BAB II

LANDASAN TEORI

2.1 Landasan Teori

2.1.1 Pencemaran Udara

Perkembangan ekonomi yang berkelanjutan, bertambahnya volume kendaraan pribadi, dan industrialisasi modern telah meningkatkan pencemaran udara di daerah perkotaan secara signifikan. Menurut *World Health Organization* (WHO), pencemaran udara adalah kontaminasi lingkungan didalam atau luar ruangan oleh unsur kimia, fisik, atau biologis apa pun yang mengubah karakteristik alami atmosfer (*World Health Organization*. "*Air Pollution*"). Dilansir dari laman WHO, polutan yang menjadi perhatian utama terhadap kesehatan masyarakat meliputi:

1) Partikulat atau *Particulate Matter* (PM) mengacu pada partikel yang dapat terhirup, terdiri dari campuran kompleks partikel padat dan cair dari zat organik dan anorganik yang tersuspensi di udara. *Particulate Matter* dibedakan menjadi PM_{10} yaitu partikel dengan diameter 10 mikron atau lebih kecil ($\leq PM_{10}$) yang dapat menembus dan bersarang jauh di dalam paru-paru, dan partikel dengan diameter 2,5 mikron atau lebih kecil ($\leq PM_{2.5}$). $PM_{2.5}$ adalah partikel yang lebih merusak kesehatan karena dapat menembus pembatas paru dan masuk ke sistem darah.
2) Karbon monoksida (CO) adalah gas beracun yang tidak berwarna, tidak berbau dan tidak berasa yang dihasilkan oleh pembakaran tidak sempurna

bahan bakar berkarbon seperti kayu, bensin, arang, gas alam, dan minyak tanah. Paparan karbon monoksida dapat menyebabkan kesulitan bernapas, kelelahan, pusing, dan gejala mirip flu lainnya. Paparan kadar karbon monoksida yang sangat tinggi dapat menyebabkan kematian.

3) Ozon (O_3) adalah gas yang terbentuk dari reaksi fotokimia dengan polutan seperti nitrogen oksida (NO_x) yang dipancarkan dari kendaraan, dan industri. Paparan ozon yang berlebihan dapat menyebabkan masalah pernapasan, memicu asma, mengurangi fungsi paru-paru dan menyebabkan penyakit paru-paru.
4) Nitrogen dioksida (NO_2) adalah gas yang dihasilkan dari proses pembakaran seperti yang digunakan untuk pemanasan, mesin pada kendaraan, dan pembangkit listrik. Paparan nitrogen dioksida dapat mengiritasi saluran udara dan memperburuk penyakit pernapasan.
5) Sulfur dioksida (SO_2) adalah gas tidak berwarna dengan bau yang tajam. Ini dihasilkan dari pembakaran bahan bakar fosil (batubara dan minyak) dan peleburan bijih mineral yang mengandung belerang. Sulfur dioksida dapat mempengaruhi sistem pernapasan dan fungsi paru-paru, serta menyebabkan iritasi pada mata.

2.1.2 Indeks Standar Pencemaran Udara

Indeks Standar Pencemar Udara (ISPU) adalah indeks standar kualitas udara yang dipergunakan secara resmi di Indonesia. Hal ini sesuai dengan Peraturan

Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 tentang Indeks Standar Pencemar Udara.

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. Parameter yang meliputi ISPU yaitu:

1) Partikulat (PM_{10})
2) Partikulat ($PM_{2.5}$)
3) Karbon monoksida (CO)
4) Nitrogen dioksida (NO_2)
5) Sulfur dioksida (SO_2)
6) Ozon (O_3)
7) Hidrokarbon (HC)

Perhitungan ISPU dilakukan melalui kegiatan pemantauan dan konversi konsentrasi parameter menjadi nilai ISPU. Konversi nilai konsentrasi parameter ISPU dapat dilihat pada Tabel 2.1.

Tabel 2.1 Konversi Nilai Konsentrasi Parameter ISPU

ISPU	24 Jam Partikulat (PM_{10}) µg/m³	24 Jam Partikulat ($PM_{2.5}$) µg/m³	24 Jam Sulfur dioksida (SO_2) µg/m³	24 Jam Karbon monoksida (CO) µg/m³	24 Jam Ozom (O_3) µg/m³	24 Jam Nitrogen dioksida (NO_2) µg/m³	24 Jam Hidrokarbon (HC) µg/m³
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	3000	648

Konversi konsentrasi parameter polutan menjadi nilai ISPU dilakukan melalui persaman 2.1 berikut:

$$I = \frac{(I_a - I_b)}{(X_a - X_b)} (X_x - X_b) + I_b \qquad (2.1)$$

Keterangan:

I = ISPU terhitung

I_a = ISPU batas atas

I_b = ISPU batas bawah

X_a = Konsentrasi ambien batas atas (µg/m³)

X_b = Konsentrasi ambien batas bawah (µg/m³)

X_x = Konsentrasi ambien nyata hasil pengukuran (µg/m³)

Nilai ISPU yang digunakan sebagai kategori akhir adalah salah satu nilai ISPU parameter yang tertinggi dari hasil perhitungan konversi terhadap semua parameter yang kemudian dimasukkan kedalam kelas kategori ISPU. Rentang kategori ISPU dapat dilihat pada Tabel 2.2.

Tabel 2.2 Rentang kategori nilai ISPU

Angka rentang	Kategori	Keterangan	Apa yang harus dilakukan	Status warna
1 – 50	Baik	Tingkat kualitas udara yang sangat baik, tidak memberikan efek negatif terhadap manusia, hewan, tumbuhan.	Sangat baik melakukan kegiatan diluar.	Hijau
51 – 100	Sedang	Tingkat kualitas udara masih dapat diterima pada kesehatan manusia, hewan, dan tumbuhan.	Kelompok sensitif: Kurangi aktvitas fisik yang terlalu lama atau berat. Setiap orang: Masih dapat beraktivitas diluar.	Biru
101 – 200	Tidak sehat	Tingkat kualitas udara yang bersifat merugikan pada manusia, hewan, dan tumbuhan.	Kelompok sensitif: Boleh melakukan aktivitas diluar, tetapi mengambil rehat lebih sering dan melakukan aktivitas ringan. Amati gejala berupa batuk atau nafas sesak. Penderita asma harus mengikuti petunjuk kesehatan untuk asma dan menyimpan obat asma. Penderita penyakit jantung: Gejala seperti palpitasi/jantung berdetak lebih cepat, sesak nafas, atau kelelahan yang tidak biasa mungkin mengindikasi masalah serius. Setiap orang: Mengurangi aktivitas fisik yang terlalu lama diluar ruangan.	Kuning

Tabel 2.2 Rentang kategori nilai ISPU (Lanjutan 1)

Angka rentang	Kategori	Keterangan	Apa yang harus dilakukan	Status warna
200 – 300	Sangat tidak sehat	Tingkat kualitas udara yang dapat meningkatkan resiko kesehatan pada sejumlah segmen populasi yang terpapar.	Kelompok sensitif: Hindari semua aktivitas diluar. Perbanyak aktivitas didalam ruangan atau lakukan penjadwalan ulang pada waktu dengan kualitas udara yang baik. Setiap orang: Hindari aktivitas fisisk yang terlalu lama diluar ruangan, pertimbangkan untuk melakukan aktivitas didalam ruangan.	Merah
≥ 301	Berbahaya	Tingkat kualitas udara yang dapat merugikan kesehatan serius pada populasi dan perlu penanganan cepat.	Kelompok sensitif: Tetap didalam ruangan dan hanya melakukan sedikit aktivitas. Setiap orang: Hindari semua aktivitas diluar.	Hitam

2.1.3 *Machine Learning*

Machine learning atau pembelajaran mesin dapat didefinisikan secara luas sebagai metode komputasi berdasarkan pengalaman untuk meningkatkan kinerja atau membuat prediksi yang akurat. Pengalaman mengacu pada informasi masa lalu yang tersedia dan dapat dijadikan data pembelajaran, yang biasanya berbentuk data elektronik yang dikumpulkan dan tersedia untuk analisis (Mohri, et al., 2012:1).

Istilah *machine learning* pertama kali didefinisikan oleh Arthur Samuel pada tahun 1959. Menurut Arthur Samuel, *machine learning* adalah suatu bidang studi

yang memberikan kemampuan pembelajaran kepada komputer untuk mengetahui sesuatu tanpa pemrograman yang eksplisit.

Tujuan utama dari *machine learning* adalah menghasilkan prediksi yang akurat untuk item yang tidak terlihat dan merancang algoritma yang efisien dan kuat untuk menghasilkan prediksi, bahkan untuk masalah dalam skala besar (Mohri, et al., 2012:2).

Metode *machine learning* dapat dikategorikan berdasarkan jenis pembelajarannya (Sarkar, et al., 2018:35-42), yaitu:

1) *Supervised learning* berhubungan pada sampel data dimana respon/label keluaran yang diinginkan sudah diketahui sebelumnya. *Supervised learning* memodelkan hubungan antara *input* dan *output* yang sesuai dari data pelatihan sehingga dapat memprediksi respon *output* untuk input data baru berdasarkan pengetahuan yang diperoleh sebelumnya.
2) *Unsupervised learning* mencoba mempelajari struktur, pola, dan hubungan yang melekat dari data tanpa label keluaran yang digunakan. *Unsupervised learning* lebih berkaitan dengan mencoba mengekstraksi informasi yang bermakna dari data yang sebelumnya tidak memiliki pola yang jelas.
3) *Semi-supervised learning* berada diantara metode *supervised* dan *unsupervised learning*. Metode ini biasanya menggunakan banyak data pelatihan yang tidak berlabel (membentuk komponen *unsupervised learning*) dan sejumlah kecil data pra-label dan beranotasi (membentuk komponen *supervised learning*). Pendekatan berupa membangun model *supervised*

berdasarkan data berlabel, yang terbatas, dan kemudian menerapkan hal yang sama untuk sejumlah besar data tidak berlabel untuk mendapatkan lebih banyak sampel berlabel, melatih model dan mengulangi prosesnya.

4) *Reinforcement learning* merupakan metode *machine learning* berbasis umpan balik dimana agen belajar berperilaku di lingkungan dengan melakukan tindakan dan melihat hasil tindakan. Untuk setiap tindakan baik, agen mendapat umpan balik positif, dan untuk setiap tindakan buruk, agen mendapat umpan balik negatif atau penalti.

2.1.4 Klasifikasi

Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target *f* yang memetakan setiap set atribut (fitur) *x* ke satu dari sejumlah label kelas *y* yang tersedia (Prasetyo, 2012:45).

Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu:

1) Pembangunan model sebagai prototipe untuk disimpan sebagai memori, dan
2) Penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui dikelas mana objek data tersbut dalam model yang sudah disimpannya.

2.1.5 *Artificial Neural Network*

Menurut Fausett (1994:3), *Artificial Neural Network* (ANN) atau jaringan syaraf tiruan adalah sistem pemrosesan informasi yang memiliki karakteristik

kinerja tertentu menyerupai jaringan saraf biologis. Sebuah *Neural Network* memiliki karakteristik, yaitu:

1) Pola koneksi antar *neuron* (arsitektur),
2) Metode penentuan bobot pada koneksi (algoritma pelatihan atau pembelajaran),
3) Fungsi aktivasi.

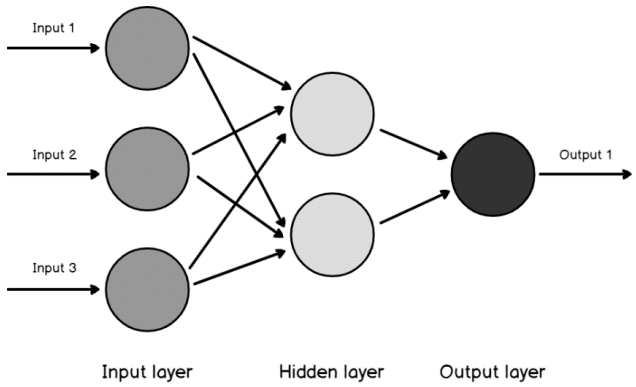
2.1.5.1 Karakteristik *Artificial Neural Network*

Karakteristik ANN dijelaskan sebagai berikut (Arhami & Nasir, 2020:123-124):

A. Arsitektur ANN

Secara umum, arsitektur ANN terdiri atas beberapa lapisan seperti pada gambar 2.1, yaitu:

(1) Lapisan masukan (*input layer*) merupakan lapisan yang terdiri dari beberapa neuron yang akan menerima sinyal dari luar dan kemudian meneruskan ke lapisan lain dalam jaringan.
(2) Lapisan tersembunyi (*hidden layer*) terletak lapisan masukan dan lapisan keluaran, berfungsi untuk meningkatkan kemampuan jaringan dalam memecahkan masalah.
(3) Lapisan keluaran (*output layer*) berfungsi untuk menyalurkan sinyal-sinyal keluaran hasil pemrosesan jaringan.



Gambar 2.1 Arsitektur ANN

(Sumber: https://medium.com/analytics-vidhya/vanishing-gradient-problem-in-deep-learning-dafb9caf2f3a)

B. Algoritma Pembelajaran

Proses pembelajaran pada ANN dapat di klasifikasikan menjadi dua bagian (Arhami & Nasir, 2020:131-132), yaitu:

(1) *Supervised learning*, yang menggunakan sejumlah pasangan data masukan dan keluaran yang diharapkan

(2) *Unsupervised learning*, yang hanya menggunakan sejumlah pasangan data masukan tanpa ada contoh keluaran yang diharapkan.

C. Fungsi Aktivasi

ANN menggunakan fungsi aktivasi yang dipakai untuk membatasi keluaran dari neuron agar sesuai dengan batasan sinyal/nilai keluarannya (Prasetyo, 2012:73).

Secara umum, ada empat macam fungsi aktivasi yang dipakai di berbagai jenis ANN (Arhami & Nasir, 2020:129-130), yaitu:

(1) Fungsi Linear

Fungsi linear atau fungsi identitas biasanya digunakan pada jaringan lapis tunggal. Fungsi linear akan menghasilkan nilai yang sama dengan nilai masukannya. Fungsi linear dituliskan pada persamaan 2.2.

$$f(x) = x \tag{2.2}$$

(2) Fungsi Tangga Biner (*Step*)

Fungsi tangga biner merupakan fungsi identitas dengan pembulatan yang bergantung pada parameter pembulatan θ. Untuk θ=1 fungsi ini hanya akan menghasilkan nilai 1 atau 0. Fungsi tangga biner dituliskan pada persamaan 2.3.

$$f\,(x) = \begin{cases} 0, & jika\ x\ \geq 0 \\ 1, & jika\ x\ < 0 \end{cases} \tag{2.3}$$

(3) Fungsi *Sigmoid* Biner

Fungsi *sigmoid* biner tergantung pada *steepness* parameter (σ). Agar fungsi ini menghasilkan nilai yang dibatasi oleh bilangan biner (0 sampai 1) maka σ = 1. Fungsi sigmoid biner dituliskan pada persamaan 2.4.

$$f(x) = \frac{1}{1+ e^{-\sigma x}} \tag{2.4}$$

(4) Fungsi *Sigmoid* Bipolar

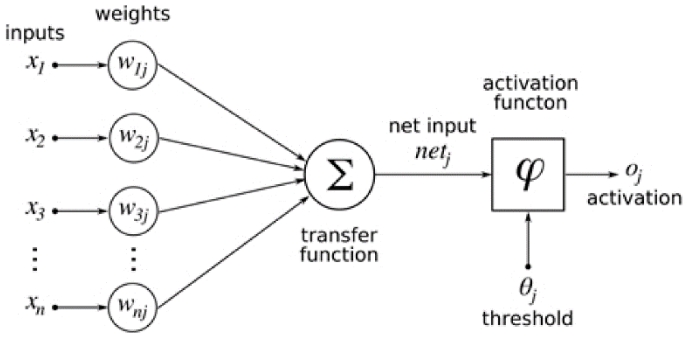
Fungsi *sigmoid* bipolar juga tergantung pada *steepness* parameter (σ). Fungsi sigmoid bipolar merupakan fungsi *sigmoid* biner yang diperluas sampai mencapai nilai negatif melalui sumbu x. Untuk σ = 1, fungsi ini akan

menghasilkan nilai keluaran antara -1 sampai +1. Fungsi sigmoid bipolar dituliskan pada persamaan 2.5.

$$g(x) = 2f(x) - 1 = \frac{2}{1+ e^{-\sigma x}} - 1 = \frac{1- e^{-\sigma x}}{1+ e^{-\sigma x}} \quad (2.5)$$

2.1.5.2 Cara Kerja *Artificial Neural Network*

Desain cara kerja ANN secara umum ditunjukkan pada Gambar 2.2.



Gambar 2.2 Desain umum ANN

(Sumber: https://www.mdpi.com/1911-8074/12/2/76)

Pada Gambar 2.2, vektor masukan terdiri atas sejumlah nilai yang diberikan sebagai nilai masukan pada ANN. Vektor masukan tersebut mempunyai tiga nilai (x_1, x_2, x_3) sebagai fitur dalam data yang akan diproses di ANN. Masing-masing nilai masukan melewati sebuah hubungan berbobot *w*, kemudian semua nilai digabungkan. Nilai gabungan tersebut kemudian diproses oleh fungsi aktivasi untuk menghasilkan sinyal *y* sebagai keluaran. Fungsi aktivasi menggunakan sebuah nilai ambang batas (*threshold*) untuk membatasi nilai keluaran agar selalu dalam batas nilai yang ditetapkan (Prasetyo, 2012:73).

2.1.5.3 Multilayer Perceptron

Multilayer Perceptron (MLP) merupakan turunan dari *Perceptron*, berupa ANN *Feedforward* dengan satu atau lebih *hidden layer*. MLP biasanya terdiri atas satu *input layer*, setidaknya satu *hidden layer*, dan satu *output layer*. Sinyal masukan dirambatkan dengan arah maju pada layer-per-layer, dari *input* ke *output* (Prasetyo, 2012:86). Tidak ada koneksi umpan balik dari *output* ke dirinya sendiri (*loop*). Pada *Feedforward* koneksi antar node tidak membentuk siklus (*loop*).

Langkah-langkah proses *feedforward* dijelaskan sebagai berikut (Arhami &Natsir, 2020:138):

1) Setiap neuron pada *input layer* (X_i, I = 1, 2, …, n) menerima sinyal masukan x_i dan menjalankan sinyal tersebut ke semua *neuron* pada *layer* selanjutnya (*hidden layer*)
2) Untuk setiap neuron pada hidden layer (Z_j, j = 1, 2, …, p) jumlahkan bobotnya dengan sinyal inputnya masing-masing

$$Z_{in_j} = v_{0\,j} + \sum_{i=1}^{n} x_i\, v_j \tag{2.6}$$

Terapkan fungsi aktivasi untuk menghitung nilai sinyal keluaran

$$Z_j = f(Z_{in_j}) \tag{2.7}$$

Kemudian kirimkan sinyal ini ke semua *neuron* pada *layer* selanjutnya (dalam hal ini adalah *output layer*)

3) Untuk setiap neuron pada output layer (Y_k, k = 1, 2, …, m) jumlahkan bobotnya dengan sinyal masukannya masing-masing

$$Y_{in_k} = w_{0\,j} + \sum_{i=1}^{p} z_j\, w_k \tag{2.8}$$

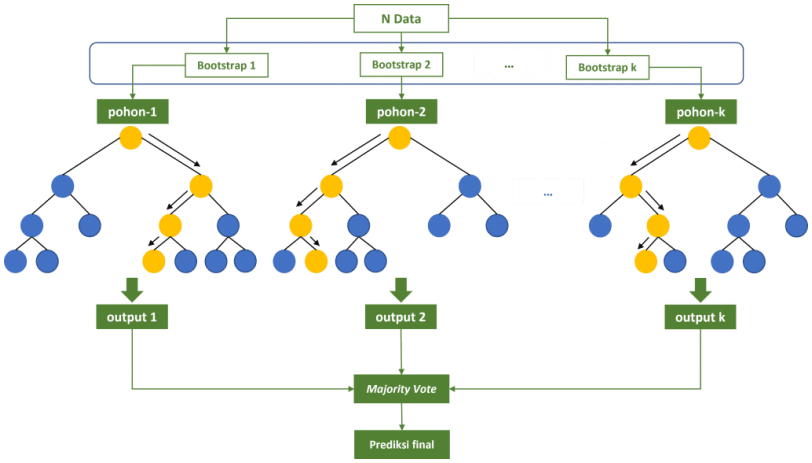
4) Terapkan fungsi aktivasi untuk menghitung sinyal *output*

$$Y_k = f(Y_{in_k}) \tag{2.9}$$

2.1.6 *Random Forest*

Random Forest adalah teknik *ensemble learning* yang bisa digunakan baik dalam kasus regresi maupun klasifikasi. *Random Forest* memanfaatkan *Decision Tree* dalam melakukan pembelajaran (Amruthnath & Gupta, 2019). *Random Forest* merupakan metode pengembangan dari CART (*Classification and Regression Tree*) dengan menerapkan proses *bootstrap aggregating* (*bagging*) dan *random feature selection* (Breiman, 2001).

Random Forest menggunakan teknik *bagging* yang merupakan singkatan dari *bootstrap aggregating*. Teknik tersebut mengambil subsampel acak dari data pelatihan dan memasukannya kedalam versi berbeda dengan model yang sama dan membiarkan semua model memilih hasil akhir. *Random Forest* menggunakan banyak *Decision Tree* yang berbeda menggunakan sampel acak berbeda dari data pelatihan untuk dilatih, dan kemudian semuanya bersatu untuk memberika suara pada hasil akhir (Kane, 2017).



Gambar 2.3 Random forest

(Sumber: https://sainsdata.id/machine-learning/893/python-random-forest-untuk-model-klasifikasi-menggunakan-scikitlearn/)

Langkah-langkah dalam menerapkan algoritma *Random Forest* dijelaskan sebagai berikut (Amruthnath & Gupta, 2019):

1) Membuat sampel *bootstrap* sebanyak n_{tree} berdasarkan data asli. *Bootstrap* adalah proses *resampling dataset* dengan mengambil sejumlah baris dan kolom secara acak dengan metode *row sampling with replacement*. Artinya antara satu *bootstrap* dan *bootstrap* yang lain akan berbeda.
2) Untuk setiap sampel *bootstrap* akan dikembangkan *decision tree* yang tidak dipangkas (*unprunned*). Pada umumnya ketika membentuk *decision tree*, *node* diambil dari nilai variabel terbaik dengan mempertimbangkan seluruh variabel yang dihitung menggunakan metode tertentu, salah satunya (*Gini Impurity*). Akan tetapi, pada *Random Forest node* diambil secara acak dengan mengambil sejumlah m_{txy} variabel.

3) Prediksi data dengan melakukan agregasi dari setiap *decision tree* yang telah terbentuk (pada klasifikasi menggunakan *majority vote* sementara pada kasus regresi menggunakan *average*).

2.1.7 Evaluasi Model

Evaluasi adalah cara standar untuk mengukur kinerja model. Ukuran kinerja model klasifikasi berfungsi untuk mengevaluasi seberapa baik model klasifikasi dapat mengklasifikasikan amatan dengat tepat.

2.1.7.1 *Confusion Matrix*

Confusion matrix merupakan tabel yang menggambarkan performa dari sebuah model atau algoritma secara spesifik. Setiap baris dari *matrix* tersebut merepresentasikan kelas aktual dari data dan setiap kolom merepresentasikan kelas prediksi dari data (atau sebaliknya) (Saputro & Sari, 2019). *Confusion matrix* dijelaskan pada Tabel 2.3.

Tabel 2.3 *Confusion Matrix*

	Predicted Positive	*Predicted Negative*
Actual Positive	*True Positive* (TP)	*False Negative* (FN)
Actual Negative	*False Positive* (FP)	*True Negative* (TN)

(1) *True Positive* = Banyak data yang aktual kelas positif, dan model memprediksi positif.

(2) *True Negative* = Banyak data aktual kelas negatif, dan model memprediksi negatif.

(3) *False Positive* = Banyak data aktual kelas negatif, namun model memprediksi positif.

(4) *False Negative* = Banyak data aktual kelas positif, namun model memprediksi negatif.

Melalui 4 data tersebut, dapat diperoleh data-data lain yang sangat berguna untuk mengukur perfoma sebuah model, diantaranya:

(1) *Accuracy* menggambarkan seberapa akurat model dalam mengklasifikasikan dengan benar.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.10)$$

(2) *Precision* merupakan jumlah prediksi yang bernilai benar atau relevan dari semua prediksi berdasarkan kelas positif.

$$Precision = \frac{TP}{TP+FP} \quad (2.11)$$

(3) *Recall* atau sensitivity merupakan pengukuran proporsi positif yang diidentifikasi dengan benar dan juga disebut true positive rate.

$$Recall = \frac{TP}{TP+FN} \quad (2.12)$$

(4) *F1-score* merupakan rata-rata harmonis dari precision dan recall, serta membantu mengoptimalkan pengklasifikasi untuk kinerja precision dan recall yang seimbang.

$$f1 - score = \frac{2*precision*recall}{precision+recall} \quad (2.13)$$

2.2 Penelitian Terkait

Penelitian terkait yang menjadi acuan penelitian dijelaskan pada Tabel 2.4

Tabel 2.4 Penelitian terkait

Penelitian 1	
Penulis	Faqih Hamami, Inayatul Fithriyah
Tahun	2020
Judul	*Classification of Air Polution Levels using Artificial Neural Network*
Jurnal	2020 International Conference on Information Technology Systems and Innovation (ICITSI) Bandung - Padang, October 19 - 23, 2020
Metode	Feedforward Neural Network
Hasil	Artificial Neural Network digunakan untuk klasifikasi polusi udara tahun 2017 di DKI Jakarta. Model terbaik adalah Neural Network dengan 1 hidden layer dan 1500 epoch yang memperoleh akurasi 96.61%. Hampir semua percobaan memperoleh nilai sensitivitas dan spesifisitas di atas 90%.
Penelitian 2	
Penulis	O F Althuwaynee, A L Balogun, A Aydda, T Gumbo
Tahun	2019
Judul	*Classification of air pollutants API Inter-Correlation using decision tree algorithms*
Jurnal	ICCEE 2019 IOP Conf. Series: Earth and Environmental Science 419 (2020) 012022
Metode	Bosted C5.0 Random Forest PART Naive Bayes Tree
Hasil	Beberapa jenis algoritma berbasis Decision tree digunakan untuk mencari keterkaitan antara indeks polusi udara (API) nilai persentil PM_{10} dengan empat polutan udara lainnya. Hasil penelitian menemukan bahwa CO, SO_2, dan O_3 adalah polutan utama yang secara signifikan berhubungan dengan PM_{10} dan berkontribusi terhadap penurunan kualitas udara di Pulau Pinang. PART dan Random forest merupakan algoritma yang paling stabil dengan menghasilkan keputusan yang jelas tentang variabel terpenting yang saling berkorelasi dengan nilai API PM_{10}.
Penelitian 3	
Penulis	Hualing Yi, Qingyu Xiong, Qinghong Zou, Rui Xu, Kai Wang, Min Gao
Tahun	2019
Judul	*A Novel Random Forest and its Application on Classification of Air Quality*

Tabel 2.4 Penelitian terkait (Lanjutan 1)

Jurnal	2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)
Metode	Random Forest Random Forest + Sample Grouped Bootstrap (SGB-RF)
Hasil	Penelitian dilakukan pada tiga set percobaan, disimpulkan bahwa kinerja Random Forest tidak bekerja baik pada data yang tidak seimbang. SGB-RF bekerja lebih baik dari Random Forest ketika keduanya diterapkan pada data set yang seimbang. Peningkatan signifikan ketika SGB-RF diterapkan pada kumpulan data yang tidak seimbang, dimana SGB-RF jauh lebih baik daripada RF dengan kemampuan yang kuat untuk mengklasifikasikan sampel minoritas dengan benar.
Penelitian 4	
Penulis	Abdul Aziiz Hendrie Kirono, Ibnu Asror, Yanuar Firdaus Arie Wibowo
Tahun	2022
Judul	Klasifikasi Tingkat Kualitas Udara DKI Jakarta Menggunakan Algoritma Naive Bayes
Jurnal	e-Proceeding of Engineering : Vol.9, No.3 Juni 2022 (Page 1962)
Metode	Naïve Bayes
Hasil	Pengklasifikasian pada data Indeks Standar Pencemaran Udara (ISPU) di DKI Jakarta dengan metode Naïve Bayes menghasilkan hasil klasifikasi dengan rata-rata akurasi 88%, precision 85%, recall 96%, f1-score 90%. Penggunaan seleksi fitur Pearson Correlation terhadap data ISPU di kota DKI Jakarta sangat berperan penting terhadap peningkatan akurasi dari permodelan klasifikasi dengan algoritma Naive Bayes.
Penelitian 5	
Penulis	Krittakom Srijiranon, Narissara Eiamkanitchat
Tahun	2018
Judul	*Collective Neural Networks System for PM_{10} Classification in the North of Thailand*
Jurnal	2018 22nd International Computer Science and Engineering Conference (ICSEC)
Metode	Collective Neural Network (Backpropagation Neural Network)
Hasil	Dilakukan beberapa kali percobaan dengan mengubah proporsi data training dan data testing untuk setiap percobaan. Hasil percobaan dataset sepanjang tahun lebih baik dibandingkan dengan dataset periode kritis karena periode kritis hanya 4 bulan. Hasil rata-rata tingkat akurasi dari semua percobaan adalah 90%. Sistem jaringan saraf kolektif yang diusulkan menunjukkan bahwa jaringan saraf lebih akurat dibandingkan algoritma lain yang telah digunakan pada percobaan sebelumya.

Tabel 2.4 Penelitian terkait (Lanjutan 2)

Penelitian 6	
Penulis	Antipas T. Teologo Jr, Elmer P. Dadios, Renann G. Baldovino, Romano Q. Neyra and Irister M. Javel
Tahun	2018
Judul	*Air Quality Index (AQI) Classification using CO and NO_2 Pollutants: A Fuzzy-based Approach*
Jurnal	Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference (Jeju, Korea, 28-31 October 2018)
Metode	Mamdani Fuzzy Inference Systen (FIS)
Hasil	Model Fuzzy Logic yang digunakan dalam perhitungan nilai AQI dan penentuan kategori AQI yang sesuai telah terbukti handal dan efektif. Semua hasil uji coba sesuai dengan output AQI yang diharapkan. Hasil ilustrasi menunjukkan Pada ilustrasi 3D pemetaan input-output menunjukkan bahwa variabel input dan output saling berhubungan, maka semakin tinggi konsentrasi polutan maka nilai AQI juga semakin meningkat. Kekurangan dari penelitian tersebut adalah hanya dengan menggunakan 2 variabel input. Tetapi, Penambahan input lain berarti membutuhkan basis pengetahuan yang lebih kompleks.
Penelitian 7	
Penulis	Chou-Yuan Lee, Zne-Jung Lee, Jian-Qiong Huang, Fu-Lan Ye, Zheng-Yuan Ning, Cheng-Fu Yang
Tahun	2019
Judul	*Urban Air Quality Analysis and Forecast Based on Intelligent Algorithm with Parameter Optimization and Decision Rules*
Jurnal	Appl. Sci. 2019, 9, 5445; doi:10.3390/app9245445
Metode	Decision Tree dengan Simulated Annealing
Hasil	Dengan kombinasi Decision Tree dan Simulated Annealing (SA), dari percobaan yang dilakukan menghasilkan akurasi klasifikasi data pelatihan adalah 99,92%. Hasil percobaan menunjukkan bahwa SA dapat digunakan untuk menyesuaikan pengaturan parameter terbaik untuk DT dan mencapai akurasi yang lebih baik untuk klasifikasi.
Penelitian 8	
Penulis	Wan Nur Shaziayani, Ahmad Zia Ul-Saufie, Sofianita Mutalib, Norazian Mohamad Noor, Nazatul Syadia Zainordin
Tahun	2022
Judul	Classification Prediction of PM10 Concentration Using a Tree-Based Machine Learning Approach
Jurnal	Atmosphere 2022, 13, 538. https://doi.org/10.3390/atmos13040538

Tabel 2.4 Penelitian terkait (Lanjutan 3)

Metode	Decision Tree Boosted Regression Tree (Boosting) Random Forest (Bagging)
Hasil	Hasil penelitian menunjukkan bahwa algoritma pembelajaran mesin dapat membantu mengklasifikasikan konsentrasi PM_{10} untuk hari berikutnya sebagai indikasi kualitas udara tingkat rendah atau tinggi. Tiga metode klasifikasi yang diterapkan dalam penelitian ini juga menunjukkan fitur yang paling relevan dalam prediksi konsentrasi PM_{10}. Random Forest menunjukkan hasil paling baik dalam memprediksi klasifikasi konsentrasi PM_{10} hari berikutnya, dengan akurasi 98,37, sensitivitas 97,19, spesifisitas 99,55, dan presisi 99,54. Pada RF dan DT wind speed adalah fitur paling relevan untuk mengklasifikasikan konsentrasi PM_{10} hari berikutnya, tetapi untuk BRT, PM_{10} adalah fitur yang paling relevan.
Penelitian 9	
Penulis	K. Kumar, B. P. Pande
Tahun	2022
Judul	*Air pollution prediction with machine learning: a case study of Indian cities*
Jurnal	International Journal of Environmental Science and Technology https://doi.org/10.1007/s13762-022-04241-5
Metode	k-NN Gaussian Naive Bayes (GNB) SVM Random Forest XGBoost
Hasil	Untuk fase pelatihan dan pengujian, XGBoost menunjukkan akurasi paling tinggi dan model SVM menunjukkan akurasi terendah. Model XGBoost menunjukkan performa terbaik secara keseluruhan dengan mencapai nilai optimal dalam set pelatihan dan pengujian. Model GNB menunjukkan nilai terbaik untuk R2 dalam set pengujian. Semua model ML menunjukkan peningkatan di hampir semua metrik penilaian ketika diterapkan dengan teknik resampling SMOTE.
Penelitian 10	
Penulis	Brainvendra Widi Dionova, M.N. Mohammed, S. Al-Zubaidi, Eddy Yusuf
Tahun	2020
Judul	Environment indoor air quality assessment using fuzzy inference system
Jurnal	ICT Express Volume 6, Issue 3, September 2020, Pages 185-194 https://doi.org/10.1016/j.icte.2020.05.007
Metode	Fuzzy Inference System (FIS)

Tabel 2.4 Penelitian terkait (Lanjutan 4)

Hasil	Peran utama dari Fuzzy Inference System adalah untuk mengembangkan Environment of Indoor Air Quality Index (EIAQI) dengan memanfaatkan teori fuzzy, yang meningkatkan efisiensi penilaian kualitas udara dan koordinasi tingkat konsentrasi tertentu dalam indeks fuzzy. Sistem ini menggunakan sistem clustering untuk menghitung Indoor Air Quality Index (IAQI) dan Thermal Comfort Index (TCI) secara terpisah, karena kedua parameter indeks ini memiliki karakteristik dan dampak bagi kesehatan manusia yang berbeda. Hasil klasifikasi EIAQI digunakan sebagai dasar untuk menentukan output berupa tindakan untuk mengurangi polusi udara dalam ruangan dan meningkatkan kualitas udara dengan memanfaatkan metode Fuzzy Logic Controller.
Penelitian 11	
Penulis	Syekh S A Umri, Muhammad S Firdaus, Aji Primajaya
Tahun	2021
Judul	Analisis dan Komparasi Algoritma Klasifikasi dalam Indeks Pencemaran Udara di DKI Jakarta
Jurnal	JIKO (Jurnal Informatika dan Komputer) Vol. 4, No. 2, Agustus 2021, hlm. 98-104 DOI: 10.33387/jiko
Metode	Neural Network Backpropagation Support Vector Machine K-Nearest Neighbor Naive Bayes Decision Tree
Hasil	Algoritma dengan performa terbaik yakni Decission Tree dengan nilai akurasi sebesar 99.80%, nilai kappa yang hampir sempurna yakni 0.996, nilai RMSE terkecil dan di bawah 0.1 yakni 0.039, serta waktu yang dibutuhkan hanya 0.8 detik. Meskipun begitu, Neural Network Backpropagation, KNN, SVM, dan Naive Bayes juga masih dapat digunakan sebagai model klasifikasi yang baik karena mendapatkan nilai akurasi yang tinggi di atas 90% dan nilai kappa di atas 0.8.
Penelitian 12	
Penulis	Adinda Inez Sang, Edi Sutoyo, Irfan Darmawan
Tahun	2021
Judul	Analisis Data Mining untuk Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma Decision Tree dan Support Vector Machine
Jurnal	e-Proceeding of Engineering : Vol.8, No.5 Oktober 2021 Page 8954
Metode	Decision Tree SVM

Tabel 2.4 Penelitian terkait (Lanjutan 5)

Hasil	Algoritma Decision Tree dengan akurasi terbaik dengan rasio perbandingan data training dan data adalah 90:10 yaitu sebanyak 99,40%. Algoritma SVM dengan akurasi terbaik dengan rasio perbandingan data training dan data adalah 60:40 yaitu sebanyak 94,93%. Algoritma Decision Tree memiliki nilai akurasi yang lebih unggul daripada algoritma SVM untuk melakukan klasifikasi kualitas udara di DKI Jakarta, dibuktikan dengan nilai akurasi disetiap rasio data yang dilakukan oleh algoritma Decision Tree lebih tinggi daripada algoritma SVM. Decision Tree juga menghasilkan performa yang lebih baik dibandingkan dengan algoritma SVM, baik dari nilai Precision, Recall dan F1-Measure.
Penelitian 13	
Penulis	Sami Tlais, Hassan Hajj Hussein, Fouad Sakr, Mohamad Hallani, Abdel-Mehsen Ahmad, Zouhair El-Bazzal
Tahun	2020
Judul	*Air Quality Monitoring and Classification Using Machine Learning*
Jurnal	S. M. Thampi et al. (Eds.): SoMMA 2019, CCIS 1203, pp. 135–143, 2020. https://doi.org/10.1007/978-981-15-4301-2_11
Metode	K-Nearest Neighbor (KNN) Support Vector Machine (SVM) Multilayer Perceptron (MLP) Naïve Bayes Classifier
Hasil	Algoritma klasifikasi yang digunakan dalam percobaan menunjukkan hasil yang baik, semuanya memiliki akurasi di atas 90%. Artificial Neural Network memiliki ukuran kinerja terbaik pada kumpulan data polusi udara, dan akurasi tertinggi di antara KNN, SVM, dan Naïve Bayes Classifier. Data cuaca dan polutan dikumpulkan melalui sensor (MQ-135, MQ-131, MQ-7, DHT21). Data yang terkumpul kemudian diolah dengan model MLP untuk memprediksi polusi udara di beberapa lokasi dengan menggunakan aplikasi WEKA. Akurasi MLP yang diatas 98% mewakili model yang baik yang akan membantu dalam memprediksi kelas data baru.
Penelitian 14	
Penulis	Zulfia Sari Permana
Tahun	2021
Judul	Implementasi Algoritma Decision Tree dan K-Nearest Neighbor Pada Klasifikasi Indeks Standar Pencemar Udara (ISPU)
Metode	Decision Tree K-Nearest Neighbour

Tabel 2.4 Penelitian terkait (Lanjutan 6)

Hasil	Berdasarkan hasil klasifikasi dan uji coba pada Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta menggunakan algoritma Decision Tree dan K-Nearest Neighbor, disimpulkan bahwa dengan menggunakan algoritma Decision Tree maka didapatkan akurasi yang cukup tinggi, hasil pengukuran akurasi data yang diperoleh mencapai 99.75% dan hasil akurasi pada algoritma K-Nearest Neighbor lebih kecil dari Decision tree dengan hasil akurasi K = 5 yaitu 93.92%, K = 7 yaitu 93.97% dan K = 9 yaitu 93.88%.
Penelitian 15	
Penulis	Bo Liu, Xingrui Li, Jianqiang Li, Yong Li, Jianlei Lang, Rentao Gu, Fei Wang
Tahun	2018
Judul	*Comparison of Machine Learning Classifiers for Breast Cancer Diagnosis Based on Feature Selection*
Jurnal	2018 IEEE International Conference on Systems, Man, and Cybernetics
Metode	Decision Tree SVM Random Forest Adaboost Classifiers
Hasil	Mengklasifikasikan data sell kanker payudara kedalam kanker jinak atau ganas. Dari hasil dengan menggunakan 3 metode Feature Selection, Random Forest menunjukkan akurasi, sensitifitas, dan spesifikasi rata-rata diatas 0,90. Dari hasil PCA dengan hanya mempertahankan fitur yang dihasilkan melalui transformasi PCA, Random Forest menunjukkan hasil paling baik diantara yang lain yaitu tingkat akurasi 99%. Secara keseluruhan, Random forest menunjukkan hasil yang paling baik.
Penelitian 16	
Penulis	Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-Viedma, Hamido Fujita, Nor Azura Md. Ghani
Tahun	2021
Judul	*Multiclass Prediction Model for Student Grade Prediction Using Machine Learning*
Jurnal	IEEE Access Volume: 9, 2021 DOI: 10.1109/ACCESS.2021.3093563
Metode	Naive Bayes Linear Regression SVM k-NN J48 Random forest

Tabel 2.4 Penelitian terkait (Lanjutan 7)

Hasil	Integrasi metode oversampling SMOTE dan Featue Selection meningkatkan kinerja multi-klasifikasi yang tidak seimbang pada dataset peneltian. Akurasi terbaik yang diperoleh Random Forest dengan 99,5% sedikit lebih tinggi dari kNN dan J48 menunjukkan bahwa algoritma RF adalah algoritma solusi yang ideal untuk memprediksi nilai akhir siswa.
Penelitian 17	
Penulis	Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni
Tahun	2021
Judul	*Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison*
Jurnal	Computers in Biology and Medicine Volume 136, September 2021, 104672 https://doi.org/10.1016/j.compbiomed.2021.104672
Metode	Multilayer Perceptron K-Nearest Neighbours Random Forest Decision Tree Logistic Regression AdaboostM1
Hasil	Tiga algoritma klasifikasi KNN, RF dan DT bekerja sangat baik dengan akurasi 100%. Selain itu, feature importance scores untuk setiap fitur diterapkan untuk semua algoritma kecuali MLP dan KNN. Fitur ini memberi peringkat pada setiap variabel berdasarkan skor kepentingan fitur. Penelitian ini bertujuan untuk menemukan teknik machine learning terbaik, di antara sejumlah algoritma yang banyak digunakan dan mudah diimplementasikan. Untuk dataset yang didapatkan dari Kaggle, semua algoritma bekerja baik.

2.3 Matrik Penelitian

Matrik penelitian menjelaskan hubungan penelitian terdahulu yang terkait dengan penelitian yang dilakukan. Dalam matrik penelitian menunjukan pendekatan yang berbeda-beda dalam menemukan sebuah solusi. Matrik penelitian tertera pada Tabel 2.5.

Tabel 2.5 Matrik Penelitian

Peneliti	Dataset / Variabel	Praproses	Metode	Pengujian
Hamami & Fitriyah, 2020	Data polusi udara Jakarta tahun 2017 Parameter PM_{10}, SO_2, CO, O_3, NO_2	Normalisasi kelas target *Split* kolom data Menghapus nilai string *Encoding* data kategori	Feedforward Neural Network	*Sensitivity* *Specificity* *Accuracy*
Althuwaynee, et al., 2019	Data stasiun pengawasan dari 1 Maret 2005 - 31 Desember 2006 di Penang Island, Malaysia utara. Nilai API perjam untuk PM_{10}, SO_2, O_3, NO_2, dan CO. Parameter metereologi suhu (^oC) dan kelembaban (%).		Bosted C5.0 Random Forest PART Naive Bayes Tree	*Accuracy* RSME MAE *Run time*
Yi, et al., 2019	dataset1 data udara Beijing. dataset2 dan dataset3 data udara Fangchenggang, provinsi Guangxi. Dataset4 data udara Beijing dan Fangchenggang.		Random Forest Random Forest + Sample Grouped Bootstrap (SGB-RF)	Out of Bag Score

Tabel 2.5 Matrik Penelitian (Lanjutan 1)

Peneliti	Dataset / Variabel	Praproses	Metode	Pengujian
Kirono, et al., 2022	Data Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta bulan Februari - Oktober 2021 PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2	Pembersihan data yang tidak lengkap, kosong, noise, duplikat, dan data yang tidak konsisten	Naïve Bayes	*Precisiom* *Recall* *F1-score* *Support*
Srijiranon & Eiamkanitchat, 2018	Data dari Pollution Control Department (PCD) Ministry of Natural Resources and Environment Thailand Lampang Meteorogical Station (LM-sta) 20 Januari 2013 - 31 Desember 2017 Knowledge Park Station (KP-sta) 01 Januari 2011 - 31 Desember 2017	Eliminasi data kosong Normalisasi min-max	Collective Neural Network	*Accuracy*
Teologo, et al., 2018	Parameter CO dan NO_2		Mamdani Fuzzy Inference Systen (FIS)	
Lee, et al., 2019	AQI 01 Januari 2017 - 06 Oktober 2018 di daerah Dongcheng, Beijing		Decision Tree dengan Simulated Annealing	*Accuracy*
Shaziayani, et al., 2022	Data kualitas udara dari Januari 2002 - Desember 2017 di Kota Bharu, Kelantan, Malaysia.	Mengisi data kosong : Linear interpolation Mengatasi *imbalance* data : Synthetic Minority Oversampling Technique (SMOTE)	Decision Tree Boosted Regression Tree (Boosting) Random Forest (Bagging)	*Accuracy* *Sensitivity* *Specificity* *Precision*

Tabel 2.5 Matrik Penelitian (Lanjutan 2)

Peneliti	Dataset / Variabel	Praproses	Metode	Pengujian
Kumar & Pande, 2022	Data polutan udara 23 kota di India dari Januari 2015 - Juli 2020.	Mengisi data kosong dengan median Deteksi dan penghapusan outlier Normalisasi data Algoritma pemilihan fitur berbasis korelasi statistik SMOTE (Synthetic Minority Oversampling Technique)	k-NN Gaussian Naive Bayes (GNB) SVM Random Forest XGBoost	*Accuracy* *Precision* *Recall* *F1-score* *Training time* MAE RMSE *Root Mean Squared Logarithmic Error* R^2
Dionova, et al., 2020	Parameter CO_2, CO, NO_2, O_3, VOC, $PM_{2.5}$, suhu, dan kelembaban		Fuzzy Inference System	
Umri, et al., 2021	Dataset Indeks Pencemar Standar Udara sejak tahun 2017 hingga Juni 2020. CO, SO_2, NO_2, 0_3, dan PM_{10}	Menghapus data yang memiliki nilai kosong, data duplikat, data yang terisi namun tidak memiliki data secara lengkap, dan penghapusan atribut yang tidak digunakan	Neural Network Backpropagation Support Vector Machine K-Nearest Neighbor Naive Bayes Decision Tree	*Accuracy* Kappa RMSE Waktu
Sang, et al., 2021	Data ISPU DKI Jakarta Dari Januari - Desember 2020. PM_{10}, SO_2, CO, O_3, NO_2	Penggabungan data Data *selection* Data *cleaning*	Decision Tree SVM	Accuracy Precisiom Recall F1-score
Tlais, et al., 2020	20000 records data 6 polutan udara utama di Australia.	Data *cleaning*	KNN, SVM, MLP, Naïve Bayes Classifier	*Accuracy* MAE RMSE

Tabel 2.5 Matrik Penelitian (Lanjutan 3)

Peneliti	Dataset / Variabel	Praproses	Metode	Pengujian
Permana, 2021	Data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta dari 1 Januari 2017 - 28 Februari 2021.	Data selection Data cleaning	Decision Tree K-Nearest Neighbour	*Accuracy*
Liu, et al., 2018	Dataset gambar digital proses Fine Needle Aspiration (FNA) dari massa jaringan payudara.	Mengganti data kosong dengan nilai tengah (median) Standarisasi dan normalisasi data	Decision Tree SVM Random Forest Adaboost Classifiers	*Accuracy* *Sensitivity* *Specificity*
Bujang, et al., 2021	Jumlah nilai mata kuliah mahasiswa semester I dari ujian akhir semester bulan Juni 2016 - Desember 2019	Oversampling teknik : SMOTE	Naive Bayes Linear Regression SVM k-NN J48 RF	*Accuracy* *Precision* *Recall* *F-Measure*
Ali, et al., 2021	Dataset penyakit jantung dari Kaggle.	Synthetic minority oversampling technique (SMOTE)	Multilayer Perceptron K-Nearest Neighbour Random Forest Decision Tree Logistic Regression AdaboostM1	*Accuracy* *Precision* Recall *F-Measure* Kappa MCC
Siti Haliza Nur Shofa, 2023	Data Indeks Standar Pencemaran Udara DKI Jakarta dari 1 Januari – 31 Desember 2021 PM_{10}, $PM_{2.5}$, CO, NO_2, SO_2, O_3	Penggabungan data Menghapus data kosong Seleksi data Normalisasi data Label Encoder	Multilayer Perceptron Random Forest	*Accuracy* *Precision* *Recall* *F1-score*

Berdasarkan studi pustaka yang telah dilakukan, metode *machine learning* telah banyak digunakan pada berbagai bidang, salah satunya adalah klasifikasi kualitas udara.

Penelitian ini akan melakukan perbandingan model klasifikasi dengan metode *Multilayer Perceptron* dan *Random Forest*. Mengacu pada penelitian sebelumnya yang dilakukan oleh Tlais (2020), metode *Multilayer Perceptron* (MLP) menghasilkan nilai akurasi tertinggi dibandingkan dengan 3 metode *machine learning* lain. Kemudian penelitian yang dilakukan Shaziayani (2022), *Random Forest* menghasilkan nilai akurasi paling tinggi diantara *machine learning* berbasis *tree* lainnya. Model klasifikasi dengan metode *Multilayer Perceptron* dan *Random Forest* akan dibandingkan pada proses evaluasi kinerja model klasifikasi melalui perhitungan nilai *accuracy*, *precision*, *recall*, dan *f1-score*. Model klasifikasi dengan kinerja paling optimal akan digunakan untuk evaluasi prediksi kategori ISPU pada data polutan yang belum diketahui kategorinya..

Penelitian ini memiliki kesamaan dengan penelitian yang dilakukan oleh Permana (2021) dan Kirono (2022) dalam penggunaan data set ISPU di DKI Jakarta, tetapi terdapat perbedaan pada metode *machine learning* yang digunakan dan keduanya tidak melakukan proses prediksi pada data baru. Perbedaan pada tahapan praproses data dan pemilihan *hyperparameter* dalam membangun model klasifikasi menjadi perbedaan antara penelitian ini dan penelitian yang telah dilakukan sebelumnya.