Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan *Open Government Data* Menggunakan Algoritma *Random Forest*

1st Adityo Nugroho
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
adityonugroho@students.telkomuniversity.ac.id

2nd Ibnu Asror
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
ibnu@telkomuniversity.ac.id

3rd Yanuar Firdaus Arie Wibowo
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
yanuar@telkomuniversity.ac.id

Abstrak-Kualitas udara pada kota yang memiliki padat penduduk tinggi dengan banyaknya pabrik industri serta padatnya jalan raya oleh kendaraan seperti DKI Jakarta harus diperhatikan kondisinya. Terdapat banyak data yang muncul mengenai kualitas udara di DKI Jakarta yang selalu menurun disebabkan oleh pencemaran udara. Data yang di dapatkan berasal dari website Jakarta Open Data yang menampilkan Indeks Standar Penecamaran Udara (ISPU) pada setiap harinya. Metode yang digunakan yaitu *data mining* klasifikasi, karena metode tersebut dapat digunakan untuk mengetahui informasi mengenai pencemaran udara berdasarkan pengolahan data parameter yang ada pada ISPU. Klasifikasi yang dilakukan yaitu menggunakan algoritma r*andom forest.* Membuat sebuah model uji pada klasifikasi menggunakan *random forest* bertujuan untuk mencari hasil terbaik. Hasil dari klasifikasi berdasarkan data Indeks Pencemaran Udara di DKI Jakarta memiliki performa terbaik yang menghasilkan akurasi tertinggi yaitu 90%. Pada kategori Sedang memiliki nilai *precision* 90,9%, *recall 89,28% f1-score* 90,09%, dan categori Tidak Sehat memiliki nilai *precision* 89,09%, *recall* 90,74%, dan *f1-score* 89,9%.

Kata kunci- kualitas udara, klasifikasi, *random forest*

Abstract-Air quality in a city that has a high population density with many industrial factories and dense highways by vehicles such as DKI Jakarta must be considered for its condition. There is a lot of data that appears regarding air quality in DKI Jakarta which is always decreasing due to air pollution. The data obtained comes from the Jakarta Open Data website which displays the AirCam Standard Index (ISPU) on a daily basis. The method used is data mining classification, because the method can be used to find out information about air pollution based on the processing of parameter data in the ISPU. The classification carried out is using the random forest algorithm Creating a test model on the classification using random forest aims to find the best results. The results of the classification based on data from the Air Pollution Index in DKI Jakarta have the best performance which produces the highest accuracy, which is 90%. In the Medium category, it has a precision value of 90,9%, recall 89,28% f1-score is 90,09%, and the Unhealthy category has a precision value of 89,09%, recall 90,74%, and an f1-score of 89,9%.

Keywords- ***air quality, classification, random forest***

I. PENDAHULUAN

A. Latar Belakang

Salah satu faktor lingkungan yang baik bagi kehidupan manusia adalah kualitas udara yang baik [1]. Dengan kualitas udara yang baik maka dapat diperoleh berbagai manfaat bagi kehidupan manusia dan makhluk hidup. Akan tetapi saat ini kualitas udara di berbagai kota sudah mulai menurun diakibatkan berbagai faktor. Suatu kota dengan pertumbuhan penduduk yang sangat pesat dengan banyaknya aktifitas pabrik industri serta padatnya jalan raya oleh kendaraan pribadi sangat berpengaruh terhadap kualitas udara yang ada sehingga terjadinya pencemaran udara.

Pencemaran udara dapat terjadi karena udara yang sudah tercampur oleh berbagai zat, yaitu: Karbon monoksida (CO), sulfur dioksida (SO_2), Nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan partikel debu (PM_{10}) [2] . Dengan adanya pencemaran udara tersebut dapat menimbulkan berbagai efek negatif bagi kelangsungan hidup manusia dan makhluk hidup. Salah satu kota yang cukup serius terhadap pencemaran udara adalah DKI Jakarta. Menurut data AirVisual yang ditampilkan oleh AQI, Indonesia berada di urutan ke-9 sebagai negara paling berpolusi di dunia dan DKI Jakarta merupakan urutan ke 202 sebagai kota yang paling berpolusi di dunia[3]. Oleh sebab itu menurut data tersebut bahwa DKI Jakarta dikategorikan sebagai kota yang memiliki kualitas udara sedang - tidak sehat.

Berdasarkan permasalahan tersebut, maka dibutuhkan proses olah data kualitas udara pada DKI Jakarta yang akan menghasilkan nilai informasi menggunakan data mining. Dengan data mining, maka dapat mengetahui informasi yang lebih rinci berdasarkan data yang berjumlah besar. Metode data mining yang digunakan untuk klasifikasi kualitas udara DKI Jakarta adalah dengan menggunakan klasifikasi. Klasifikasi digunakan untuk memprediksi suatu nilai yang akan datang berdasarkan data yang dimiliki sebelumnya. Klasifikasi yang digunakan yaitu menggunakan *random forest. Random Forest* merupakan pengembangan dari metode klasifikasi dan regresi[4].

Pada penelitian sebelumnya yang dilakukan oleh Azhar, Yufis, Galang Aji Mahesa, and Moch Chamdani Mustaqim. "Prediksi pembatalan pemesanan hotel menggunakan optimalisasi hiperparameter pada algoritme Random Forest." (2021) membahas penentuan jumlah parameter tree dalam metode random forest, menghasilkan akurasi tertinggi 87%[18]. Lalu pada penelitian oleh Sang, A. I., Sutoyo, E., & Darmawan, I. (2021) yang berjudul Analisis *Data Mining* Untuk Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma *Decision Tree* Dan *Support Vector Machine* membahas tentang pembagian rasio pada *data training* dan *data* untuk menghasilkan nilai evaluasi, menghasilkan nilai *Precision* sebesar 99,02%, *Recall* 99,73%, *F1-Measure* 99,37%, akurasi 99,40% pada algoritma *Decision Tree* dan pada algoritma *Support Vector Machine* mendapatkan nilai mendapatkan nilai *Precision* sebesar 95,82%, *Recall* 88,89%, *F1-Measure* 92,22% dan Akurasi 94,93%[14]. Pada penelitian yang dilakukan oleh Kirono, A. A. H., Asror, I., & Wibowo, Y. F. A. (2022) yang berjudul Klasifikasi Tingkat Kualitas Udara DKI Jakarta Dengan Algoritma *Naive Bayes* membahas tentang klasifikasi berdasarkan kelas kategori tingkat kualitas udara,menghasilkan nilai dengan rata – rata akurasi 88%, *precision* 85%, *recall* 96%, *f1-score* 90%[15].

B. Tujuan

Tujuan yang akan dicapai adalah untuk pengklasifikasian data Indeks Standar Pencemaran Udara (ISPU) menggunakan algoritma *Random Forest* sebagai bahan analisis, dan menjadi sebuah informasi khususnya mengenai data kualitas udara di DKI Jakarta.

C. Topik dan Batasan

Berdasarkan tujuan tersebut, maka batasan untuk penelitian yang akan dilakukan yaitu:

1. Menggunakan data yang bersumber dari *Open Government Data* DKI Jakarta
2. Data yang digunakan dimulai dari bulan Februari 2021 hingga Oktober 2021
3. Menggunakan algoritma *random forest*
4. Parameter yang digunakan adalah karbon monoksida (CO), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), ozon permukaan (O_3), partikel debu (PM_{10}), partikel debu (PM_{25}), dan *location* (lokasi)

II. KAJIAN TEORI

A. Kualitas Udara

Kualitas udara adalah ukuran baik buruknya suatu campuran gas yang ada pada lapisan toposfer yang dibutuhkan dan dapat mempengaruhi kesehatan manusia, makhluk hidup, dan unsur – unsur yang ada pada lingkungan hidup[9]

B. Pencemaran Udara

Menurut Undang – Undang Pokok Pengolahan Lingkungan Hidup No. 4 Tahun 1982, pencemaran udara adalah masuknya atau dimasukannya makhluk hidup, zat energi, dan atau komponen lain ke dalam lingkungan, atau berubahnya tatanan lingkungan oleh kegiatan manusia atau oleh proses alam sehingga kualitas lingkungan menjadi kurang atau tidak dapat berfungsi lagi sesuai peruntukannya[5]. Pencemaran udara dapat terjadi karena adanya beberapa senyawa yang tercampur dengan udara. Senyawa yang menyebabkan terjadinya pencemaran udara diantaranya adalah Karbon monoksida (CO), sulfur dioksida (SO_2), Nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan partikel debu (PM_{10}) [2]. Pada intensitas tertentu pencemaran udara dapat langsung bereaksi terhadap kesehatan manusia, diantaranya masalah pernapasan, iritasi mata, dan alergi kulit[6].

C. Indeks Standar Pencemaran Udara

Indeks standar pencemaran udara (ISPU) adalahangka yang tidak memiliki nilai satuan yang menggambarkan kondisi kualitas udara di lokasi dan waktu tertentu berdasarkan pada dampak kesehatan manusia , nilai estetika, dan makhluk hidup lainnya[11]. Berdasarkan keputusan Menteri Lingkungan Hidup Nomor: KEP 45/MENLH/1997 Tentang Indeks

Standar Pencemaran Udara, bahwa ISPU digunakan sebagai standar kualitas udara yang resmi di Indonesia[12]. Pada ISPU terdapat konvensi nilai konsentrasi, yang ditampilkan pada tabel dibawah ini:

TABEL 1
KONVENSI NILAI KONSENTRASI

ISPU	24 Jam PM_{10} ($\mu g/m^3$)	24 Jam PM_{25} ($\mu g/m^3$)	24 Jam SO_2 ($\mu g/m^3$)	24 Jam CO ($\mu g/m^3$)	24 Jam O_3 ($\mu g/m^3$)	24 Jam NO_2 ($\mu g/m^3$)	24 Jam HC ($\mu g/m^3$)
0-50	50	15,5	52	4000	120	80	45
51-100	150	55,4	180	8000	235	200	100
101-200	350	150,4	400	15000	400	1130	215
201-300	420	250,4	800	30000	800	2260	432
>300	500	500	1200	45000	1000	300	648

Keterangan:

1. Data pengukuran dilakukan selama 24 jam tanpa henti
2. Hasil perhitungan ISPU pada parameter partikulat (PM_{25}) disampaikan tiap jam selama 24 jam
3. Hasil perhitungan ISPU pada parameter partikel debu (PM_{10}), Karbon monoksida (CO), Sulfur oksida (SO_2), Nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan Hidrokarbon (HC), diambil nilai ISPU parameter tertinggi dan paling sedikit disampaikan setiap jam 09.00 dan jam 15.00

Untuk mendapatkan hasil berdasarkan tabel tersebut, terdapat rumus perhitungan nya yaitu pada persamaan berikut:

$$I = \frac{Ia - Ib}{Xa - X} b\ (X_x - X_b) + I_b \qquad (1)$$

Keterangan:

I = ISPU terhitung X_a = Konsentrasi ambien batas atas ($\mu g/m^3$)

I_a = ISPU batas atas X_b = Konsentrasi ambien batas bawah ($\mu g/m^3$)

I_b = ISPU batas bawah X_x = Konsentrasi ambien nyata hasil pengukuran ($\mu g/m^3$)

Dibawah ini adalah tabel keterangan nilai yang ada pada Indeks Standar Pencemaran Udara (ISPU) dan juga level pencemaran udara serta dampak kesehatan yang terjadi berdasarkan nilai ISPU:

TABEL 2
SKALA PENCEMARAN UDARA

ISPU	Pencemaran Udara Level
0-50	Baik
51-100	Sedang
101-199	Tidak Sehat
200-299	Sangat Tidak Sehat
>300	Berbahaya

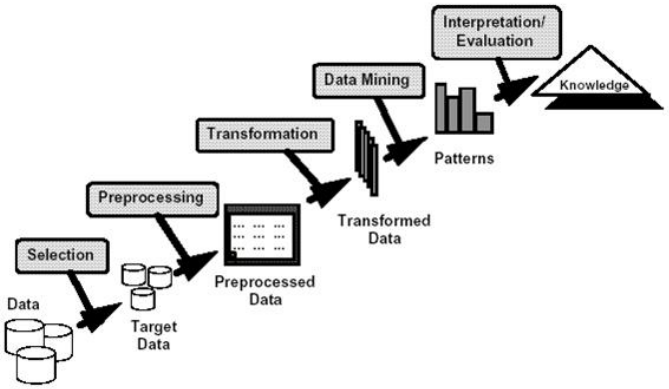
Keterangan:

1. Skala 0-50 (Baik): Skala yang tidak memberikan dampak negatif yang dihasilkan kepada manusia dan makhluk hidup lainnya.
2. Skala 51-100(Sedang): Skala yang masih dapat diterima oleh kesehatan manusia dan makhluk hidup lainnya.
3. Skala 1011-100(Tidak Sehat): Skala yang sudah mulai sedikit memberikan dampak negatif pada manusia dan makhluk hidup lainnya.
4. Skala 200-299(Sangat Tidak Sehat): Skala yang sudah dapat memberikan dampak negatif kepada kesehatan manusia dan makhluk hidup pada populasi.
5. Skala >300(Berbahaya): Skala yang dapat memberikan dampak negatif kepada kesehatan manusia dan makhluk hidup secara serius dan harus segera dilakukan penangan cepat untuk penyembuhan.

D. Data Mining

Data mining adalah suatu proses mengidentifikasi data untuk mendapatkan informasi yang berguna dalam kumpulan data yang sangat besar, yang kemudian digunakan untuk proses pengolahan data[7]. Data yang akan diperoleh menggunakan d*ata mining*

harus sesuai dengan alur *Knowledge Discovery in Database* (KDD) dan dibagi menjadi 2 bagian yaitu *data training* dan *data testing*.



GAMBAR 1
PROSES KDD

Berikut adalah penjelasan dari setiap tahapan pada proses KDD:

1. Selection
   Pemilihan data dilakukan sebelum masuk ke tahapan penggalian informasi dalam proses KDD dimulai. Hasil yang sudah di seleksi di simpan dalam database yang berbeda dengan data base operasional, yang kemudian akan digunakan dalam proses data mining[8].
2. Preprocessing (cleaning)
   Proses *cleaning* ini dilakukan untuk menghilangkan noise pada data supaya tidak terjadinya kesalahan dalam proses data mining.
3. Transformation
   Data yang sudah di seleksi (tahap 1) kemudian data di proses yang nantinya dapat digunakan untuk proses data mining, dengan cara memodifikasi data ke model analitis.
4. Data Mining
   Memilih metode *data mining* yang sesuai dengan tujuan tertentu yang telah di tentukan pada tahap pertama(seleksi).
5. Interpretation / Evaluation
   Pada tahap ini melakukan pemeriksaan apakah informasi yang ditemukan, apakah bertentangan dengan hipotesa yang sudah ada.
6. Knowledge
   Hasil pada tahap ini digunakan untuk melakukan suatu tindakan terhadap informasi yang sudah selesai didapatkan dan bagaimana membuat keputusan terhadap hasil analisis yang telah dilakukan.

Berdasarkan fungsionalitasnya *data mining* dikelompokkan menjadi 6 bagian, yaitu:

1. Klasifikasi(*Classification*)
2. Klastering(*Clustering*)
3. Regresi(*Regression*)
4. *Association Rule*
5. *Anomaly Detection*
6. *Summarization*

E. Klasifikasi

Klasifikasi adalah suatu proses untuk mencari nilai fungsi yang menjelaskan suatu kelas data, yang bertujuan untuk memperkirakan kelas dari suatu objek atau label yang nilainya belum diketahui[12].

F. Random Forest

Random Forest adalah algoritma dalam *machine learning* yang digunakan untuk pengklasifikasian *dataset* dalam jumlah besar. Karena fungsinya bisa digunakan untuk banyak dimensi dengan berbagai skala dan performa yang tinggi[11]. Klasifikasi ini dilakukan melalui penggabungan *tree* dalam *decision tree* dengan cara *training dataset* yang dimiliki. *Decision tree* terdiri dari *root node*, *internal node*, dan *leaf node* dengan mengambil atribut dan data secara aacak sesuai ketentuan yang berlaku[16].*Root node* adalah simpul yang terletak paling atas, yang biasa disebut akar dari keputusan. *Internal node* adalah simpul pada percabangan, yang hanya mempunyai satu input dan jumlah minimal outputnya dua. *Leaf node* adalah simpul yang terletak di akhir yang memiliki satu input dan tidak memiliki output. *Decision tree* dimulai dari menghitung nilai *entropy* dan nilai *information gain*. Untuk menghitung nilai *entropy* dan nilai *information gain* dapat

dilihat pada rumus dibawah ini[17]:

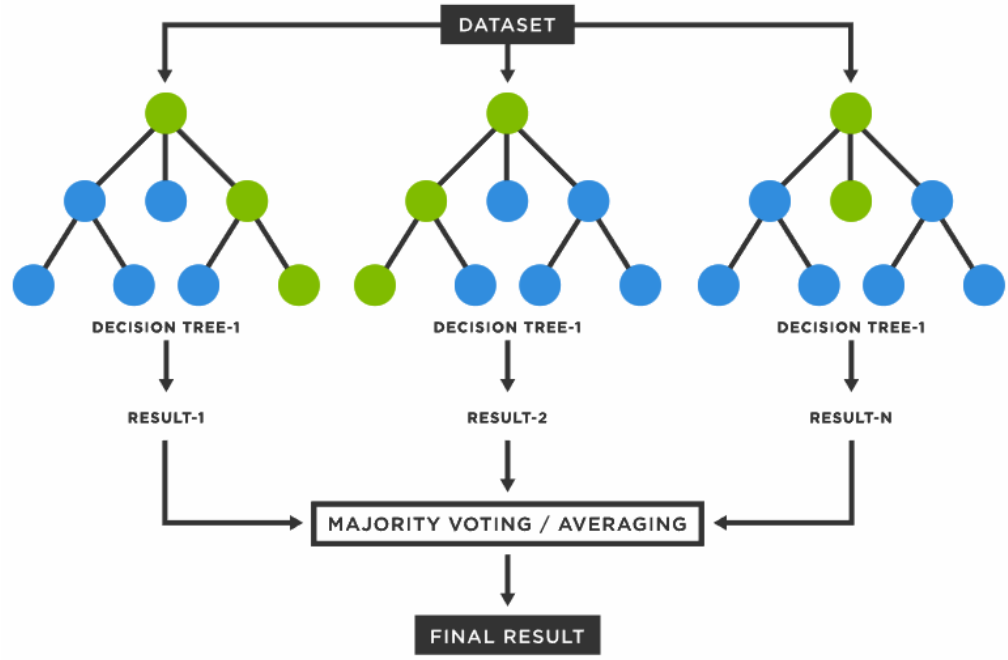
$$\boldsymbol{Entropy}\ (\boldsymbol{Y}) = -\Sigma \mathbf{i}\ \mathbf{p}\ (\mathbf{c}|\mathbf{Y})\ \mathbf{log2}\ \mathbf{p}\ (\mathbf{c}|\mathbf{Y}) \quad (2)$$

Dimana Y adalah himpunan kasus dan p(c|Y) adalah proporsi nilai Y terhadap kelas c.

$$\boldsymbol{Information}\ (\boldsymbol{Y},) = \boldsymbol{Entropy}(\boldsymbol{Y}) - \Sigma_{\boldsymbol{veValues}} \frac{|Yv|}{|Ya|} \boldsymbol{Entropy}(\boldsymbol{Yv}) \quad (3)$$

Dimana *Values*(a) merupakan semua nilai yang mungkin pada himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a.

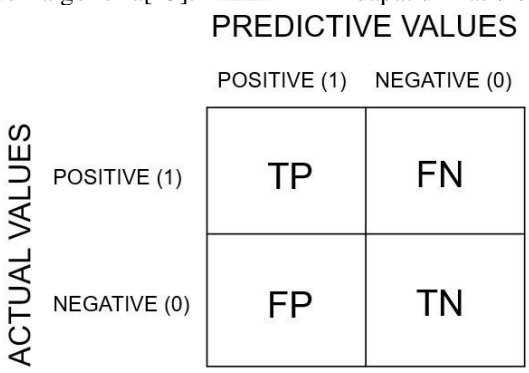
Pada setiap node yang dibuat pada *decision tree* hanyalah sebagian saja yang terpilih untuk kemudian dilakukan pemisahan yang terbaik[19]. Jadi, pengklasifikasian *random forest* terdiri dari jumlah n *tree*, dimana n merupakan jumlah *tree* yang akan di tentukan oleh pengguna. Untuk mengklasifikasikan kumpulan data baru, setiap kumpulan data yang ada diturunkan ke masing – masing *tree* dengan jumlah n. Selanjutnya hutan melakukan *vote* untuk memilih kelas yang memiliki n suara terbanyak.



GAMBAR 2
RANDOM FOREST

G. Confusion Matrix

Confusion Matrix adalah suatu table matriks yang berfungsi untuk perhitungan performa pada suatu model algoritma[13]. Setiap baris pada matriks tersebut menunjukan kelas aktual berdasarkan data, dan setiap kolom menunjukan prediksi yang dihasilkan pada data Tabel *Confusion Matrix* dapat dilihat dibawah ini:



GAMBAR 3
CONFUSION MATRIX

1. TP (*True Positive*): Banyaknya data yang nilai aktual kelasnya positif, dan nilai prediksinya positif
2. TN (*True Negative*): Banyaknya data yang nilai aktual kelasnya negatif, dan nilai prediksinya negatif
3. FP (*False Positive*): Banyaknya data yang nilai aktual kelasnya negatif, tetapi nilai prediksinya positif
4. FN (*False Negative*): Banyaknya data yang nilai aktual kelasnya positif, tetapi nilai prediksinya negatif

Berdasarkan data tersebut, dapat diperoleh data-data yang berguna untuk mengukur perfomansi dalam suatu model, diantaranya adalah:

1. Akurasi, adalah total keseluruhan seberapa sering model benar mengklasifikasi, dihitung dengan membandingkan jumlah data yang benar terklasifikasi dengan jumlah data keseluruhan. Persamaan akurasi adalah sebagai berikut:

$$\frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

2. *Precision*, diartikan sebagai ukuran ketepatan. Jika data diprediksi positif, seberapa sering data prediksi itu benar. Persamaan *precision* adalah sebagai berikut:

$$\frac{TP}{TP+FP} \quad (5)$$

3. *Recall,* diartikan sebagai ukuran kelengkapan. Dari jumlah data sebenarnya yang bernilai positif, sebanyak apakah data yang diprediksi positif. Persamaan *recall* adalah sebagai berikut:

$$\frac{TP}{TP+FP} \quad (6)$$

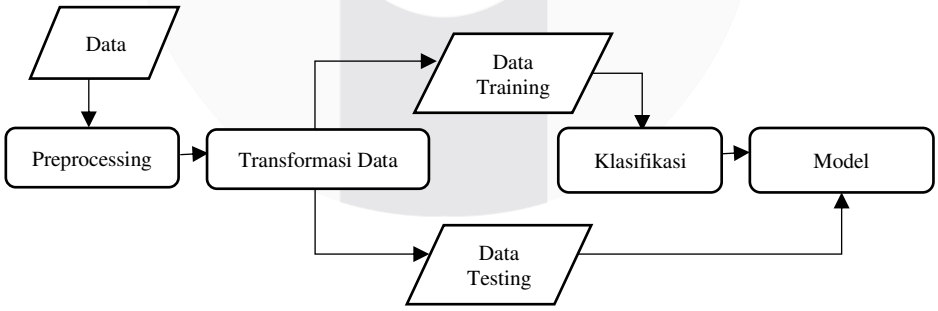
4. F1- Score, menggambarkan perbandingan rata-rata precision dan recall yang dibobotkan. Persamaan F1-Score adalah sebagai berikut:

$$\frac{2\,(Recall * Precision)}{(Recall+Precision)} \quad (7)$$

III. METODE

A. Alur Sistem

Penelitian ini dilakukan untuk mengukur tingkat kualitas udara dengan metode klasifikasi menggunakan algoritma *random forest*. Tahapan proses dari penelitian ini dapat dilihat berdasarkan flowchart dibawah ini:



GAMBAR 4
ALUR SISTEM

B. Sumber Data

Data yang didapatkan berdasarkan dari website *open government data* DKI Jakarta https://data.jakarta.go.id/. Data yang diambil yaitu data kualitas udara DKI Jakarta dari bulan Februari 2021 hingga Oktober 2021. Terdapat total record data yang ada sebanyak 3003 *dataset*.Pada atribut kelas categori terdapat total 273 data yang dimana terdapat dua categori, yaitu categori sedang sebanyak 145 data(53,1%) dan categori tidak sehat (46,9%).Dibawah ini adalah contoh data pada bulan Februari 2021

```
tanggal,pm10,pm25,so2,co,o3,no2,max,critica
Feb,73,126,38,26,46,34,126,PM25,TIDAK SEHAT
Feb,53,70,40,14,55,25,70,PM25,SEDANG,DKI3
Feb,32,53,40,11,42,19,53,PM25,SEDANG,DKI3
Feb,36,59,40,14,47,24,59,PM25,SEDANG,DKI5
Feb,29,51,40,14,45,35,51,PM25,SEDANG,DKI3
Feb,34,53,40,8,57,15,57,O3,SEDANG,DKI2
Feb,33,55,40,10,57,13,57,O3,SEDANG,DKI2
Feb,26,44,39,10,54,17,54,O3,SEDANG,DKI2
Feb,33,57,40,13,47,22,57,PM25,SEDANG,DKI4
Feb,50,64,40,13,49,16,64,PM25,SEDANG,DKI3
Feb,38,57,43,13,35,17,57,PM25,SEDANG,DKI3
Feb,63,98,43,16,33,42,98,PM25,SEDANG,DKI3
Feb,59,89,40,12,40,16,89,PM25,SEDANG,DKI3
Feb,55,73,40,11,42,19,73,PM25,SEDANG,DKI3
Feb,42,66,40,13,37,25,66,PM25,SEDANG,DKI4
Feb,43,63,40,11,42,26,63,PM25,SEDANG,DKI4
Feb,46,71,40,25,41,37,71,PM25,SEDANG,DKI5
Feb,53,70,44,13,41,29,70,PM25,SEDANG,DKI3
Feb,32,52,42,20,41,33,52,PM25,SEDANG,DKI5
Feb,45,63,39,13,53,20,63,PM25,SEDANG,DKI4
Feb,36,52,39,10,48,17,52,PM25,SEDANG,DKI5
Feb,68,103,42,22,41,40,103,PM25,TIDAK SEHAT
Feb,66,90,40,16,54,23,90,PM25,SEDANG,DKI3
Feb,42,61,40,10,33,16,61,PM25,SEDANG,DKI4
Feb,31,54,43,12,45,23,54,PM25,SEDANG,DKI3
Feb,48,75,43,13,40,20,75,PM25,SEDANG,DKI4
Feb,59,94,44,15,50,31,94,PM25,SEDANG,DKI4
Feb,68,113,40,18,60,18,113,PM25,TIDAK SEHAT
```

GAMBAR 5
DATA BULAN FEBRUARI 2021

Pada dataset tersebut memiliki 11 atribut,yaitu:

1. Tanggal: Waktu pengukuran kualitas udara
2. PM_{10}: Partikulat berukuran 10 micron
3. PM_{25}: Partikulat berukuran 2,5 micron
4. SO_2: Sulfur Oksida
5. CO: Carbon Monoksida
6. O_3: Ozon
7. NO_2: Nitrogen dioksida salah satu parameter yang diukur
8. Max: Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama
9. Critical: Parameter berdasarkan pengukuran indeks paling tinggi
10. Categori: Kategori hasil perhitungan indeks standar pencemaran udara dan mencakup sebagai
nilai dari kelas
11. Location: Lokasi pengukuran berdasarkan SPKU (Stasiun Pemantauan Kualitas Udara)

C. Proses KDD

1. Input Data

Data yang sudah di dapatkan dari sumber website https://data.jakarta.go.id/ kemudian dilakukan input ke dalam jupyter notebook untuk dilakukan proses pengolahan data.

	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	categori	location
0	Feb	73	126	38	26	46	34	126	PM25	TIDAK SEHAT	DKI5
1	Feb	53	70	40	14	55	25	70	PM25	SEDANG	DKI3
2	Feb	32	53	40	11	42	19	53	PM25	SEDANG	DKI3
3	Feb	36	59	40	14	47	24	59	PM25	SEDANG	DKI5
4	Feb	29	51	40	14	45	35	51	PM25	SEDANG	DKI3
...	...	...	...	...	...	...	...	...	...	...	...
268	Oct	62	90	64	15	50	39	90	PM25	SEDANG	DKI4
269	Oct	54	78	67	16	56	39	78	PM25	SEDANG	DKI4
270	Oct	54	79	80	19	49	35	80	SO2	SEDANG	DKI2
271	Oct	64	103	81	15	58	40	103	PM25	TIDAK SEHAT	DKI4
272	Oct	56	79	63	28	56	32	79	PM25	SEDANG	DKI4

273 rows × 11 columns

GAMBAR 6
INPUT DATA

2. Preprocessing

Proses *preprocessing* dilakukan untuk membersihkan data yang terdapat noise, duplikat, ataupun data yang tidak lengkap. Selanjutnya atribut yang tidak diperlukan untuk proses *data mining* akan di buang. Atribut parameter yang dibuang adalah max, critical, dan categori.

```
X = df.drop (['max', 'critical', 'tanggal'], axis = 1)
X
```

	pm10	pm25	so2	co	o3	no2	categori	location
0	73	126	38	26	46	34	TIDAK SEHAT	DKI5
1	53	70	40	14	55	25	SEDANG	DKI3
2	32	53	40	11	42	19	SEDANG	DKI3
3	36	59	40	14	47	24	SEDANG	DKI5
4	29	51	40	14	45	35	SEDANG	DKI3
...	...	...	...	...	...	...	...	...
268	62	90	64	15	50	39	SEDANG	DKI4
269	54	78	67	16	56	39	SEDANG	DKI4
270	54	79	80	19	49	35	SEDANG	DKI2
271	64	103	81	15	58	40	TIDAK SEHAT	DKI4
272	56	79	63	28	56	32	SEDANG	DKI4

273 rows × 8 columns

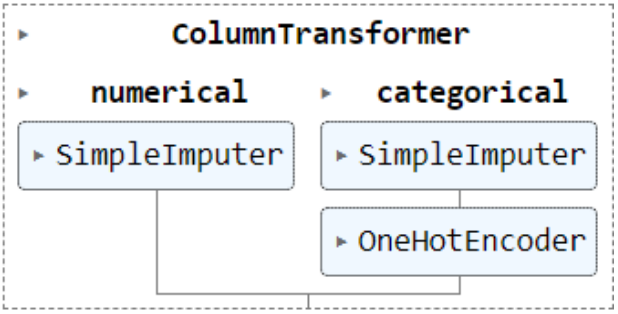
GAMBAR 7
PREPROCESSING

3. Transformasi Data

Proses transformasi dimulai dari pemisahan dua tipe data, yaitu data *numerical* dan *categorical* dengan menggunakan *ColumnTransformer*, karena pada dataset yang digunakan merupakan kumpulan data yang berisi tipe data heterogen atau lebih dari satu tipe data. Pemisahan tersebut dilakukan untuk melihat skala pada *numerical* dan *categorical*. Pada penelitian ini atribut yang termasuk *numerical* adalah 'pm10', 'so2', 'co', 'o3', 'no2' , dan untuk *cateogircal* adalah *location*. Kedua tipe data tersebut di skalakan standar setelah imputasi rata-rata, dan pada tipe data *categorical* ditambah dengan melakukan *onehot encode*,maksudnya adalah dengan merepresentasikan data yang ada pada atribut *location* ke dalam biner yang bernilai 0 dan 1, dimana semua elemen bernilai 0, kecuali elemen yang bernilai pada nilai kategori tersebut. Data yang ada pada atribut *location* yaitu, DKI1, DKI2, DKI3, DKI4, DKI5. Berikut adalah proses *onehot encode* pada atribut parameter *location* dan proses keselurahan pada transformasi data menggunakan *ColumnTransformer*

TABEL 3
ONEHOT ENCODER

LOCATION		DKI1	DKI2	DKI3	DKI4	DKI5
DKI1		1	0	0	0	0
DKI2		0	1	0	0	0
DKI3		0	0	1	0	0
DKI4		0	0	0	1	0
DKI5		0	0	0	0	1



GAMBAR 8
TRANSFORMASI DATA

4. Data Training dan Data Testing

Sebelum masuk ke proses klasifikasi data yang sudah di lakukan seleksi *data* , maka *dataset* dibagi terlebih dahulu menjadi 2 bagian, yaitu *data training* dan *data testing* agar data dapat diolah pada proses klasifikasi.

5. Klasifikasi

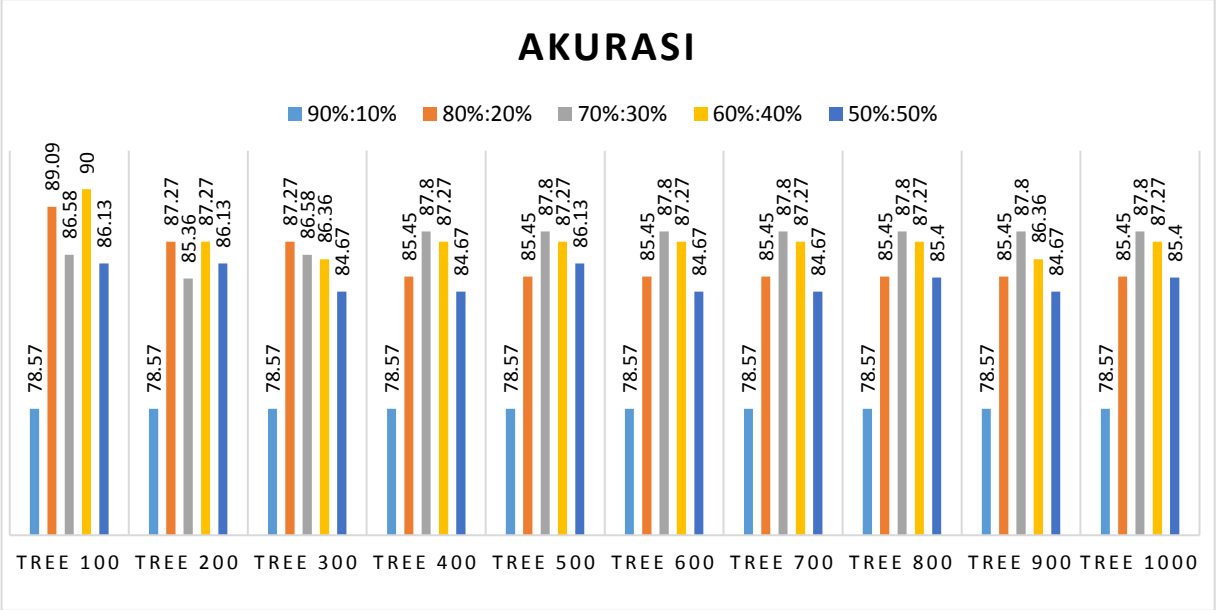
Data yang sudah siap diolah selanjutnya akan diuji dengan metode klasifikasi yang menggunakan algoritma *random forest*. Pengujian yang dilakukan yaitu untuk memunculkan kelas pada attribut categori yang menunjukan tingkat kualitas udara DKI Jakarta berdasarkan *dataset* yang digunakan. Metode klasifikasi dilakukan menggunakan bahasa pemrograman phyton yang dimana sudah tersedia library *random* forest didalamnya.

6. Model

Model yang dihasilkan pada penelitian ini yaitu, untuk melihat hasil *akurasi* dengan melakukan penentuan jumlah *tree* pada interval kelipatan 100, yang dimulai dari 100 hingga 1000 dan dengan menentukan perbandingan rasio pada pembagian *data training* dan *data testing*, yaitu 90%:10%, 80%:20%, 70%:30%, 60%:40% dan 50%:50%. Selanjutnya hasil pengujian tersebut di tampilkan pada evaluasi.

7. Evaluasi

Evaluasi melakukan pembentukan output bersumber dari hasil proses klasifikasi pada model yang dihasilkan. Hasil dari evaluasi di tampilkan dalam grafik untuk mengetahui hasil akurasinya.Perhitungan nilai dari hasil evaluasi dengan menguji dari hasil proses *klasifikasi* menggunakan confusion matrix. Berikut adalah grafik hasil pengujian yang telah dilakukan dengan menunjukan hasil persentase dari akurasi.



GAMBAR 9
GRAFIK HASIL AKURASI

IV. HASIL DAN PEMBAHASAN

Berdasarkan grafik hasil skenario uji yang sudah dilakukan dengan merubah jumlah tree yang digunakan, yaitu dengan interval kelipatan 100, dari 100 hingga 1000 serta merubah rasio perbandingan *data training* dan *data testing* yaitu 90%:10%, 80%:20%, 70%:30%, 60%:40%, 50%:50% mendapatkan hasil terbaik dengan akurasi 90% pada tree yang berjumlah 100 dengan pembagian *dataset* 60% *data training* dan 40% *data testing*. Pada hasil akurasi tersebut menunjukkan tingkat kualitas udara yang ada di DKI Jakarta dari bulan Februari 2021 hingga Oktober 2021. Terdapat pula hasil *precision,recall, f1-score* pada masing – masing tingkat kualitas udara tersebut, yang dapat dilihat pada tabel dibawah ini.

TABEL 4
HASIL AKURASI TERTINGGI

	Precision	Recall	F1-Score
SEDANG	90,9%	89,28%	90,09%
TIDAK SEHAT	89,09%	90,74%	89,9%
Accuracy			90%

V. KESIMPULAN

Berdasarkan analisis hasil pengujian yang dilakukan berdasarkan data Indeks Standar Pencemaran Udara DKI Jakarta yang di dapatkan dari website data.jakarta.go.id pada data harian yang tercatat dari bulan Februari 2021 hingga Oktober 2021 dengan adanya atribut parameter pengukuran tingkat kualitas udara yaitu PM10, PM25, SO2, CO, O3 NO2 serta atribut *location* dan categori yang dilakukan *preprocessing* dengan membuang atribut parameter *max,* critical, dan tanggal menggunakan klasifikasi *data mining* dengan algoritma *random forest* menghasilkan nilai terbaik pada akurasi 90%. Pada kategori SEDANG memiliki nilai *precision* 90,09%, *recall 89,28% f1-score* 90,09%, dan pada kategori TIDAK SEHAT memiliki nilai *precision* 89,09%, *recall* 90,74%, dan *f1-score* 89,9%. Berdasarkan hasil evaluasi tersebut dapat disimpulkan bahwa perubahan nilai tree dan perbandingan rasio pada *data training* dan *data testing* memberikan perngaruh pada hasil akurasi.

REFERENSI

[1]. K. Prabowo and B. Muslim, *Penyehat Udara*, Pertama. Jakarta Selatan: Pusat Pendidikan Sumber Daya Manusia Kesehatan, 2018.

[2]. Kementrian Lingkungan Hidup dan Kehutanan DITJEN Pengedalian Pencemaran Dan Kerusakan Lingkungan DIREKTORAT Pengendalian Pencemaran Udara "INDEKS STANDAR PENCEMARAN UDARA (ISPU) SEBAGAI INFORMASI MUTU UDARA AMBIEN DI INDONESIA", 2020 [online].

[3]. IQ*Air* "Kualitas udara di Indonesia", 2021 [online].

Avaliable: https://www.iqair.com/id/indonesia

[4]. Suci Cahaya H.N., Fibri Rakhmawati, Riri Syafitri Lubis, "KLASIFIKASI TINGKAT PENCEMARAN UDARA PADA SEKTOR INDUSTRI DENGAN METODE RANDOM FOREST"

[5]. Undang-Undang Pokok Pengolahan Lingkungan Hidup No.4 Tahun 1982.

[6]. A. Budiyono, "Pencemaran Udara : Dampak Pencemaran Udara Pada Lingkungan," *Dirgantara*, vol. 2,no. 1, pp. 21–27, 2010.

[7]. Osman, Abdullahi Sidow. "Data mining techniques." (2019).

[8]. Kusrini, Emha Taufiq Lutfi "Algoritma Data Mining", CV ANDI OFFSET, Yogyakarta.

[9]. Vallero, D. (2014), Fundamentals of Air Pollution - Fifth Edition, Cambridge: Elsevier Academic Pres.

[10]. Keputusan Menteri Negara Lingkungan Hidup Nomor: KEP 45 / MENLH / 1997.

[11]. J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, vol. 84. 2013.

[12]. J. Han, Data mining: Data mining concepts and techniques. 2014.

[13]. Irkham Widhi Saputro, Bety Wulan Sari, "Uji Performa Algoritma Naive Bayes untuk Prediksi Masa Studi Mahasiswa", Citec Journal, Vol. 6, No. 1 ISSN: 2460-4259, Universitas AMIKOM Yogyakarta, 2019.

[14]. Sang, A. I., Sutoyo, E., & Darmawan, I. (2021). Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine. *eProceedings of Engineering*, *8*(5)

[15]. Kirono, A. A. H., Asror, I., & Wibowo,

Y. F. A. (2022). Klasifikasi Tingkat Kualitas Udara Dki Jakara Dengan Algoritma Naive Bayes. *eProceedings of Engineering*, *9*(3).

[16]. Siburian, Vanissa Wanika, and Ika Elvina Mulyana. "Prediksi Harga Ponsel Menggunakan Metode Random Forest." *Annual Research Seminar (ARS)*. Vol. 4. No. 1. 2019.

[17]. Schouten, Kim, Flavius Frasincar, and Rommert Dekker. "An information gain-driven feature study for aspect-based sentiment analysis." *International Conference on Applications of Natural Language to Information Systems*. Springer, Cham, 2016.

[18]. Azhar, Yufis, Galang Aji Mahesa, and Moch Chamdani Mustaqim. "Prediksi pembatalan pemesanan hotel menggunakan optimalisasi hiperparameter pada algoritme Random Forest." *Jurnal Teknologi dan Sistem Komputer* 9.1 (2021)

[19]. Pal, Mahesh. "Random forest classifier for remote sensing classification." *International journal of remote sensing* 26.1 (2005): 217-222.