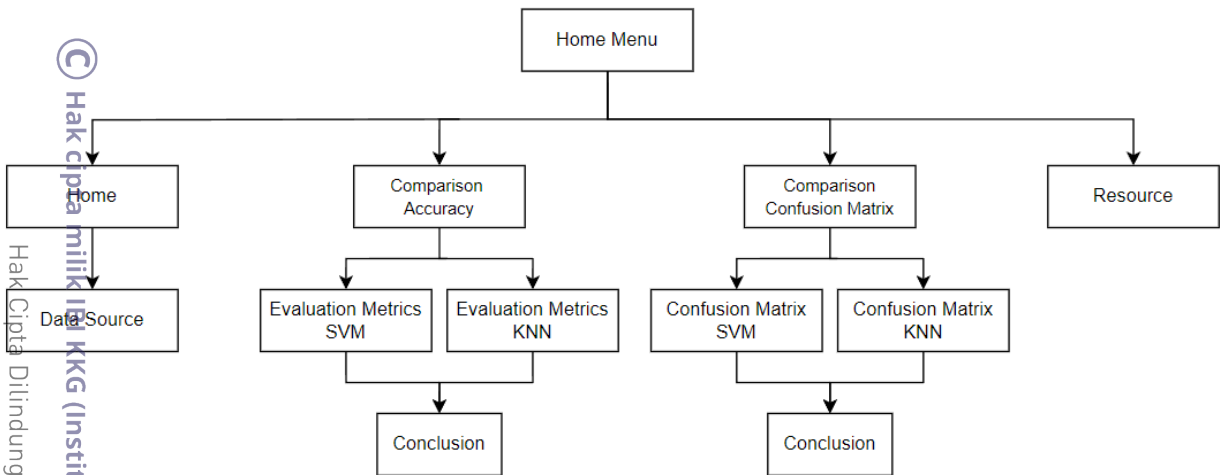
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.

### C. Teknik Perancangan Graphic User Interface (GUI)

Pada perancangan *Graphic User Interface* (GUI), peneliti menggunakan Google Sites untuk menyajikan hasil proses data mining dalam bentuk dashboard. Pada halaman website terdapat page home, Page Prediction, Page Comparison, dan Page Resource. Detail Rancangan GUI yang ada meliputi *Home, Prediction, Comparison, Resource*. Berikut gambar hirarki dari rancangan GUI tersebut:



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik dan tinjauan suatu masalah.
  - b. Pengutipan tidak merugikan kepentingan yang wajar IBIKKG.
2. Dilarang menggunakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IBIKKG.



**Gambar 3.3**  
**Hierarki Rancangan GUI**

Sumber: Olahan Penulis