BAB III
METODE PENELITIAN

3.1 Objek Penelitian

3.1.1. Studi Pustaka

Penulis menggunakan metode studi pustaka dalam proses pengumpulan data. Data yang diperoleh dari metode ini berfungsi sebagai dasar pendukung dalam pelaksanaan penelitian serta penyusunan skripsi. Studi pustaka yang dilakukan mencakup referensi dari sebanyak 18 jurnal ilmiah.

3.1.2. Deskripsi Data

Penilitian ini menggunakan dataset kualitas udara (Air Quality) di Yogyakarta tahun 2021. Data tersebut mencakup seluruh 12 bulan dalam satu tahun, dengan total sebanyak 8.784 entri. Data ini akan digunakan untuk melakukan klasifikasi terhadap kualitas udara, sebagaimana ditampilkan pada Gambar 3.1 dibawah ini

	Date	Time	PM10	PM2.5	SO2	CO	O3	NO2	Max	Critical Component	Category
0	1/1/2021	00:00:00	13.0	40.0	0.0	25.0	0.0	0.0	40	PM2.5	Good
1	1/1/2021	01:00:00	12.0	38.0	0.0	24.0	0.0	0.0	38	PM2.5	Good
2	1/1/2021	02:00:00	11.0	35.0	0.0	23.0	0.0	0.0	35	PM2.5	Good
3	1/1/2021	03:00:00	10.0	32.0	0.0	22.0	0.0	0.0	32	PM2.5	Good
4	1/1/2021	04:00:00	9.0	29.0	0.0	21.0	0.0	0.0	29	PM2.5	Good
...	...	...	...	...	...	...	...	...	...	...	...
8779	12/31/2021	19:00:00	15.0	0.0	10.0	21.0	1.0	11.0	21	CO	Good
8780	12/31/2021	20:00:00	16.0	0.0	10.0	21.0	1.0	11.0	21	CO	Good
8781	12/31/2021	21:00:00	17.0	0.0	10.0	21.0	1.0	11.0	21	CO	Good
8782	12/31/2021	22:00:00	18.0	0.0	10.0	21.0	1.0	11.0	21	CO	Good
8783	12/31/2021	23:00:00	18.0	0.0	10.0	21.0	1.0	11.0	21	CO	Good

8784 rows × 11 columns

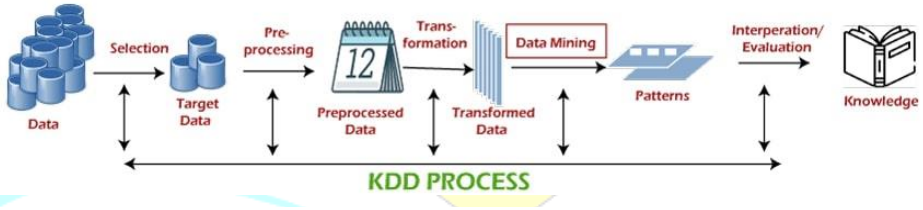
Gambar 3. 1. Deskripsi

Tahap diatas merupakan proses untuk melihat keseluruhan dataset Air Quality In Yogyakrta 2021. Pada hasil deskripsi data gambar 3.1 menampilkan $Date, Time, PM_{10}, PM_{2.5}, SO_2, CO, NO_2, Max, Critical\ Component, Category$ dan keseluruhan data Air Quality In Yogyakrta 2021 sebesar 8.784.

3.2 Prosedur Penelitian

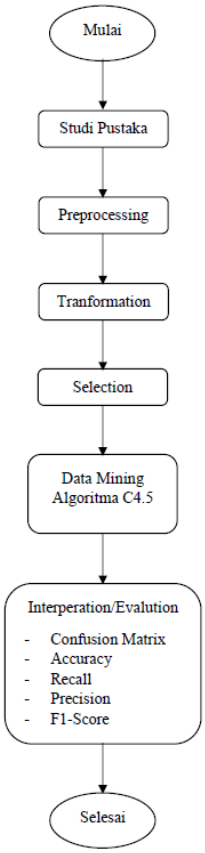
3.2.1. Metode Knowledge Discovery in Database (KDD)

Penelitian ini, penulis menerapkan metode atau tahapan KDD (*Knowledge Discovery In Database)* yang digambar secara ringkas pada Gambar 3.2 mengenai proses *KDD* proses.



Gambar 3. 2. Knowledge Discovery In Database

Pada gambar 3.2 peneliti menerapkan metode KDD, bertujuan untuk menggali wawasan mendalam dari data yang kompleks, menemukan pola-pola penting, dan menghasilkan pengetahuan baru yang dapat digunakan untuk pengambil keputusan.



Gambar 3. 3. Alur Penelitian

Pada Gambar 3.3, alur penelitian menunjukan bahwa peneliti mejalankan tahap secara sistematis dan terorganisir guna menjamin kualitas data yang baik serta memastikan keandalan hasil analisis yang diperoleh.

3.2.2. Preprocessing

Langkah berikutnya adalah tahap praproses (*preprocessing*), yang mencakup pembersihan data, seperti menghapus *noise*, memastikan tidak ada data yang terduplikasi, memeriksa inkonsitensi, serta memperbaiki kesalahan seperti salah penulisan. Hal ini dilakukan agar data siap digunakan dalam proses data mining. Contohnya dapat dilihat pada Gambar berikut:

1. Pemeriksaan terhadap nilai yang hilang (*missing value*) pada data Kualitas Udara di Yogyakarta ditampilkan pada Gambar 3.4 dibawah ini:

```
Missing Values Count:
Date                    0
Time                    0
PM10                  171
PM2.5                 171
SO2                  2398
CO                    171
O3                   3132
NO2                   171
Max                     0
Critical Component    171
Category                0
dtype: int64
```

Gambar 3. 4. Missing Value

Tahap diatas proses melihat *missing values* pada dataset Air Quality In Yogyakarta Pada Tahun 2021.

2. Pemeriksaan terhadap data duplikat pada data Kualitas Udara di Yogyakarta tahun 2021 dapat dilihat pada Gambar 3.5 berikut ini:

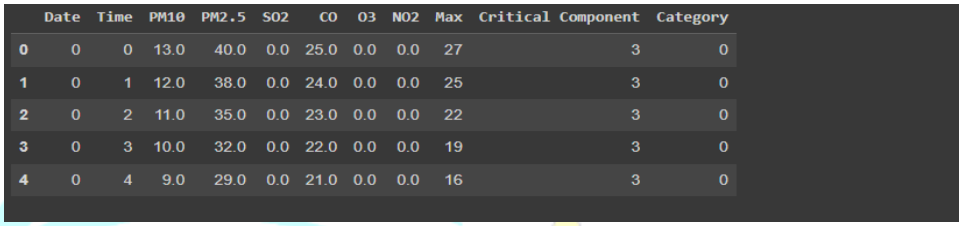
```
0.0
```

Gambar 3. 5. Duplikat

Langkah diatas menunjukan proses pengecekan duplikat pada dataset Kualitas Udara di Yogyakarta tahun 2021, dimana hasil dari pemeriksaan tersebut menunjukan bahwa tidak ada data duplikat (0.0 data duplikat).

3.2.3. Tranformation

Pada tahap *transformation*, data akan diubah menjadi format yang sesuai untuk diproses lebih lanjut pada tahap data mining. Pada tahap ini, hanya terjadi perubahan nama atribut yang akan digunakan. Contoh proses ini dapat dilihat pada Gambar 3.6 dibawah ini:



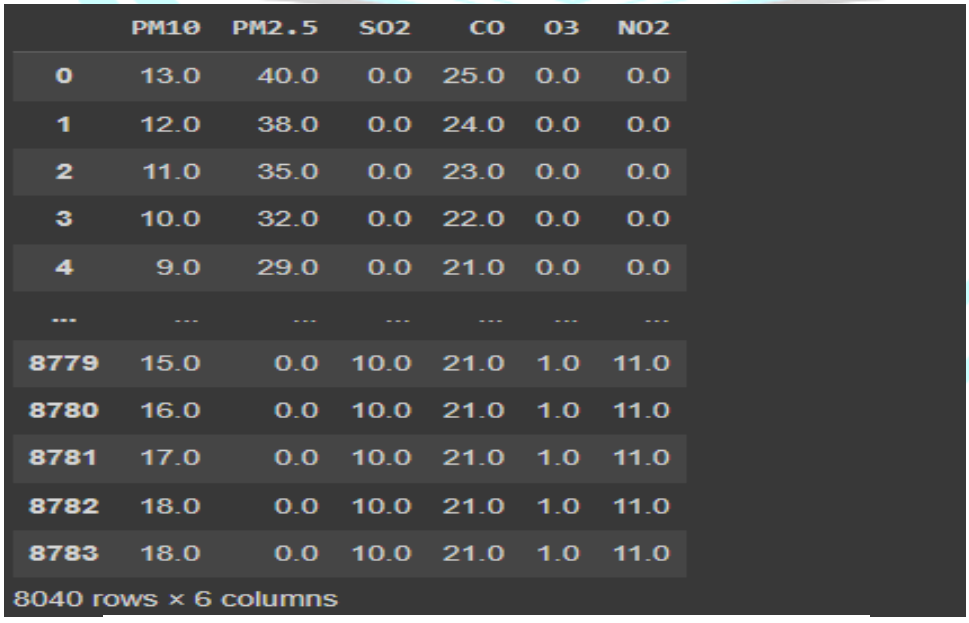
	Date	Time	PM10	PM2.5	SO2	CO	O3	NO2	Max	Critical	Component	Category
0	0	0	13.0	40.0	0.0	25.0	0.0	0.0	27		3	0
1	0	1	12.0	38.0	0.0	24.0	0.0	0.0	25		3	0
2	0	2	11.0	35.0	0.0	23.0	0.0	0.0	22		3	0
3	0	3	10.0	32.0	0.0	22.0	0.0	0.0	19		3	0
4	0	4	9.0	29.0	0.0	21.0	0.0	0.0	16		3	0

Gambar 3. 6. Tranformation

Pada tahap proses diatas merubah nilai kolom yang akan diklasifikasi gambar diatas merubah kolom "*Category*" dari *string* menjadi *numerik*.

3.2.4. Selection

Pada tahap *selection*, data akan disesuaikan agar siap untuk diproses pada tahap data mining berikutnya. Pada tahap ini, kolom atau fitur yang tidak relevan akan dihapus dari dataset. Contoh proses ini dapat dilihat pada Gambar 3.7 dibawah ini:



	PM10	PM2.5	SO2	CO	O3	NO2
0	13.0	40.0	0.0	25.0	0.0	0.0
1	12.0	38.0	0.0	24.0	0.0	0.0
2	11.0	35.0	0.0	23.0	0.0	0.0
3	10.0	32.0	0.0	22.0	0.0	0.0
4	9.0	29.0	0.0	21.0	0.0	0.0
...	...	...	...	...	...	...
8779	15.0	0.0	10.0	21.0	1.0	11.0
8780	16.0	0.0	10.0	21.0	1.0	11.0
8781	17.0	0.0	10.0	21.0	1.0	11.0
8782	18.0	0.0	10.0	21.0	1.0	11.0
8783	18.0	0.0	10.0	21.0	1.0	11.0

8040 rows × 6 columns

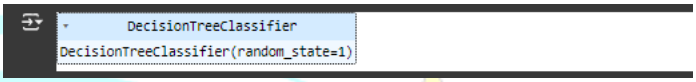
Gambar 3. 7. Selection

3.2.5. Data Mining

Pada tahap data mining, dataset Air Quality In Yogyakarta tahun 2021 dianalisis dan dihitung menggunakan algoritma C4.5 untuk membuat model

serta aturan dari pohon keputusan (*Decision Tree*). Pada proses ini, data dibagi menjadi dua bagian: data latih (*Traning*) dan data uji (*Testing*) dengan rasio pembagian 80:20, yaitu 80% untuk pelatihan dan 20% untuk pengujian. "Sebenarnya tidak ada batasan tetap mengenai proporsi pembagian antara data *traning* dan *testing*"(Jatmiko, et al., 2019). Proses ini dapat dilihat pad Gambar 3.8 berikut:

1. Tahap proses algoritma C4.5



Gambar 3. 8. Algoritma

Tahap diatas merupakan proses pembentukan model dengan menggunakan algoritma C4.5. Algoritma ini memiliki keunggulan dalam menangani data bertipe *numerik*, data yang mengandung *noise*, serta nilai yang hilang (*missing value*). Dalam algoritma ini, digunakan *gain ration* sebagai kriteria pemisah dalam membangun *decision tree*, menggantikan *information gain*. *Decision tree* dibentuk dengan prinsip pemisah data, termasuk data kontinu, dan mampu menangani ketidaksepurnaan data.

3.2.6. Interpretion/Evaluation

Pada tahap ini dilakukan proses evaluasi dan penafsiran terhadap *rule* yang telah diperoleh, disesuaikan dengan tujuan yang telah ditetapkan. Proses ini juga melibatkan penerapan data mining untuk melakukan pengujian serta penilaian. Pengujian dilakukan dengan menggunakan Confusion Matrix yang mencakup nilai *Accuracy*, *Recall*, dan *Precision*, sedangkan evaluasi kinerja algoritma dilakukan melalui analisis *kurva ROC* dan nilai *AUC*.