

***CLUSTERING POLUSI UDARA BERDASARKAN FAKTOR METEOROLOGI
BERBASIS *PRINCIPAL COMPONENT*
*ANALYSIS****

TESIS

**Oleh :
MAIMUNA SYARIEFATUSSHOLEHAH
NIM 230605220005**



**PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

HALAMAN PENGAJUAN

***CLUSTERING* POLUSI UDARA BERDASARKAN FAKTOR METEOROLOGI
BERBASIS *PRINCIPAL COMPONENT*
*ANALYSIS***

TESIS

Diajukan kepada:

**Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)**

Oleh :

**MAIMUNA SYARIEFATUSSHOLEHAH
NIM 230605220005**

**PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

***CLUSTERING POLUSI UDARA BERDASARKAN FAKTOR METEOROLOGI
BERBASIS *PRINCIPAL COMPONENT*
ANALYSIS***

TESIS

Diajukan kepada:

**Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Magister Komputer (M.Kom)**

**Oleh :
MAIMUNA SYARIEFATUSSHOLEHAH
NIM 230605220005**

**PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

HALAMAN PERSETUJUAN

**CLUSTERING POLUSI UDARA BERDASARKAN FAKTOR METEOROLOGI
BERBASIS *PRINCIPAL COMPONENT*
*ANALYSIS***

TESIS

Oleh :
MAIMUNA SYARIEFATUSSHOLEHAH
NIM 230605220005

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 5 Juni 2026

Pembimbing I,

Dr. Agung Teguh Wibowo Almais, S.kom., M.T
NIP. 198603012023211016

Pembimbing II,

Dr. M. Imamudin Lc, MA
NIP. 19740602 200901 1 010

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T
NIP.19740510 200501 1 007

**CLUSTERING POLUSI UDARA BERDASARKAN FAKTOR METEOROLOGI
BERBASIS *PRINCIPAL COMPONENT*
*ANALYSIS***

TESIS

Oleh :
MAIMUNA SYARIEFATUSSHOLEHAH
NIM 230605220005

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 5 Juni 2026

Susunan Dewan Penguji

Penguji I : Dr. Mokhamad Amin Hariyadi, M.T.
NIP. 19670018 200501 1 001

Penguji II : Prof. Dr. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

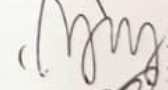
Pembimbing I : Dr. Agung Teguh Wibowo Almais, M.T
NIP. 198603012023211016

Pembimbing II : Dr. M. Imamudin Lc, MA
NIP. 19740602 200901 1 010

Tanda Tangan

()

()

()

()

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Prof. Dr. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

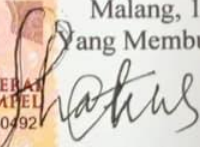
Nama : Maimuna Syariefatussholehah
NIM : 230605220005
Program Studi : Magister Informatika
Fakultas : Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Tesis yang saya tulis ini benar-banar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Tesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.



Malang, 19 Juni 2026
Yang Membuat Pernyataan,


MAIMUNA SYARIEFATUSSHOLEHAH
230605220005

MOTTO

"Success is not final, failure is not fatal: it is the courage to continue that counts."

Kesuksesan bukanlah akhir, kegagalan bukanlah kehancuran yang terpenting adalah keberanian untuk terus melangkah.

"Every delay carries a wisdom, and every struggle shapes a purpose."

Setiap penundaan mengandung hikmah, dan setiap perjuangan membentuk tujuan.

HALAMAN PERSEMBAHAN

Assalamu'alaikum Warahmatullahi Wabarakatuh.

Kupersembahkan karya Tesis terbaikku ini kepada:

1. Kedua orang tua tercinta, Bapak Noli Satroli dan Ibu Umi Kulsum, yang senantiasa menjadi sumber doa, kasih sayang, dukungan, dan kekuatan terbesar dalam setiap langkah kehidupan penulis. Terima kasih atas setiap pengorbanan yang tidak pernah dihitung, setiap doa yang tidak pernah terputus, serta setiap cinta yang selalu menguatkan penulis dalam menghadapi berbagai tantangan. Semoga karya sederhana ini dapat menjadi salah satu bentuk bakti, rasa syukur, dan kebanggaan untuk Bapak dan Ibu yang telah menjadi alasan terbesar penulis untuk terus berjuang hingga sampai pada titik ini.
2. Keluarga besar tercinta, yang senantiasa memberikan doa, perhatian, dan dukungan sehingga penulis dapat menyelesaikan tesis ini dengan baik.
3. Erra Regita Melati Putri, sahabat sekaligus keluarga yang selalu hadir dengan ketulusan, dukungan, dan kebersamaan dalam berbagai situasi. Terima kasih telah menjadi tempat berbagi cerita, tawa, dan semangat selama perjalanan ini.
4. Nicholas Evan Sitanggang, yang telah menebarkan kebahagiaan, menghadirkan energi positif, serta menjadi salah satu alasan bagi penulis untuk terus bangkit dan melangkah dengan lebih semangat dalam menyelesaikan perjalanan ini.
5. Nareswari Tarisa Kirana, yang selalu membawa kebahagiaan, energi positif, dan motivasi untuk terus berjuang. Terima kasih atas energi baik yang diberikan sehingga penulis dapat melalui berbagai tantangan dengan lebih optimis dan penuh keyakinan.
6. Dan yang terakhir, untuk diri saya sendiri, Maimuna Syariefatussholehah. Terima kasih telah bertahan melewati hari-hari yang tidak pernah diketahui orang lain. Terima kasih karena tetap melangkah meskipun harus menghadapi ketakutan, keraguan, dan kegagalan yang datang silih berganti. Hanya diri ini dan Allah yang tahu berapa banyak doa, air mata, serta perjuangan yang telah dilalui untuk sampai pada titik ini. Hari ini menjadi bukti bahwa apa yang Allah janjikan tidak pernah salah alamat. Terima kasih karena tidak menyerah dan terus percaya bahwa setiap proses memiliki tujuan yang indah pada waktunya.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh.

Puji syukur kehadiran Allah *Subhānahu wa Ta'ālā* atas segala rahmat, taufik, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan studi pada Program Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang serta menyelesaikan tesis ini dengan baik. Shalawat serta salam senantiasa tercurah kepada Nabi Muhammad SAW yang telah membawa umat manusia menuju jalan yang penuh ilmu pengetahuan dan keberkahan.

Penyusunan tesis ini tidak terlepas dari bantuan, dukungan, dan bimbingan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

1. Allah *Subhānahu wa Ta'ālā* atas segala rahmat, karunia, kesehatan, dan kemudahan yang diberikan kepada penulis.
2. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Prof. Dr. H. Agus Mulyono, S.Pd., M.Kes., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T., selaku Ketua Program Studi Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
5. Dr. Agung Teguh Wibowo Almais, M.T., selaku Dosen Pembimbing I yang telah memberikan bimbingan, arahan, dan motivasi selama proses penyusunan tesis.
6. Dr. M. Imamudin Lc, MA., selaku Dosen Pembimbing II yang telah memberikan bimbingan, masukan, dan arahan selama penyusunan tesis.
7. Dr. Mokhamad Amin Hariyadi, M.T., selaku Dosen Penguji I yang telah memberikan kritik dan saran yang membangun.
8. Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T., selaku Dosen Penguji II yang telah memberikan masukan untuk penyempurnaan tesis ini.

9. Bapak Ahmad Zarkoni, S.Kom., dan Bapak Mohammad Mamba'ul Ulum selaku pembimbing lapangan dari BMKG yang banyak membantu penulis dalam pengumpulan dan pemahaman data penelitian, dan Saudari Nor Faizah dan Saudara Imam Agus Maulana, sahabat sekaligus rekan seperjuangan yang selalu hadir memberikan semangat, dukungan, doa, dan kebersamaan selama penulis menempuh studi hingga penyelesaian tesis ini.
10. Seluruh rekan mahasiswa Magister Informatika dan semua pihak yang telah membantu serta memberikan dukungan kepada penulis selama masa studi dan penyusunan tesis ini.
11. Seluruh pihak serta orang-orang terdekat penulis yang tidak dapat disebutkan satu per satu, atas segala doa, dukungan, perhatian, dan kebersamaan yang telah diberikan kepada penulis selama proses studi dan penyusunan tesis ini.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi penyempurnaan di masa mendatang. Semoga tesis ini dapat memberikan manfaat bagi pembaca serta menjadi referensi dalam pengembangan penelitian di bidang analisis data dan kajian polusi udara berdasarkan faktor-faktor meteorologi.

Akhir kata, penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan bantuan, bimbingan, dukungan, dan doa selama proses studi serta penyusunan tesis ini. Semoga segala kebaikan yang telah diberikan mendapat balasan dan keberkahan dari Allah *Subhānahu wa Ta'ālā*. Aamiin Ya Rabbal 'Alamin.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Malang, 19 Juni 2026

Penulis,

Maimuna Syarifatussholehah

DAFTAR ISI

HALAMAN PENGAJUAN	i
HALAMAN PERSETUJUAN	iii
MOTTO	vi
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR.....	viii
DAFTAR ISI	x
DAFTAR TABEL	xii
DAFTAR GAMBAR.....	xiii
Abstrak	xiv
Abstract.....	xv
الملخص.....	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Pernyataan Masalah	7
1.3 Tujuan Penelitian	7
1.4 Hipotesis	8
1.5 Manfaat Penelitian	8
1.6 Batasan Masalah	8
1.7 Sistematika Penulisan	9
BAB II STUDI PUSTAKA	11
2.1 <i>Analisis Kluster Polusi Udara</i>	11
2.2 Kerangka Teori	21
2.3 <i>Principal Component Analysis (PCA) Clustering</i>	24
BAB III METODOLOGI PENELITIAN	27
3.1 Kerangka Konsep.....	27

3.2 Desain Sistem	29
3.2 Data Preparation	34
3.3 Skenario Evaluasi.....	36
BAB IV HASIL PENELITIAN.....	39
4.1 Standarisasi Data.....	39
4.2 Hasil <i>Analisis Principal Component Analysis</i> (PCA).....	41
4.2.1 Nilai <i>Eigenvalue</i> dan <i>Explained Variance Ratio</i>	41
4.3 Hasil Pengelompokan Analisis Komponen Utama.....	43
4.4 Penentuan Label dan Evaluasi <i>Silhouette Score</i>	44
4.4.1 Analisis Kualitas Klaster dengan <i>Silhouette Score</i>	45
4.4.2 Penetapan Ambang Batas (<i>Threshold</i>) Pengelompokan	48
4.5 Visualisasi Hasil Pengelompokan Data	49
4.6 Perbandingan Kinerja PCA- <i>Clustering</i> dan K-Means.....	52
4.6.1 Implementasi Metode K-Means	53
4.6.2 Hasil <i>Clustering</i> Menggunakan K-Means.....	54
4.6.3 Perbandingan Kinerja PCA- <i>Clustering</i> dan K-Means	57
BAB V PEMBAHASAN	60
5.1 Analisis Pola Pengelompokan Polusi Udara.....	60
5.2 Pengaruh Kondisi Cuaca terhadap Kejadian Polusi Ekstrem	62
5.3 Ketepatan Garis Batas (<i>Threshold</i>) dalam Pemberian Label.....	64
5.4 Implikasi Hasil Pengelompokan terhadap Kualitas Udara	66
5.5 Perspektif Islam Terhadap Hasil Pengelompokan Polusi Udara	69
BAB VI KESIMPULAN DAN SARAN.....	74
6.1 Kesimpulan	74
6.2 Saran	75
DAFTAR PUSTAKA.....	76

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu.....	18
Tabel 3.1 Keterangan Variabel Dataset.....	35
Tabel 3.2 Data Partikulat dan Meteorologi	36
Tabel 4.1 Hasil Standarisasi Variabel Menggunakan <i>StandardScaler</i>	40
Tabel 4.2 Nilai <i>Eigenvalue</i> dan Rasio Variansi Komponen Utama	41
Tabel 4.3 Hasil Evaluasi Jumlah Klaster Berdasarkan <i>Silhouette Score</i>	46
Tabel 4.8 Distribusi Data Berdasarkan Klaster	51
Tabel 4.9 Distribusi Jumlah Data Hasil <i>Clustering</i> K-Means.....	56
Tabel 4.10 Perbandingan Nilai <i>Silhouette Score</i> PCA- <i>Clustering</i> dan K-Means	57

DAFTAR GAMBAR

Gambar 2.1 Kerangka Teori.....	22
Gambar 3.1 Kerangka Konsep	27
Gambar 3.2 Desain Sistem.....	30
Gambar 4.1 Eigenvalues dan Explained Variance Ratio	42
Gambar 4.2 Graph of PC1 and PC2 Variance ratio Value.....	44
Gambar 4.3 Coordinate Points of PC1 and PC2 Data Distribution	50
Gambar 4.4 Hasil Pengelompokan Data Menggunakan K-Means	55

Abstrak

Syariefatussholehah Maimuna 2026. ***Clustering Polusi Udara Berdasarkan Faktor Meteorologi Berbasis Principal Component Analysis***. Tesis. Program Study Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing (I) Dr. Agung Teguh Wibowo Almais, S.kom., M.T Pembimbing (II) Dr. M. Imamudin Lc, MA

Kata kunci: Polusi Udara, PM_{2.5}, *Principal Component Analysis*, *Clustering*, Faktor Meteorologi, Kota Malang.

Polusi udara merupakan salah satu permasalahan lingkungan yang dipengaruhi oleh berbagai faktor meteorologi, seperti suhu udara, kelembapan udara, curah hujan, penyinaran matahari, dan kecepatan angin. Interaksi antara faktor-faktor tersebut dapat memengaruhi variasi konsentrasi PM_{2.5} sehingga diperlukan pendekatan analisis untuk mengidentifikasi pola kondisi kualitas udara. Penelitian ini bertujuan mengidentifikasi pola kondisi polusi udara berdasarkan faktor meteorologi menggunakan pendekatan *PCA-Clustering*. Data yang digunakan berupa data harian PM_{2.5} dan parameter meteorologi yang diperoleh dari BMKG Kota Malang selama tahun 2024. Tahapan penelitian meliputi praproses data menggunakan *StandardScaler*, reduksi dimensi menggunakan *Principal Component Analysis* (PCA), serta pelabelan data berdasarkan distribusi nilai komponen utama pertama (PC1). Hasil PCA menunjukkan bahwa dua komponen utama pertama (PC1 dan PC2) mampu menjelaskan 67.15% variasi total data. Hasil *PCA-Clustering* mengidentifikasi dua kelompok kondisi kualitas udara, yaitu kondisi normal sebanyak 357 data dan kondisi ekstrem sebanyak 9 data. Evaluasi menggunakan *silhouette score* menghasilkan nilai sebesar 0.87 yang menunjukkan bahwa kelompok yang terbentuk memiliki tingkat kekompakan dan keterpisahan yang sangat baik. Hasil perbandingan dengan metode K-Means menunjukkan bahwa *PCA-Clustering* menghasilkan kualitas pengelompokan yang lebih baik dengan nilai *silhouette score* sebesar 0.87 dibandingkan K-Means sebesar 0.47. Hasil penelitian menunjukkan bahwa pendekatan *PCA-Clustering* efektif dalam mengidentifikasi pola kondisi kualitas udara berdasarkan faktor meteorologi dan dapat dimanfaatkan sebagai informasi pendukung dalam kegiatan pemantauan kualitas udara di Kota Malang.

Abstract

Syariefatussholehah Maimuna. 2026. *Air Pollution Clustering Based on Meteorological Factors Using Principal Component Analysis*. Thesis. Master of Informatics Program, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Supervisors: (I) Dr. Agung Teguh Wibowo Almais, S.Kom., M.T.; (II) Dr. M. Imamudin, Lc., M.A.

Air pollution is a significant environmental issue influenced by various meteorological factors, including air temperature, humidity, rainfall, sunshine duration, and wind speed. The interactions among these factors may affect variations in PM_{2.5} concentrations, necessitating an analytical approach to identify air quality condition patterns. This study aims to identify air pollution condition patterns based on meteorological factors using a PCA-Clustering approach. The dataset consisted of daily PM_{2.5} measurements and meteorological parameters obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) in Malang City during 2024. The research procedure involved data preprocessing using StandardScaler, dimensionality reduction through Principal Component Analysis (PCA), and data labeling based on the distribution of the first principal component (PC1) values. The PCA results revealed that the first two principal components (PC1 and PC2) accounted for 67.15% of the total data variance. The PCA-Clustering approach identified two air quality condition groups, namely a normal condition group comprising 357 observations and an extreme condition group comprising 9 observations. Cluster evaluation using the Silhouette Score yielded a value of 0.87, indicating excellent cluster cohesion and separation. Furthermore, a comparison with the K-Means algorithm demonstrated that PCA-Clustering achieved superior clustering performance, with a silhouette score of 0.87 compared to 0.47 for K-Means. The findings indicate that the PCA-Clustering approach is effective in identifying air quality condition patterns based on meteorological factors and can serve as supporting information for air quality monitoring activities in Malang City.

Keywords: Air Pollution, PM_{2.5}, Principal Component Analysis, Clustering, Meteorological Factors, Malang City.

الملخص

شريعة الصالحة ميمونة، 2026. تجمع تلوث الهواء بناء على عوامل الأرصاد الجوية بناء على تحليل المكونات الرئيسية. رسالة جامعية، برنامج ماجستير المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية، مالانغ. المشرف (I) الدكتور أغونغ تيغوه ويبيو أليس، S.Kom.، M.T.، المشرف (II) الدكتور محمد إمام الدين، ماجستير.

الكلمات المفتاحية: تلوث الهواء، $PM_{2.5}$ ، تحليل المكونات الرئيسية، التجمع، العوامل الجوية في مدينة مالانغ.

تلوث الهواء هو أحد المشكلات البيئية التي تتأثر بعوامل جوية مختلفة، مثل درجة حرارة الهواء، والرطوبة، والأمطار، والإشعاع الشمسي، وسرعة الرياح. يمكن أن يؤثر التفاعل بين هذه العوامل على التغيرات في تركيزات $PM_{2.5}$ ، لذا هناك حاجة إلى نهج تحليلي لتحديد أنماط ظروف جودة الهواء. تهدف هذه الدراسة إلى تحديد نمط ظروف تلوث الهواء بناء على عوامل الأرصاد الجوية باستخدام نهج *PCA-Clustering*. البيانات المستخدمة هي بيانات $PM_{2.5}$ اليومية والمعايير الجوية التي تم الحصول عليها من BMKG مدينة مالانغ خلال عام 2024. تشمل مراحل البحث معالجة البيانات مسبقاً باستخدام *StandardScaler*، وتقليل الأبعاد باستخدام تحليل المكونات الرئيسية (PCA)، وتصنيف البيانات بناء على توزيع قيمة المكون الرئيسي الأول (PC1). تظهر نتائج PCA أن المكونين الرئيسيين الأولين (PC1 و PC2) قادران على تفسير 67.15٪ من إجمالي تباين البيانات. حددت نتائج تجميع PCA مجموعتين من ظروف جودة الهواء، وهما الظروف الطبيعية التي تصل إلى 357 بيانه، والظروف القصوى التي تصل إلى 9 بيانات. أظهر التقييم باستخدام درجة *الظل* قيمة 0.87 أظهرت أن المجموعة التي تشكلت تتمتع بمستوى ممتاز من التماسك والفصل. أظهرت نتائج المقارنة مع طريقة K-Means أن تجميع PCA أنتج جودة تجميع أفضل بدرجة *ظلال* 0.87 مقارنة بمتوسط 0.47 K. تظهر نتائج الدراسة أن نهج PCA-التجميع فعال في تحديد أنماط ظروف جودة الهواء بناء على العوامل الجوية ويمكن استخدامه كمعلومات داعمة في أنشطة مراقبة جودة الهواء في مدينة مالانغ.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Polusi udara merupakan salah satu permasalahan lingkungan global yang memberikan dampak serius terhadap kesehatan manusia serta keberlanjutan ekosistem. Peningkatan konsentrasi polutan di atmosfer telah berkembang menjadi krisis lingkungan yang mengancam kualitas hidup masyarakat di berbagai negara. Xu et al., (2025) menjelaskan bahwa pencemaran udara merupakan salah satu tantangan lingkungan yang memiliki implikasi luas terhadap aspek kesehatan, sosial, dan ekonomi. *World Health Organization*, (2021) melaporkan bahwa sekitar 99% populasi dunia terpapar udara dengan konsentrasi polutan yang melebihi batas aman yang direkomendasikan, sehingga menyebabkan sekitar tujuh juta kematian prematur setiap tahunnya akibat penyakit pernapasan dan kardiovaskular (World Health Organization, 2023). Di antara berbagai jenis polutan, *Particulate Matter* berukuran kurang dari 2,5 μm (PM_{2.5}) merupakan salah satu polutan yang paling berbahaya karena partikulat tersebut mampu menembus sistem pernapasan hingga memasuki aliran darah dan meningkatkan risiko berbagai gangguan kesehatan.

Di Indonesia, permasalahan polusi udara semakin mendapat perhatian seiring dengan meningkatnya aktivitas transportasi, pertumbuhan sektor industri, serta perubahan tata guna lahan di wilayah perkotaan. Badan Pusat Statistik Kota Malang, (2023) melaporkan bahwa rata-rata konsentrasi PM_{2.5} di kota-kota besar Indonesia dapat mencapai 45 $\mu\text{g}/\text{m}^3$ atau sekitar tujuh kali lebih tinggi dibandingkan

nilai ambang batas yang direkomendasikan oleh WHO. Tingginya konsentrasi polutan tersebut tidak hanya dipengaruhi oleh sumber emisi yang berasal dari kendaraan bermotor, aktivitas industri, dan pembakaran terbuka, tetapi juga oleh kondisi meteorologi, seperti suhu udara, kecepatan angin, curah hujan, kelembapan relatif, serta penyinaran matahari yang berperan dalam proses pembentukan, dispersi, dan akumulasi polutan di atmosfer. Shafi & Waheed, (2020) mengungkapkan bahwa konsentrasi PM_{2.5} dapat meningkat hingga mencapai 242 µg/m³ hanya dalam waktu tujuh menit setelah terjadinya kebakaran. Sementara itu, Huang et al., (2025) menjelaskan bahwa interaksi antara kanopi vegetasi dan fenomena inversi termal dapat menyebabkan akumulasi PM_{2.5} yang tinggi pada kawasan ruang terbuka hijau di Tongxiang, Tiongkok. Temuan-temuan tersebut menunjukkan bahwa dinamika polusi udara dipengaruhi oleh interaksi yang kompleks antara sumber emisi dan faktor meteorologi.

Dalam perspektif Islam, kerusakan lingkungan yang ditimbulkan oleh aktivitas manusia merupakan konsekuensi dari terganggunya keseimbangan hubungan antara manusia dan alam. Allah SWT berfirman dalam Surah Ar-Rum ayat 41:

ظَهَرَ الْفَسَادُ فِي الْبَرِّ وَالْبَحْرِ بِمَا كَسَبَتْ أَيْدِي النَّاسِ لِيُذِيقَهُمْ بَعْضَ الَّذِي عَمِلُوا لَعَلَّهُمْ يَرْجِعُونَ ﴿٤١﴾

Artinya: “Telah tampak kerusakan di darat dan di laut disebabkan karena perbuatan tangan manusia; Allah menghendaki agar mereka merasakan sebagian dari akibat perbuatan mereka agar mereka kembali ke jalan yang benar” (QS. Ar-Rum: 41).

Ayat tersebut memberikan landasan bahwa fenomena *al-fasad* (kerusakan) yang terjadi di darat dan di laut merupakan konsekuensi dari aktivitas manusia sebagaimana ditegaskan dalam frasa "*bimā kasabat aydin-nās*" yang berarti "disebabkan karena perbuatan tangan manusia". Dalam konteks lingkungan, konsep *al-fasad* dapat dimaknai sebagai berbagai bentuk kerusakan ekosistem yang mengakibatkan terganggunya keseimbangan alam, termasuk pencemaran udara yang terjadi akibat aktivitas antropogenik seperti transportasi, industri, pembakaran terbuka, dan perubahan tata guna lahan. Dengan demikian, ayat tersebut menunjukkan bahwa manusia memiliki tanggung jawab terhadap kondisi lingkungan yang ditimbulkan oleh aktivitasnya, sekaligus berkewajiban untuk melakukan upaya perbaikan dan pelestarian lingkungan.

Oleh karena itu, upaya memahami karakteristik polusi udara serta faktor-faktor yang memengaruhinya tidak hanya memiliki nilai ilmiah, tetapi juga merupakan bagian dari tanggung jawab manusia dalam menjaga keseimbangan lingkungan dan mewujudkan kemaslahatan bagi kehidupan masyarakat. Penelitian mengenai pola pengelompokan polusi udara berdasarkan faktor meteorologi diharapkan dapat memberikan kontribusi dalam mendukung upaya mitigasi pencemaran udara dan pengelolaan lingkungan yang lebih berkelanjutan sesuai dengan nilai-nilai Islam (Indonesia, 2019).

Hubungan antara polusi udara dan parameter meteorologi memiliki karakteristik yang kompleks dan saling berkorelasi sehingga memerlukan pendekatan analisis yang mampu mengidentifikasi pola-pola yang terbentuk di

dalam data. Seiring dengan perkembangan teknologi, berbagai penelitian telah memanfaatkan metode *machine learning* untuk menganalisis dan memprediksi kualitas udara. Roy et al., (2022) menjelaskan bahwa suhu udara dan *planetary boundary height* merupakan faktor yang berpengaruh dominan terhadap variasi konsentrasi PM_{2.5}. Montazeri et al., (2021) mengungkapkan bahwa tata guna lahan dan kecepatan angin memiliki peran penting dalam distribusi konsentrasi NO₂ di wilayah perkotaan. Liu et al., (2022) berhasil mengintegrasikan *Principal Component Analysis* (PCA) dengan *Support Vector Regression* (SVR) untuk meningkatkan akurasi prediksi PM_{2.5} di Beijing. Kalajdjieski et al., (2023) mengembangkan model *Long Short-Term Memory* (LSTM) berbasis *attention mechanism* untuk pemantauan kualitas udara secara real-time di Skopje, Makedonia. Sementara itu, Zareba et al., (2023) memanfaatkan K-Means *Clustering* dengan metrik *Dynamic Time Warping* (DTW) untuk mengidentifikasi pola temporal polusi PM₁₀ di Krakow, Polandia. Berbagai penelitian tersebut menunjukkan bahwa metode machine learning memiliki kemampuan yang baik dalam menganalisis kualitas udara.

Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada aspek prediksi konsentrasi polutan, sedangkan kajian yang berorientasi pada identifikasi pola kondisi polusi udara berdasarkan interaksi parameter meteorologi masih relatif terbatas. *Principal Component Analysis* (PCA) pada umumnya dimanfaatkan sebagai teknik reduksi dimensi untuk mendukung model prediksi, sedangkan penerapan integrasi PCA dan *Clustering* sebagai metode utama untuk mengidentifikasi pola kondisi polusi udara masih relatif terbatas. Kondisi tersebut

menunjukkan bahwa pemanfaatan *PCA-Clustering* sebagai pendekatan utama dalam mengungkap karakteristik kondisi polusi udara berdasarkan faktor meteorologi masih memerlukan pengembangan lebih lanjut.

Selain itu, sebagian besar penelitian terdahulu dilakukan pada wilayah beriklim subtropis, seperti Beijing, Cincinnati, Oakland, Krakow, dan Skopje, yang memiliki karakteristik suhu, kelembapan, curah hujan, serta dinamika atmosfer yang berbeda dengan wilayah tropis. Perbedaan karakteristik tersebut menyebabkan hasil penelitian sebelumnya belum tentu mampu merepresentasikan dinamika polusi udara pada wilayah tropis secara optimal. Oleh karena itu, kajian mengenai pola kondisi polusi udara berbasis faktor meteorologi pada wilayah beriklim tropis masih memerlukan pengembangan lebih lanjut.

Di Indonesia, penelitian mengenai kualitas udara masih didominasi oleh pendekatan statistik konvensional maupun model prediksi berbasis machine learning. Kajian yang mengintegrasikan *PCA-Clustering* untuk mengidentifikasi pola kondisi polusi udara berdasarkan parameter meteorologi pada wilayah perkotaan masih relatif terbatas. Hingga saat ini, penelitian yang secara khusus menerapkan pendekatan *PCA-Clustering* untuk menganalisis pola pengelompokan kondisi polusi udara berdasarkan faktor meteorologi di Kota Malang masih sangat terbatas. Kondisi tersebut menunjukkan masih terbatasnya bukti empiris mengenai karakteristik pengelompokan kondisi polusi udara berdasarkan faktor meteorologi pada wilayah perkotaan beriklim tropis, khususnya Kota Malang.

Kota Malang merupakan salah satu wilayah perkotaan yang mengalami pertumbuhan jumlah kendaraan bermotor sekitar 9% per tahun dan memiliki karakteristik topografi cekung yang berpotensi menyebabkan akumulasi polutan pada lapisan atmosfer bawah. Nor Asri et al., (2022) menjelaskan bahwa konsentrasi $PM_{2.5}$ di kawasan padat aktivitas, seperti Jalan Basuki Rahmat, dapat mencapai $68 \mu g/m^3$ pada jam sibuk atau sekitar dua kali lebih tinggi dibandingkan ambang batas nasional. Kondisi tersebut menunjukkan bahwa kualitas udara di Kota Malang dipengaruhi oleh interaksi kompleks antara sumber emisi dan faktor meteorologi. Oleh karena itu, diperlukan suatu pendekatan yang mampu menyederhanakan data multidimensi sekaligus mengidentifikasi pola-pola kondisi polusi udara yang terbentuk akibat interaksi antarparameter tersebut.

Berdasarkan permasalahan dan kesenjangan penelitian tersebut, penelitian ini menerapkan metode *PCA-Clustering* untuk mengelompokkan kondisi polusi udara berdasarkan parameter meteorologi yang meliputi suhu udara, kelembapan relatif, curah hujan, kecepatan angin, dan penyinaran matahari. Berbeda dengan penelitian sebelumnya yang lebih berorientasi pada prediksi konsentrasi polutan (Liu et al., 2022) penelitian ini menitikberatkan pada identifikasi kelompok-kelompok kondisi polusi udara yang terbentuk berdasarkan karakteristik meteorologi. Kebaruan penelitian ini terletak pada penerapan integrasi *PCA-Clustering* sebagai metode utama untuk mengidentifikasi pola kondisi polusi udara berbasis faktor meteorologi pada wilayah perkotaan beriklim tropis, khususnya Kota Malang. Melalui pendekatan tersebut, penelitian ini diharapkan dapat memberikan bukti empiris mengenai karakteristik kondisi polusi udara yang

terbentuk berdasarkan faktor meteorologi serta memperkaya penerapan metode *PCA-Clustering* dalam bidang analisis kualitas udara. Hasil penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan metode analisis kualitas udara berbasis machine learning, serta mendukung pengembangan sistem peringatan dini, mitigasi risiko kesehatan masyarakat, dan penyusunan kebijakan pengendalian kualitas udara yang lebih efektif dan berkelanjutan sesuai dengan tujuan (SDGs) poin ke-11 dan ke-13 (Pusat, 2023; The Global Goals, 2024).

1.2 Pernyataan Masalah

1. Bagaimana faktor meteorologi memengaruhi pola variasi polusi udara di wilayah urban?
2. Seberapa efektif metode *PCA-Clustering* dalam mengidentifikasi kluster polusi udara berdasarkan interaksi kompleks antara polusi dan parameter cuaca?
3. Apa implikasi hasil segmentasi polusi udara terhadap kebijakan pengendalian kualitas udara dan perlindungan kesehatan masyarakat?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Menganalisis pengaruh faktor meteorologi (seperti kecepatan angin, curah hujan, dan suhu udara) terhadap variasi polusi udara di Kota Malang.
2. Mengembangkan model *PCA-Clustering* untuk mengelompokkan wilayah berdasarkan pola polusi udara dan dinamika cuaca.

3. Memberikan rekomendasi strategis berbasis data kepada pemangku kebijakan dalam merancang program mitigasi polusi udara yang efektif dan berkelanjutan.

1.4 Hipotesis

1. H1: Penerapan metode *PCA-Clustering* dapat mengelompokkan data polusi udara berdasarkan karakteristik faktor meteorologi ke dalam kelompok-kelompok yang memiliki pola serupa.
2. H2: *PCA-Clustering* memiliki tingkat performa yang baik dalam mengelompokkan kondisi polusi udara, yang ditunjukkan oleh nilai *Silhouette Score* positif (> 0.5).

1.5 Manfaat Penelitian

1. Memperkaya literatur tentang integrasi machine learning dalam analisis lingkungan, khususnya penerapan *PCA-Clustering* untuk studi polusi udara di wilayah tropis.
2. Menyediakan peta kluster polusi udara sebagai alat bantu dalam perencanaan tata ruang, pengembangan sistem pemantauan real-time, dan kebijakan pengurangan emisi.
3. Meningkatkan kesadaran masyarakat tentang risiko kesehatan polusi udara dan mendorong partisipasi aktif dalam program lingkungan berbasis bukti.

1.6 Batasan Masalah

Agar penelitian tetap fokus dan terukur, batasan masalah yang diterapkan adalah:

1. Data polusi udara dan meteorologi yang digunakan bersifat historis (periode 2024) dari stasiun pemantauan tetap di Kota Malang sehingga tidak mencakup variasi real-time.

2. Fokus analisis pada Materi Partikulat ($PM_{2.5}$) serta lima parameter meteorologi utama: kecepatan angin, curah hujan, penyinaran matahari, suhu udara, dan kelembaban relatif.
3. Model PCA-*Clustering* mengasumsikan linearitas hubungan antar variabel, sehingga belum mengakomodasi pola non-linear yang mungkin muncul akibat fenomena cuaca ekstrem.

1.7 Sistematika Penulisan

BAB I PENDAHULUAN

Menjelaskan latar belakang pentingnya analisis polusi udara berdasarkan faktor meteorologi di Kota Malang, pernyataan masalah, rumusan masalah, tujuan, manfaat, hipotesis, serta batasan penelitian.

BAB II STUDI PUSTAKA

Menguraikan teori tentang *Clustering* polusi udara, metode PCA-*Clustering*, penelitian terdahulu, serta kerangka teori sebagai dasar penelitian

BAB III METODOLOGI PENELITIAN

Membahas desain sistem, sumber data BMKG 2024, tahap pembersihan dan standarisasi data, penerapan PCA, *Clustering*, serta metode evaluasi (*Silhouette Score* dan validasi ahli).

BAB IV HASIL PENELITIAN

Menyajikan hasil *preprocessing* dan standarisasi data, hasil reduksi dimensi menggunakan PCA, pembentukan komponen utama, hasil *Clustering*, serta hasil evaluasi menggunakan *Silhouette Score*.

BAB V PEMBAHASAN

Membahas hasil pengelompokan kondisi polusi udara berdasarkan faktor meteorologi, interpretasi setiap *cluster* yang terbentuk, serta perbandingan hasil penelitian dengan penelitian terdahulu.

BAB VI KESIMPULAN DAN SARAN

Merangkum hasil penelitian mengenai penerapan *PCA-Clustering* dalam mengelompokkan kondisi polusi udara berdasarkan faktor meteorologi serta memberikan saran untuk penelitian selanjutnya.

BAB II

STUDI PUSTAKA

2.1 Analisis Kluster Polusi Udara

Penelitian oleh Shafi & Waheed, (2020) berfokus dalam menganalisis dinamika polusi udara akibat kebakaran di Southampton, Inggris, melalui pendekatan machine learning. Penelitian ini menekankan urgensi pemantauan kualitas udara menyusul dampak kesehatan yang masif mulai dari iritasi mata hingga kematian dini 24.000 orang per tahun serta ancaman lingkungan seperti hujan asam. Kasus kebakaran pada 2 Desember 2019 menjadi fokus utama, di mana kadar PM_{2.5} meloncat dari 32 $\mu\text{g}/\text{m}^3$ (kategori baik) menjadi 242 $\mu\text{g}/\text{m}^3$ (berbahaya) hanya dalam 7 menit, memicu risiko akut bagi kesehatan masyarakat. Dengan memanfaatkan algoritma *K-Means Clustering* dan *Metode Elbow*, data PM_{2.5} dari sensor IoT (Plantower PMS5003 dan Sensirion SPS030) dikelompokkan ke dalam lima kategori kualitas udara (baik hingga berbahaya) serta tiga kluster temporal (sebelum, selama, pasca-kebakaran). Hasilnya mengungkap fluktuasi ekstrem PM_{2.5} yang bersifat temporer namun berdampak signifikan, dengan pemulihan kualitas udara membutuhkan waktu beberapa jam. Penelitian ini menegaskan potensi integrasi sistem IoT dan analisis kluster untuk pengembangan early warning system di wilayah urban rawan kebakaran.

Penelitian oleh Wang et al. (2021) berfokus pada prediksi kualitas air waduk menggunakan pendekatan gabungan antara algoritma klusterisasi dinamis yang dimodifikasi dan jaringan saraf dinamis regresi (NARX). Penelitian ini

menekankan pentingnya prediksi kualitas air secara akurat dan cepat untuk mendukung pengambilan keputusan dalam pengelolaan sumber daya air serta memberikan peringatan dini terhadap degradasi kualitas air akibat faktor cuaca. Kasus Waduk Shi River di China menjadi fokus utama, di mana konsentrasi total fosfor (TP) dan total nitrogen (TN) selalu tinggi dan menjadi indikator utama eutrofikasi air. Dengan memanfaatkan model CA-NARX yang menggabungkan algoritma klasterisasi berbasis koefisien siluet dan jaringan saraf NARX dengan validasi silang m-fold, data faktor cuaca (suhu udara, suhu permukaan air, hujan, dan evaporasi) serta data historis TP dan TN dianalisis untuk memprediksi perubahan kualitas air. Hasilnya menunjukkan bahwa model CA-NARX mampu memprediksi konsentrasi TP dan TN dengan akurasi tinggi (kesalahan relatif rata-rata $\sim 1\%$), unggul dalam prediksi jangka pendek (15 hari) dan unggul secara signifikan dalam prediksi jangka menengah (30 hari) dibandingkan metode SVM, GA-BP, dan pemulusan eksponensial. Penelitian ini menegaskan potensi model CA-NARX sebagai solusi praktis dan hemat biaya untuk prediksi kualitas air di waduk kecil-menengah dengan keterbatasan data dan dana, serta dapat diadaptasi untuk danau atau sungai kecil lainnya.

Penelitian oleh Montazeri et al. (2021) berfokus untuk mengeksplorasi dinamika polusi *Nitrogen Dioksida* (NO_2) di Kota Oakland, California, dengan fokus pada interaksi kompleks antara penggunaan lahan, kondisi meteorologi, dan kesehatan masyarakat. Penelitian ini menjawab keterbatasan metode pemantauan udara konvensional yang cenderung statis dan kurang mampu merekam variasi spasial polusi dengan mengintegrasikan data pengukuran mobile dari kendaraan

Google Street View (2.7 juta sampel), analisis statistik multivariat, dan algoritma *K-Means Clustering*. Hasilnya mengungkapkan bahwa penggunaan lahan berperan krusial dalam memediasi pengaruh kecepatan angin terhadap konsentrasi NO₂, terutama di klaster industri dan permukiman. Pada musim dingin, peningkatan kecepatan angin sebesar 1 m/s mengurangi kadar NO₂ sekitar 3 ppb di kedua wilayah tersebut, sementara di jalan raya dengan lalu lintas truk intensif, efek serupa tidak teramati akibat turbulensi udara dari aktivitas kendaraan. Lebih memprihatinkan, probabilitas paparan NO₂ melebihi ambang batas kesehatan (40 ppb) di koridor transportasi ini mencapai 53-84%, mengindikasikan risiko akut bagi populasi urban. Melalui pendekatan *Principal Component Analysis* (PCA) dan pemodelan spasial 30-meter, penelitian ini tidak hanya menyoroti kerentanan wilayah tertentu terhadap polusi udara tetapi juga menawarkan kerangka kerja adaptif untuk perencanaan kota berkelanjutan (Montazeri et al., 2021).

Penelitian oleh Rivera-Campoverde et al., (2021) berfokus dalam eksplorasi pendekatan inovatif untuk mengatasi keterbatasan metode konvensional dalam mengestimasi emisi kendaraan. Berangkat dari fakta bahwa 54,5% populasi global tinggal di wilayah perkotaan (PBB, 2016) dengan mesin pembakaran dalam (*Internal Combustion Engine*) sebagai kontributor utama polusi udara, penelitian ini mengkritisi ketidakakuratan pengukuran emisi laboratorium yang gagal merepresentasikan kondisi riil akibat variasi gaya berkendara, kepadatan lalu lintas, dan faktor lingkungan. Melalui integrasi data *Onboard Diagnostics* (OBD) dan GPS dengan teknik machine learning, tim peneliti mengembangkan model *Artificial Neural Network* (ANN) yang mampu memprediksi empat polusi kunci

(CO₂, CO, HC, NO_x) menggunakan parameter operasional kendaraan seperti *Throttle Position* (TPS), RPM, dan kecepatan. Validasi menggunakan data *Real Driving Emissions* (RDE) pada dua rute berbeda (total 117,9 km) menunjukkan performa mengesankan dengan koefisien determinasi (R^2) 0,985 untuk CO₂ dan 0,999 untuk HC, sekaligus mengungkap temuan kritis seperti kontribusi emisi idle sebesar 7,35% CO₂. Meskipun faktor emisi model OBDM (111,47 g/km CO₂) lebih tinggi dari hasil RDE, metodologi ini menawarkan solusi ekonomis pengganti *Portable Emissions Measurement System* (PEMS) yang mahal, dengan implikasi strategis untuk homologasi kendaraan, kebijakan pengurangan emisi perkotaan, dan pengembangan sistem IoT berbasis real-time monitoring. Implementasi teknik *K-Means Clustering* dan *Classification Tree* untuk identifikasi posisi gigi transmisi (akurasi 99,5%) serta analisis 1.218,19 km data berkendara di Cuenca, Ekuador, memperkuat validitas model dalam kondisi lalu lintas heterogen.

Penelitian yang dilakukan oleh Roy et al., (2022) mengupas dinamika hubungan antara PM_{2.5} dan parameter meteorologi di kawasan Cincinnati, Ohio. Penelitian ini menyoroti permasalahan polusi udara yang terkait dengan PM_{2.5} dan dampaknya pada kesehatan serta lingkungan, seperti penyakit kardiovaskular, asma, bronkitis kronis, hingga kematian dini. Penelitian ini berkontribusi dalam memberikan pemahaman yang lebih mendalam tentang sensitivitas PM_{2.5} terhadap perubahan parameter meteorologi, serta mengevaluasi kesesuaian distribusi gamma untuk memodelkan PM_{2.5}. Dengan memanfaatkan teknik *k-means clustering* dan metode PCA serta regresi multivariabel, penelitian ini menganalisis data PM_{2.5} dan meteorologi selama lima tahun (Agustus 2011-Desember 2015). Hasilnya

menunjukkan bahwa outdoor temperature (OT), *planetary boundary height* (HPBL), dan visibilitas (VIS) memiliki pengaruh terkuat terhadap PM_{2.5}. Penelitian ini juga menemukan bahwa PM_{2.5} memiliki variasi spasial dan temporal di sepanjang kawasan Cincinnati, serta distribusi PM_{2.5} sedikit lebih baik cocok dengan distribusi gamma daripada log-normal. Penelitian ini menegaskan pentingnya analisis spasial dan temporal dalam memahami hubungan antara PM_{2.5} dan meteorologi, serta memberikan fondasi untuk pengembangan model peramalan PM_{2.5} di masa depan.

Penelitian oleh Liu et al., (2022) berfokus pada peramalan konsentrasi PM_{2.5} di Beijing menggunakan metode pengenalan pola polusi. Penelitian ini menyoroti masalah polusi udara yang semakin parah di Beijing dan dampaknya terhadap kesehatan masyarakat, seperti peningkatan penyakit pernapasan dan kardiovaskular serta kematian dini. Dengan menggunakan data dari stasiun pemantauan Yizhuang, penelitian ini menganalisis hubungan antara polusi (PM₁₀, SO₂, NO₂, O₃, CO) dan faktor meteorologi (tekanan udara, kelembaban relatif, suhu udara, kecepatan angin, arah angin, curah hujan kumulatif) terhadap pembentukan dan penyebaran PM_{2.5}. Melalui pendekatan machine learning, penelitian ini menggabungkan metode PCA-Clustering method dan algoritma CS yang dioptimalkan untuk SVR. Data PM_{2.5}, yang mencapai lebih dari 500 µg/m³ pada hari-hari dingin, diproses menggunakan teknik ini untuk menghasilkan model peramalan yang lebih akurat. Hasilnya menunjukkan bahwa model CSSVR memberikan hasil peramalan yang lebih baik dibandingkan dengan model MNR dan ANN. Evaluasi menggunakan indikator RMSE, MAE, MAPE, dan R mengkonfirmasi bahwa ketika data polusi

dibagi menjadi tiga pola, model CSSVR memberikan akurasi peramalan terbaik. Penelitian ini menegaskan potensi metode pengenalan pola polusi dan optimasi parameter SVR untuk pengembangan sistem peramalan polusi udara yang lebih efisien dan akurat di wilayah perkotaan.

Penelitian oleh Kalajdjieski et al., (2023) berfokus dalam menyoroti kompleksitas prediksi PM_{2.5} di wilayah urban seperti Skopje, Makedonia Utara kota dengan tingkat polusi udara tertinggi di Eropa di tengah tantangan missing data 15–20% per tahun dan ketergantungan model pada variabel lokasi serta musim. Dengan memanfaatkan 865.453 data pengukuran per jam dari 40 stasiun pemantauan (2017–2020), penelitian ini mengkombinasikan empat arsitektur encoder-decoder berbasis LSTM dengan attention mechanism untuk mengekstraksi pola temporal-spasial tersembunyi, serta dua model adversarial networks (ANEncoded dan AAN) yang menghasilkan data sintesis realistis berdasarkan kondisi cuaca, koordinat geografis, dan informasi temporal. Hasilnya, *Bidirectional Stacked Attention Model* (BSAM) berhasil memprediksi PM_{2.5} dengan akurasi tertinggi (RMSE 1.98% pada pelatihan dan 3.41% pada pengujian), sementara augmentasi data menggunakan AAN yang menggabungkan encoder-decoder dengan attention mengurangi RMSE hingga 50% (menjadi 1.27%) dengan mempertahankan distribusi data asli (uji Kolmogorov-Smirnov, p-value 0.815). Tidak hanya mengatasi masalah missing data, pendekatan ini juga meningkatkan generalisasi model lintas lokasi berkat kemampuan sintesis data yang adaptif terhadap variasi cuaca dan geografi. Implikasi praktisnya terlihat dalam potensi implementasi sistem pemantauan real-time di kota-kota padat migran, sekaligus membuka jalan untuk aplikasi serupa

pada prediksi polusi udara lain (seperti NO_2 atau SO_2) atau masalah deret waktu multidimensi di sektor kesehatan lingkungan.

Penelitian oleh Zareba et al., (2023) berfokus dalam pendekatan inovatif untuk mengatasi kompleksitas analisis polusi udara partikulat (PM_{10} dan $\text{PM}_{2.5}$) di wilayah perkotaan, dengan studi kasus di Krakow, Polandia. Permasalahan utama yang dihadapi meliputi keterbatasan metode konvensional seperti analisis statistik sederhana dan GIS tradisional dalam mengungkap pola spasio-temporal tersembunyi, serta mahalnya biaya pemantauan real-time menggunakan sensor referensi. Melalui integrasi *Machine Learning* (ML) berbasis *big data* dan penggunaan 100 sensor berbiaya rendah dari Airly (Maret 2021–Februari 2022), penelitian ini berhasil mengembangkan pipeline analisis yang mencakup imputasi data hilang, normalisasi, dan interpolasi spasial menggunakan kriging. Dua algoritma *K-means Clustering* dengan metrik *Dynamic Time Warping* (DTW) untuk variasi temporal dan SKATER untuk keterkaitan spasial dikombinasikan dengan analisis hot-spots (Getis-Ord dan Moran's I) untuk mengidentifikasi pola unik polusi PM_{10} . Hasil analisis mengungkapkan fenomena "*Krakow bagel*", di mana konsentrasi PM_{10} tertinggi justru terdistribusi di wilayah non-perkotaan (khususnya lembah sungai Wisła) akibat akumulasi polusi musim dingin dari pemanasan berbasis batubara, sementara area perkotaan menunjukkan klaster polusi lebih rendah. *K-means Clustering* terbukti superior dalam menangkap variasi temporal cepat (Silhouette Coefficient: 0.62), sedangkan SKATER kurang efektif akibat dominasi faktor spasial. Temuan kritis lain menunjukkan korelasi antara peningkatan polusi musim panas dengan aktivitas lalu lintas dan efek urban heat

island. Melalui interpolasi kriging, metode ini berhasil memitigasi outlier dan meningkatkan akurasi pemetaan polusi hingga 34% dibanding pendekatan tradisional.

Penelitian oleh Huang et al., (2025) mengungkap paradoks tak terduga: kawasan hijau perkotaan justru menjadi episentrum polusi PM_{2.5} tertinggi di Tongxiang, Tiongkok. Studi ini menggabungkan teknologi LiDAR portabel dan stasiun meteorologi canggih untuk menganalisis 1,2 juta data PM_{2.5} per menit di tiga zona fungsional (taman, lalu lintas, permukiman). Hasilnya menunjukkan bahwa taman kota yang diharapkan sebagai paru-paru urban mencatat konsentrasi PM_{2.5} tertinggi (84 µg/m³ pada musim dingin), 23% lebih tinggi daripada zona lalu lintas. Fenomena ini dijelaskan melalui interaksi kompleks antara struktur kanopi pohon dan inversi termal musim dingin, di mana kanopi lebat (CA > 150 m²) justru menghambat dispersi polusi. Analisis regresi mengungkap bahwa peningkatan 1 mm curah hujan mengurangi PM_{2.5} sebesar 13,7 µg/m³, sementara perluasan 1 m² area kanopi meningkatkan PM_{2.5} sebesar 0,86 µg/m³ mengesampingkan asumsi umum bahwa ruang hijau selalu bersifat mitigatif.

Tabel 2.1 Penelitian Terdahulu

Referensi	Topik	Metode	Subjek / Kontribusi
Jana Shafi et al., 2020	Analisis Perubahan Kualitas Udara Menggunakan K-Means Clustering	K-Means Clustering	Polusi udara di lingkungan perkotaan

Referensi	Topik	Metode	Subjek / Kontribusi
Jin Wang et al., 2020	Integrasi Klaster Polusi dan Data Meteorologi	K-Means <i>Clustering</i>	Analisis temporal sumber PM10 & PM2.5 dengan pengaruh cuaca di Tiongkok
Amir Montazeri., 2021	Peran Tata Guna Lahan dalam Kaitan Meteorologi dan Kualitas Udara	Land Use Regression (LUR)	Hubungan antara parameter udara, meteorologi, dan tata guna lahan
Néstor Diego Rivera-Campoverde et al., 2021	Estimasi Emisi Polusi Berbasis Data OBD dan Pembelajaran Mesin	Random Forest, Support Vector Machine (SVM), Neural Networks	Emisi kendaraan bermotor
Madhumitaa Roy et al., 2022	Klasterisasi dan Regresi Polusi PM _{2.5} berdasarkan Kondisi Cuaca	K-Means <i>Clustering</i> , Regresi Linear	Klasterisasi hari berdasarkan faktor meteorologi dan analisis sensitivitas PM _{2.5} di Cincinnati, Ohio
Wei Liu et al., 2022	Prediksi Konsentrasi PM _{2.5} Berbasis Pola Polusi Udara dan Faktor Meteorologi	PCA- <i>Clustering</i> , Support Vector Regression (SVR)	Prediksi PM2.5 berdasarkan pola polusi udara dan parameter meteorologi
Agung Teguh Wibowo Almais, et al., 2023	Pengelompokan Data Kerusakan Sektor Pasca-Bencana Alam	PCA- <i>Clustering</i>	Kerusakan Sektor Pasca-Bencana Alam
Jovan Kalajdjieski et al., 2023	Kerangka Lengkap Pemantauan dan Prediksi Polusi Udara	LSTM	Konsentrasi PM _{2.5} dan PM ₁₀

Referensi	Topik	Metode	Subjek / Kontribusi
Mateusz Zareba et al., 2023	Analisis Pola Polusi Udara Berbasis Big Data dan Machine Learning	XGBoost, Random Forest	Variasi konsentrasi polusi udara di area urban
Yichen Huang et al., 2025	Dampak Faktor Meteorologi dan Kanopi Urban pada PM _{2.5}	Analisis Regresi Multivariat dan Pemodelan Spasial	Interaksi kanopi vegetasi, meteorologi, dan PM _{2.5}
Penelitian ini	Segmentasi Polusi Udara di Kota Malang, Indonesia	PCA- <i>Clustering</i>	Mengelompokkan polusi udara berdasarkan data meteorologi

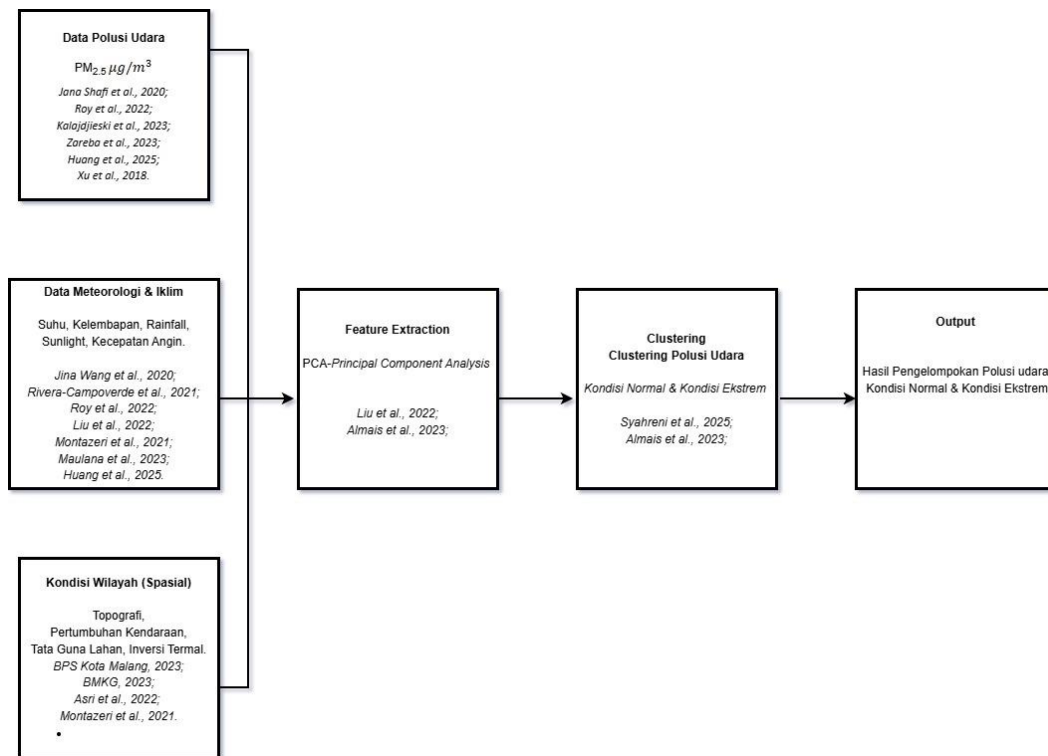
Identifikasi pola spasial-temporal polusi udara melalui pendekatan clustering telah menunjukkan kemajuan signifikan, terutama dengan pemanfaatan teknik reduksi dimensi seperti PCA. Sejumlah studi, seperti Liu et al., (2022) berhasil memadukan PCA dengan Support Vector Regression (SVR) untuk memprediksi konsentrasi PM_{2.5} di wilayah industri, sementara Almais et al., (2023) mengaplikasikan PCA-*Clustering* guna mengklasifikasikan kerusakan pasca-bencana berbasis data multidimensi. Namun demikian, implementasi dalam konteks polusi udara masih bersifat konvensional, di mana komponen utama ditentukan secara statis tanpa mempertimbangkan dinamika temporal seperti fluktuasi harian emisi kendaraan atau variasi musim. Di sisi lain, algoritma clustering klasik seperti K-Means (Shafi & Waheed, 2020) dan DBSCAN (Kalajdjieski et al., 2023) mengasumsikan stabilitas distribusi data, sehingga kurang responsif terhadap perubahan mendadak akibat fenomena cuaca ekstrem atau lockdown perkotaan. Padahal, perkembangan terkini dalam bidang machine

learning memperkenalkan metode seperti Robust PCA yang tahan terhadap noise serta *Deep Clustering* yang mampu menangkap pola non-linear (Guo et al., 2017). Untuk itu, inovasi dengan mengimplementasikan *PCA-Clustering* ke dalam konteks polusi udara pada penelitian ini belum banyak diadopsi untuk analisis polusi udara skala kota, padahal potensinya sangat besar untuk menjawab tantangan kompleksitas data dan kebutuhan real-time monitoring.

2.2 Kerangka Teori

Penelitian ini menggunakan pendekatan PCA untuk mereduksi dimensi data polusi udara dan parameter meteorologi, sehingga pola tersembunyi antar variabel dapat diidentifikasi secara efisien. Data masukan berupa parameter seperti $PM_{2.5}$, suhu, kelembapan, curah hujan, penyinaran matahari, dan kecepatan angin diolah untuk menghasilkan komponen utama yang merepresentasikan variasi dominan dalam data. Hasil PCA kemudian digunakan dalam proses klasterisasi guna mengelompokkan kondisi polusi udara berdasarkan karakteristik cuaca tertentu, yang bertujuan untuk mendukung pemahaman spasial-temporal kualitas udara di wilayah urban.

Dari studi literatur yang dilakukan diperoleh kerangka teori seperti terlihat pada Gambar 2.2.



Gambar 2.1 Kerangka Teori

Gambar 2.1 menggambarkan hubungan konseptual antara sumber data (input), metode analisis (proses), dan hasil akhir (output) berdasarkan sintesis dari berbagai penelitian terdahulu yang relevan. Pada bagian input, ditunjukkan bahwa data yang digunakan dalam penelitian-penelitian sebelumnya mencakup variabel meteorologi seperti suhu rata-rata (*temp_avg*), kelembapan (*humidity_avg*), curah hujan (*rainfall*), penyinaran matahari (*sunlight*), dan kecepatan angin (*wind_avg*), serta konsentrasi polutan partikulat (PM_{2.5}). Jenis data tersebut menjadi dasar utama dalam memahami interaksi antara kondisi fisik atmosfer dan akumulasi polutan di suatu wilayah, terutama dalam mengidentifikasi pemicu terjadinya lonjakan polusi udara.

Selanjutnya, pada bagian proses, Gambar 2.1 menampilkan dua kategori utama pendekatan analisis data, yaitu metode tanpa reduksi dimensi dan metode dengan reduksi dimensi. Metode tanpa reduksi berfokus langsung pada pengelompokan data berdasarkan jarak dan kepadatan variabel asli. Pendekatan ini efektif untuk dataset dengan dimensi rendah, namun sering mengalami kendala efisiensi ketika diterapkan pada data meteorologi dan kualitas udara yang bersifat multidimensi, kompleks, dan memiliki korelasi antar variabel yang tinggi. Sebaliknya, metode dengan reduksi dimensi, seperti PCA (*Principal Component Analysis*), digunakan untuk mengekstraksi variabel-variabel utama yang paling berpengaruh terhadap fluktuasi polusi. Dengan cara ini, informasi dari banyak variabel meteorologi dapat disederhanakan menjadi komponen utama tanpa kehilangan informasi penting, sehingga memudahkan proses visualisasi dan pengelompokan pola lingkungan.

Posisi penelitian ini terletak pada pendekatan reduksi menggunakan PCA yang kemudian diintegrasikan dengan proses *Clustering*, membentuk metode PCA-*Clustering*. Pendekatan ini dipilih karena mampu mengoptimalkan hasil analisis dengan cara mengekstraksi komponen utama dari data meteorologi, kemudian memanfaatkannya untuk mengelompokkan pola kondisi kualitas udara "Normal" dan "Ekstrem" secara lebih efisien dan akurat. Penggunaan faktor meteorologi sebagai input utama menjadikan penelitian ini sangat relevan dalam konteks analisis kesehatan lingkungan, sebab variabel seperti kecepatan angin dan kelembapan memiliki hubungan langsung terhadap proses dispersi atau penumpukan polutan di permukaan tanah.

Hasil akhir dari proses ini berupa *Clustering* polusi udara yang menggambarkan distribusi pola kualitas udara berdasarkan pengaruh cuaca. Temuan ini dapat dimanfaatkan untuk mendukung sistem peringatan dini, kebijakan pengelolaan kualitas udara perkotaan, serta pengambilan keputusan berbasis data meteorologi untuk mitigasi dampak polusi bagi kesehatan masyarakat.

2.3 *Principal Component Analysis (PCA) Clustering*

Principal Component Analysis (PCA) adalah teknik statistik yang digunakan untuk mereduksi dimensi data dengan mengidentifikasi komponen utama yang paling signifikan, yang dapat menjelaskan sebagian besar variabilitas dalam data. PCA dimulai dengan melakukan standarisasi data, jika diperlukan, untuk memastikan setiap variabel memiliki skala yang sama. Selanjutnya, dihitung matriks kovarians untuk menggambarkan hubungan antar variabel, kemudian dilakukan dekomposisi *eigen* untuk menemukan *eigenvector* dan *eigenvalue*. Eigen vector yang memiliki eigenvalue terbesar dipilih sebagai komponen utama, yang mencerminkan arah dengan variansi terbesar dalam data. Proses ini memungkinkan kita untuk memproyeksikan data ke ruang komponen utama yang lebih rendah dimensinya, tetapi tetap mempertahankan informasi yang relevan. PCA sangat berguna dalam analisis data dengan dimensi tinggi, memungkinkan pengurangan kompleksitas dan meningkatkan efisiensi dalam pemrosesan data untuk aplikasi seperti machine learning dan visualisasi data. Langkah-langkah PCA yang digunakan sebagai berikut:

Langkah 1: Normalisasi Data.

Proses normalisasi data merupakan salah satu proses untuk membuat data menjadi standar sehingga sesuai dengan standar PCA. Menurut *Susilo dkk*, standar data PCA adalah data yang digunakan harus memiliki derajat atau nilai yang sama dan seimbang untuk setiap datanya. Dalam penelitian (Almais et al., 2023) terdapat 6 metode untuk melakukan standarisasi data, yaitu Normalisasi (NR), Skala Standar (SS), MinMax (MM), MaxAbs (MA), Robust Scale (RS), dan Quantile Transformer (QT). Penelitian ini menggunakan skala Standar (SS) karena jenis data kerusakan bangunan pasca bencana alam menggunakan skala standar 1/2/3. Persamaan dari skala Standar (SS) menggunakan persamaan (1)

$$X_{standart} = \bar{X}_{mean}(x) / standartdeviation(x) \quad (1)$$

Standar deviasi menggunakan persamaan

$$\bar{X} = \sum_{i=1}^n X_i / n \quad (2)$$

Simbol $\bar{X}$ (batang X) menggambarkan nilai rata-rata dari himpunan X.

Langkah 2: Tentukan rasio varians dan nilai eigen.

Rasio varians adalah ukuran penyebaran data. Persamaan untuk menentukan rasio varians dapat menggunakan persamaan

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1) \quad (3)$$

Persamaan memiliki pengertian bahwa sebaran data memiliki ukuran tertentu yang dapat menentukan jumlah sebaran data. Sedangkan nilai eigen adalah nilai yang menempati suatu tempat dalam vektor eigen dalam bentuk matriks.

Langkah 3: Menentukan jumlah PC.

Dalam penelitian (Almais et al., 2023) dijelaskan bahwa penentuan jumlah komponen utama dalam PCA merupakan proses yang sangat penting karena dapat mempengaruhi tingkat keakuratan suatu data set. Selain itu, penentuan jumlah komponen utama yang tepat akan mendapatkan nilai eigen dan rasio varians yang paling optimal.

Langkah 4: Membuat Visualisasi

Visualisasi merupakan hal yang sangat penting dalam mempresentasikan sebuah hasil, terutama pada hasil PCA. Dengan menggunakan grafik 3 dimensi (3D) untuk mempresentasikan rentang nilai yang hasilnya menggambarkan hasil dalam gambar 2 dimensi (2D).

Langkah 5: Membuat rentang nilai.

Menurut (Almais et al., 2023) rentang nilai dapat dilakukan dengan memaksimalkan nilai tertinggi dalam kumpulan data dan meminimalkan hilangnya informasi yang disebabkan oleh reduksi data. Dalam penelitian ini, untuk membuat rentang nilai menggunakan hasil dari PC.

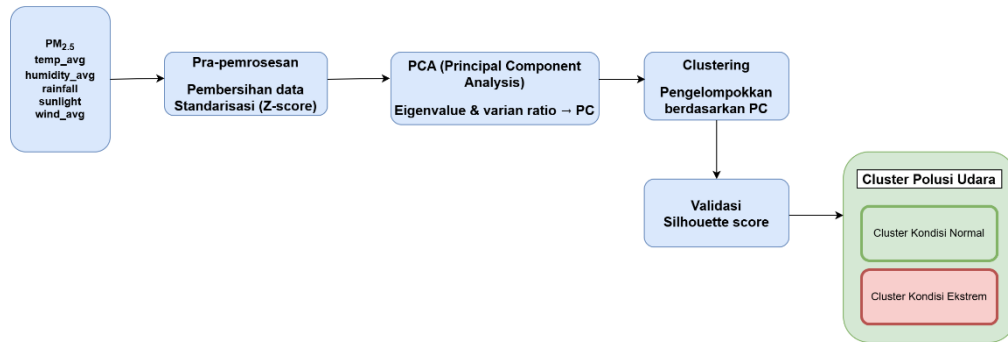
Langkah 7: Mengelompokkan data.

Clustering adalah teknik pembelajaran yang membagi data menjadi beberapa bagian unsupervised data menjadi beberapa kelompok data yang homogen. Pada penelitian (Almais et al., 2023) untuk data clustering menggunakan data titik koordinat PC berdasarkan rentang nilai yang telah ditentukan. Hasil dari clustering data memerlukan validasi untuk mengetahui tingkat kebenarannya.

BAB III METODOLOGI PENELITIAN

3.1 Kerangka Konsep

Kerangka konsep penelitian ini dirancang untuk menguraikan keterkaitan antara parameter meteorologi dan polusi udara, metode analisis yang diterapkan, serta hasil pengelompokan yang diharapkan dalam penelitian, sebagaimana telah diilustrasikan dalam Gambar 3.1.



Gambar 3.1 Kerangka Konsep

Gambar 3.1 menjelaskan secara terperinci alur analisis pola polusi udara menggunakan metode PCA–Clustering, yang terdiri dari tiga tahapan utama, yaitu Input, Proses, dan Output. Setiap tahap memiliki peran penting dalam mendukung proses analisis hingga menghasilkan pengelompokan kondisi kualitas udara "normal" dan "ekstrem" yang representatif terhadap kondisi lingkungan di wilayah penelitian.

Pada tahap Input, data yang digunakan mencakup konsentrasi polutan *Particulate Matter* (PM_{2.5}) serta beberapa variabel meteorologi utama, yaitu suhu rata-rata (*temp_avg*), kelembapan rata-rata (*humidity_avg*), curah hujan (*rainfall*),

penyinaran matahari (*sunlight*), dan kecepatan angin (*wind_avg*). Variabel-variabel tersebut diperoleh dari stasiun pemantauan yang mewakili karakteristik atmosfer harian di wilayah penelitian. Sebelum dilakukan analisis, data melalui tahap pra-pemrosesan yang meliputi pembersihan data untuk mengatasi nilai hilang (*missing values*) serta standardisasi agar seluruh variabel berada pada skala yang sebanding. Proses standardisasi penting dilakukan karena setiap variabel memiliki satuan dan rentang nilai yang berbeda, sehingga diperlukan penyamaan skala agar tidak terjadi bias dalam proses reduksi dimensi dan pengelompokan.

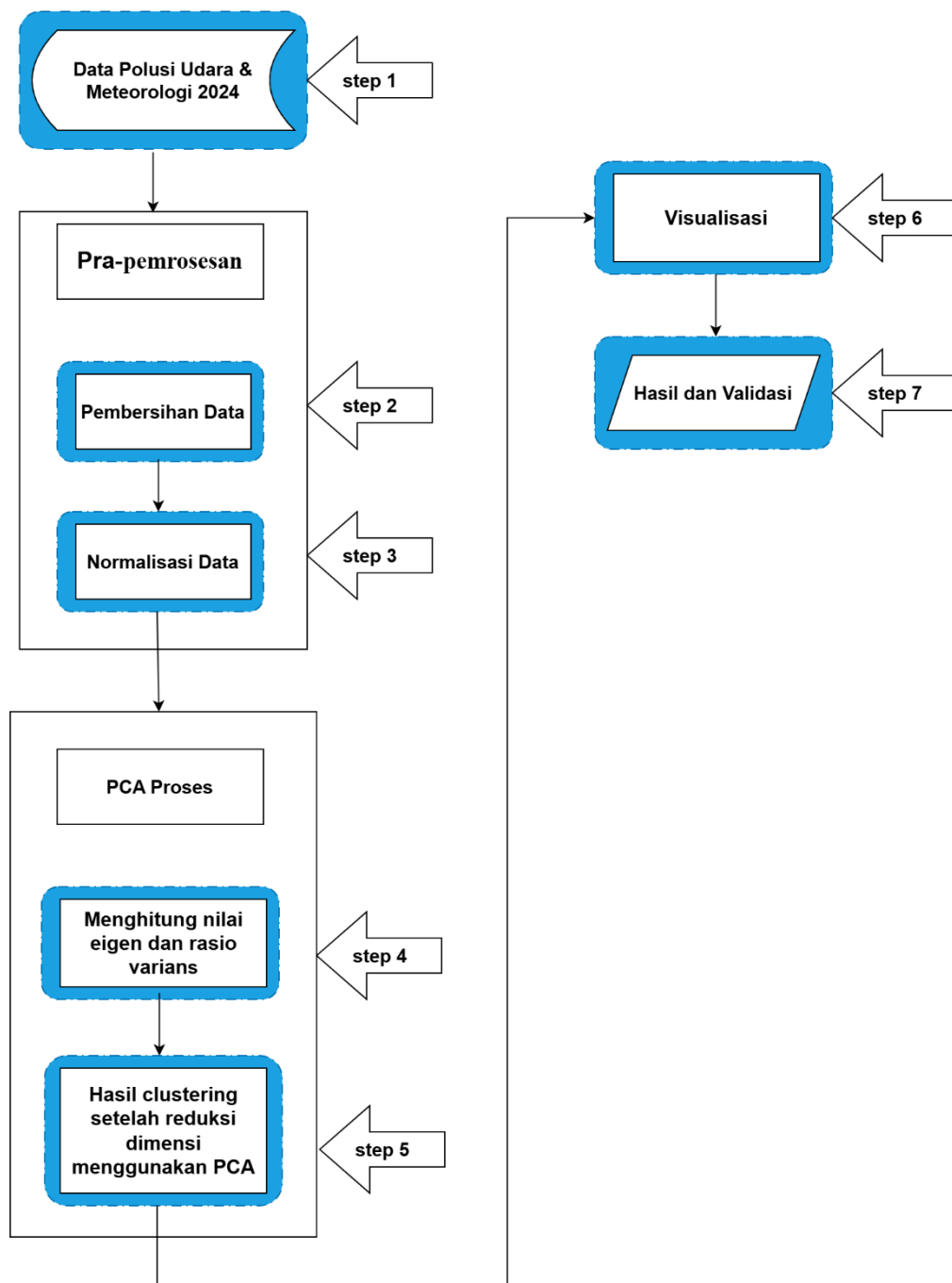
Tahap Proses merupakan inti dari analisis ini. Setelah data dinormalisasi, dilakukan perhitungan nilai *eigenvalue* dan *variance ratio* untuk menentukan komponen utama menggunakan metode *Principal Component Analysis* (PCA). Proses ini bertujuan mereduksi dimensi data meteorologi yang kompleks dengan tetap mempertahankan sebagian besar variansi informasi penting. Komponen utama (*principal components*) yang memiliki kontribusi paling besar terhadap variansi total data dipertahankan dan digunakan sebagai basis koordinat untuk tahap berikutnya, yaitu *Clustering*. Pengelompokan dilakukan menggunakan metode *PCA-Clustering*, yaitu integrasi antara hasil reduksi dimensi PCA dengan algoritma pengelompokan yang difokuskan untuk membedakan antara kondisi kualitas udara normal dan ekstrem berdasarkan kesamaan karakteristik meteorologi harian. Hasil reduksi PCA juga divisualisasikan dalam bentuk grafik dua dimensi untuk menunjukkan distribusi data pada setiap *cluster*.

Tahap Output menghasilkan pengelompokan kondisi kualitas udara yang merepresentasikan variasi pola polusi di wilayah penelitian. Melalui hasil

pengelompokan ini, dapat diidentifikasi periode atau kelompok hari dengan karakteristik cuaca tertentu yang memicu penumpukan polutan. Untuk memastikan keandalan hasil pengelompokan, dilakukan dua bentuk validasi, yaitu validasi internal menggunakan *Silhouette Score* untuk mengukur kualitas dan pemisahan antar *cluster*, serta validasi eksternal melalui perbandingan dengan data target dari BMKG (validasi ahli). Kedua validasi tersebut memastikan bahwa hasil *Clustering* yang diperoleh tidak hanya konsisten secara statistik, tetapi juga sesuai dengan kondisi empiris di lapangan. Dengan demikian, metode PCA-*Clustering* yang diterapkan dalam penelitian ini mampu memberikan hasil pengelompokan polusi udara yang lebih akurat, efisien, dan interpretatif.

3.2 Desain Sistem

Pada penelitian ini bertujuan untuk melakukan pengelompokkan polusi udara berdasarkan faktor meteorologi dengan menggunakan metode dan PCA-*Clustering*. Tujuan dari pengelompokkan ini adalah untuk mengidentifikasi pola polusi udara dalam beberapa kluster, sehingga dapat digunakan untuk mendukung mitigasi risiko kesehatan masyarakat, meningkatkan pelayanan, dan rekomendasi tata ruang berkelanjutan. Diagram alur yang diusulkan studi penelitian ini terdapat pada Gambar 3.2



Gambar 3.2 Desain Sistem

Penelitian ini menggunakan pendekatan kuantitatif dan menggunakan strategi eksploratif untuk melakukan pengelompokkan polusi udara berdasarkan faktor meteorologi. Teknik utama yang digunakan dalam penelitian ini adalah PCA-

Clustering, yang bertujuan untuk mengelompokkan polusi udara ke dalam pengelompokan tertentu berdasarkan faktor meteorologi. Tahapan penelitian meliputi:

Langkah 1: Data

Data penelitian diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Jawa Timur, yang bersumber dari Stasiun Klimatologi Karangploso dan Stasiun Geofisika Sumberpucung melalui situs resmi <https://dataonline.bmkg.go.id>. Dataset yang dianalisis mencakup faktor meteorologi harian selama periode 1 Januari sampai 31 Desember 2024.

Langkah 2: Pembersihan Data

Pembersihan data dilakukan untuk mengatasi nilai yang hilang (*missing value*), ketidakonsistenan, atau pencilan (*outlier*) yang dapat mengganggu jalannya analisis. Tahapan pra-pemrosesan ini sangat krusial untuk menjamin keakuratan serta keandalan (*reliability*) dataset sebelum digunakan dalam analisis pemodelan lebih lanjut, sehingga mampu mencegah dominasi bias dan menjaga integritas data. Proses ini menjadi syarat mutlak mengingat algoritma berbasis *Principal Component Analysis* (PCA) dan metode *clustering* sangat sensitif terhadap kekosongan maupun distorsi data pencilan (Jolliffe & Cadima, 2016)

Langkah 3: Normalisasi Data

Proses normalisasi data bertujuan untuk menyamakan skala data menggunakan metode tertentu agar tidak ada variabel yang mendominasi analisis akibat perbedaan rentang nilai dan satuan. Sebagai contoh, parameter $PM_{2.5}$ ($\mu g/m^3$) dan

rainfall atau curah hujan (mm) harus dinormalisasi untuk memastikan kontribusi yang seimbang antarvariabel (Géron, 2019). Dalam penelitian mengenai pengelompokan polusi udara berdasarkan faktor meteorologi ini, transformasi nilai ke dalam skala seragam diimplementasikan menggunakan metode *Standard Scaler* (SS) atau standarisasi *Z-score* (Khan & Maity, 2020), sehingga menghasilkan nilai rata-rata (*mean*) nol dan standar deviasi satu.

Proses standarisasi nilai dihitung berdasarkan Persamaan (3-1)

$$X_{standard} = \frac{X - \bar{X}}{\sigma} \quad (3-1)$$

Standar deviasi menggunakan persamaan (3-2)

$$\sigma = \frac{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}{n - 1} \quad (3-2)$$

Langkah 4: Penentuan Nilai Eigen (*Eigenvalue*) dan Rasio Varians

Setelah data melalui tahap normalisasi, proses dilanjutkan dengan menghitung nilai Eigen (*eigenvalue*) dan rasio varians untuk setiap komponen utama (*Principal Component* / PC). Kedua parameter ini memiliki peran krusial dalam merepresentasikan distribusi data serta mengukur seberapa besar proporsi informasi dari varians data mentah yang mampu dipertahankan oleh masing-masing PC. Semakin besar nilai Eigen dan rasio varians yang dimiliki oleh suatu komponen utama, maka semakin tinggi pula kontribusi informasi signifikan yang dibawa oleh PC tersebut dalam menggambarkan karakteristik data (Almais et al., 2023).

Dalam penelitian ini, jumlah PC yang dipilih ditetapkan berdasarkan ambang batas rasio varians kumulatif yang signifikan, yaitu berada pada rentang 85%–95% dari total varians keseluruhan. Komponen utama dibentuk dengan

memilih sejumlah nilai Eigen terbesar yang mampu memenuhi kriteria kumulatif tersebut. Reduksi dimensi melalui penetapan jumlah PC dalam konteks ini difokuskan untuk mendukung proses pelabelan data secara optimal, bukan untuk mengidentifikasi fitur dominan secara terpisah.

Besarnya nilai varians (s^2) dari masing-masing variabel harian yang menjadi dasar perhitungan matriks kovarians dalam PCA dihitung menggunakan Persamaan (3-3)

$$s^2 = \frac{\sum_{i=1}^n (X - \bar{X})^2}{n - 1} \quad (3-3)$$

Langkah 5: Hasil Pengelompokan Analisis Komponen Utama

Setelah komponen utama diperoleh melalui proses PCA (Almais et al., 2023), *Clustering* merupakan teknik untuk mengelompokkan data ke dalam beberapa kelompok homogen. Pada penelitian ini, proses *Clustering* dilakukan dengan memanfaatkan titik koordinat *principal component* (PC) berdasarkan rentang nilai tertentu. Adapun rentang nilai yang dipakai adalah sebagai berikut.

- a. Kondisi Normal
- b. Kondisi Ekstrem

Langkah 6: Visualisasi

Hasil pengelompokan (*clustering*) polusi udara selanjutnya dianalisis berdasarkan komponen utama yang diperoleh dari proses *Principal Component Analysis* (PCA) (J.M.A.U. Jayasekara, 2024). Melalui analisis berbasis visualisasi ini, pola pengelompokan dapat dikenali dengan lebih mudah karena dataset multivariat yang

telah direduksi dimensinya mampu menyajikan struktur informasi yang jauh lebih sederhana tanpa kehilangan esensi informasi utamanya. Representasi grafis ini mempermudah interpretasi sebaran data serta memperjelas jarak antar-kluster potensial yang terbentuk.

Langkah 7: Interpretasi Hasil dan Validasi

Interpretasi hasil dan validasi data merupakan tahap akhir dalam alur metode *Principal Component Analysis* (PCA) untuk *Clustering* polusi udara berdasarkan faktor meteorologi. Interpretasi dilakukan dengan menganalisis secara mendalam grafik hasil pengelompokan berdasarkan karakteristik sebaran titik koordinat komponen utama serta rentang nilai polutan yang telah ditetapkan sebagai acuan. Validasi memiliki peran krusial untuk memastikan kesesuaian, ketepatan, dan keandalan hasil pengelompokan yang terbentuk. Proses validasi dilakukan melalui dua pendekatan utama, yaitu pengukuran metrik evaluasi internal menggunakan *Silhouette Score* untuk menguji kekuatan pemisahan antar-klaster, serta analisis komparatif antara hasil pengelompokan berbasis komponen utama pertama (PC1) dengan visualisasi tren data aktual. Melalui kombinasi metode ini, tingkat kesalahan pengelompokan dapat diminimalisasi sehingga diperoleh konfigurasi klaster kualitas udara yang paling optimal dan representatif.

3.2 Data Preparation

Data yang digunakan dalam penelitian ini diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Jawa Timur, dengan fokus pengamatan utama pada Stasiun Klimatologi Karangates. Pemilihan lokasi ini

didasarkan pada ketersediaan data yang komprehensif serta kemampuannya dalam merepresentasikan karakteristik atmosfer di wilayah penelitian.

Dataset ini mencakup rentang waktu pengamatan selama satu tahun, terhitung mulai dari 1 Januari hingga 31 Desember 2024, dengan resolusi harian. Secara keseluruhan, terdapat 6 variabel meteorologi dan polutan yang dianalisis sebagaimana dirinci pada Tabel 3.1, dengan total sampel sebanyak 367 entitas data. Sebelum dilakukan analisis menggunakan metode *PCA-Clustering*, seluruh data telah melalui tahap pemeriksaan awal untuk memastikan konsistensi dan kelengkapan variabel penelitian.

Tabel 3.1 Keterangan Variabel Dataset

No	Nama Variabel	Satuan	Keterangan Ilmiah
1	Partikulat (2.5)	$\mu\text{g}/\text{m}^3$	Konsentrasi partikel udara berukuran ≤ 2.5 mikrometer, berpengaruh pada kesehatan pernapasan.
2	Temperatur Rata-rata	$^{\circ}\text{C}$	Rata-rata suhu harian; indikator kondisi termal yang memengaruhi kualitas udara.
3	Kelembapan Rata-rata	%	Rata-rata kelembapan relatif udara; kelembapan tinggi dapat meningkatkan adsorpsi polutan.
4	Curah Hujan	mm	Total presipitasi harian; hujan dapat menurunkan konsentrasi PM2.5 melalui wet deposition.
5	Penyinaran Matahari	Jam	Durasi paparan sinar matahari; memengaruhi fotokimia dan penguapan polutan.
6	Kecepatan Angin Rata-rata	m/s	Rata-rata kecepatan angin harian; indikator kemampuan angin membawa partikulat.

Tabel 3.1 menjelaskan variabel-variabel klimatologi yang digunakan dalam penelitian, yang diperoleh dari data observasi stasiun klimatologi BMKG. Setiap variabel disajikan dengan satuan pengukuran yang sesuai standar meteorologi, serta diberikan definisi ilmiah yang mengacu pada pedoman BMKG dan WMO. Variabel-variabel ini digunakan untuk analisis tren iklim, peramalan cuaca, hingga pemodelan fenomena lingkungan.

Tabel 3.2 Data Partikulat dan Meteorologi

pm_2.5	52,0	79,0	...	-	-
temp_avg	27,1	27,1	...	24,3	24,6
humidity_avg	78,0	77,0	...	83,0	84,0
rainfall	8,6	0,0	...	1,6	4,8
sunlight	6,4	6,1	...	0,0	5,7
wind_avg	180,0	80,0	...	110,0	170,0

Tabel 3.2 menjelaskan nilai-nilai aktual dari setiap variabel yang diamati setiap harinya, mulai dari konsentrasi PM_{2.5} hingga parameter cuaca lainnya seperti suhu dan kecepatan angin. Data ini menjelaskan variasi harian yang menjadi masukan utama dalam proses normalisasi dan reduksi dimensi pada tahap analisis selanjutnya. Penabelan ini sangat penting dalam tahap *data preparation* guna memastikan bahwa seluruh atribut data telah terstruktur dengan benar sebelum dilakukan pengolahan statistik menggunakan PCA-Clustering.

3.3 Skenario Evaluasi

Skenario evaluasi algoritma merupakan tahapan penting dalam penelitian ini untuk mengukur kinerja model yang dibangun. Dalam penelitian ini, evaluasi

dilakukan dengan menggunakan evaluasi khusus pengelompokan yaitu Silhouette Score. Silhouette Score adalah metrik yang digunakan untuk mengevaluasi kualitas hasil pengelompokan data (clustering). Skor ini mengukur seberapa mirip setiap titik data dengan klaster tempatnya berada dibandingkan dengan klaster lain. Semakin tinggi *Silhouette Score*, semakin baik kualitas pengelompokan (Rousseeuw, 1987).

Rumus *Silhouette Score*:

Untuk suatu data point i dalam klaster:

1. Hitung jarak rata-rata ($a(i)$) dari data point i ke semua data point lain dalam klaster yang sama (Friedman, 2009):

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j) \quad (3)$$

Keterangan:

- i : Data point ke- i yang sedang dievaluasi.
- $a(i)$: Jarak rata-rata dari data point i ke semua data point lain dalam klaster yang sama (C_i). Mengukur kepadatan intra-klaster (how compact the cluster is).
- $|C_i|$: Jumlah data point dalam klaster C_i .
- C_i : Klaster tempat data point i berada.
- $d(i, j)$: Jarak (distance) antara data point i dan data point j . Jarak ini bisa dihitung menggunakan metode seperti Euclidean, Manhattan, atau Cosine Similarity, tergantung konteks data.

2. Hitung jarak rata-rata minimum ($b(i)$) dari data point i ke semua data point di klaster terdekat lainnya:

$$b(i) = \min_{C_k \neq C_i} \left(\frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \right) \quad (4)$$

Keterangan:

- $b(i)$: Jarak rata-rata terkecil dari data point i ke semua data point di klaster terdekat lainnya (klaster selain

C_i). Mengukur pemisahan inter-klaster (how well-separated the clusters are).

3. Skor Silhouette untuk data point i didefinisikan sebagai (Tan, P.-N., Steinbach, M., & Kumar, 2006):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (5)$$

Keterangan:

$s(i)$: Skor Silhouette untuk satu data point i . Nilainya berkisar antara -1 hingga 1.

Interpretasi : Jika $b(i) \gg a(i)$ (klaster terpisah baik), maka $s(i) \rightarrow 1$.
 Jika $b(i) \approx a(i)$ (klaster tumpang tindih), maka $s(i) \rightarrow 0$.
 Jika $b(i) \ll a(i)$ (data point salah klaster), maka $s(i) \rightarrow -1$.

4. Skor Silhouette keseluruhan adalah rata-rata dari semua $s(i)$:

$$\text{Silhouette Scor} = \frac{1}{N} \sum_{i=1}^N s(i) \quad (6)$$

Keterangan:

N : Total jumlah data point dalam dataset.

BAB IV

HASIL PENELITIAN

4.1 Standarisasi Data

Penyajian hasil penelitian diawali dengan tahap pra-pemrosesan data untuk menjamin kualitas dataset. Sebelum dilakukan standarisasi, data terlebih dahulu melalui tahap interpolasi guna menangani data yang hilang (*missing values*) dan menjamin kontinuitas dataset. Langkah ini memastikan bahwa 367 data meteorologi yang digunakan dalam penelitian ini tersedia secara utuh sebelum masuk ke tahap analisis.

Setelah melalui tahap interpolasi untuk menangani data yang hilang (*missing values*) dan menjamin kontinuitas dataset, data kemudian diproses menuju tahap standarisasi. Langkah ini dilakukan sebagai prasyarat utama untuk menyesuaikan seluruh parameter masukan agar berada pada skala yang seragam. Hal ini menjadi penting karena algoritma *Principal Component Analysis* (PCA) berbasis pada perhitungan varians yang sangat sensitif terhadap besarnya magnitudo setiap variabel. Mengingat variabel dalam penelitian ini memiliki satuan yang beragam seperti $PM_{2.5}$ ($\mu g/m^3$), temperatur $^{\circ}C$, kelembapan (%), curah hujan (mm), penyinaran matahari (jam), serta kecepatan angin (m/s) standarisasi wajib dilakukan untuk menyamakan kontribusi setiap variabel dalam pembentukan komponen utama.

Proses transformasi dilakukan dengan metode *StandardScaler* melalui perhitungan skor-Z (*Z-score*) sehingga menghasilkan nilai rata-rata (*mean*) nol dan

standar deviasi satu. Hasil transformasi pada dataset penelitian yang berjumlah 367 data tersebut disajikan pada Tabel 4.1 sebagai berikut:

Tabel 4.1 Hasil Standarisasi Variabel Menggunakan *StandardScaler*

PM _{2.5}	temp_avg	humidity_avg	rainfall	sunlight	wind_avg
-1.153998388	0.035726014	0.248900674	0.20996807	-0.123126018	0.486312401
0.07520506	0.035726014	0.170899435	-0.416889083	-0.13641731	-1.152640995
0.894674026	-0.037381053	0.326901913	-0.416889083	-0.118695588	-1.316536335
0.940200079	-0.068712654	0.170899435	-0.416889083	-0.202873766	-0.988745656
0.257309275	-0.162707455	0.638906868	-0.416889083	-0.180721614	0.158521722

Data yang disajikan pada Tabel 4.1 menunjukkan hasil transformasi terhadap 367 sampel data meteorologi. Setiap baris dalam tabel ini merepresentasikan satu sampel dari dataset asli, sementara setiap kolom merepresentasikan fitur asli yang telah diseragamkan skalanya melalui metode *StandardScaler*. Penggunaan nilai negatif pada hasil standarisasi mengindikasikan bahwa data berada di bawah nilai rata-rata (*mean*), sedangkan nilai positif menunjukkan posisi data berada di atas rata-rata.

Dengan diterapkannya transformasi skala tersebut, kontribusi setiap variabel dalam proses reduksi dimensi menjadi setara. Dataset yang telah distandarisasi ini kemudian menjadi masukan utama untuk proses transformasi ke ruang komponen utama (*Principal Component Analysis*) guna meminimalkan bias yang diakibatkan oleh perbedaan magnitudo pada variabel asli.

4.2 Hasil Analisis *Principal Component Analysis* (PCA)

Analisis reduksi dimensi menggunakan metode *Principal Component Analysis* (PCA) diterapkan pada dataset yang telah memenuhi kriteria standarisasi melalui transformasi skor-Z. Prosedur ini dilakukan untuk menyarikan informasi dari berbagai variabel meteorologi dan polusi udara ke dalam bentuk komponen utama yang lebih efisien tanpa menghilangkan karakteristik esensial dari data asli. Melalui ekstraksi ini, kompleksitas dimensi data dapat direduksi sehingga hubungan antarvariabel dapat divisualisasikan dan diinterpretasikan secara lebih akurat pada ruang koordinat yang baru.

4.2.1 Nilai *Eigenvalue* dan *Explained Variance Ratio*

Penentuan jumlah komponen utama didasarkan pada nilai *Eigenvalue* yang dihasilkan dari matriks kovarians data. Berdasarkan kriteria statistik, komponen yang memiliki kontribusi paling signifikan dalam menjelaskan variabilitas dataset dipilih untuk mewakili dimensi baru. Ringkasan kontribusi masing-masing komponen disajikan pada Tabel 4.2 berikut:

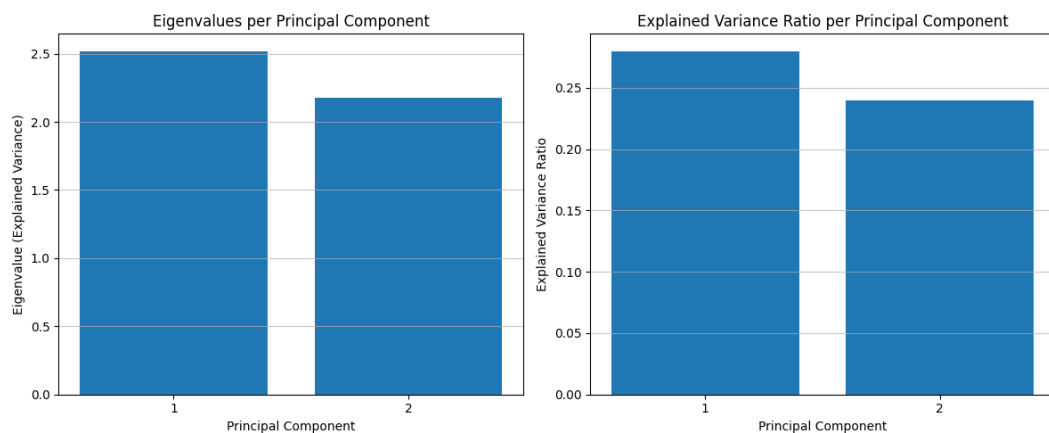
Tabel 4.2 Nilai *Eigenvalue* dan Rasio Variansi Komponen Utama

Number of component	Eigenvalue	Variance ratio (%)
1.0	2.83	47.06
2.0	1.20	20.09
Total		67.15

Berdasarkan Tabel 4.2, hasil ekstraksi menunjukkan bahwa komponen utama pertama (PC1) memiliki nilai *Eigenvalue* sebesar 2.83, yang mampu menjelaskan 47.06% variabilitas data asli. Sementara itu, komponen utama kedua

(PC2) memiliki nilai *Eigenvalue* sebesar 1.20 dengan kontribusi varians sebesar 20.09%. Secara kumulatif, kedua komponen utama ini mampu merangkum 67.15% informasi dari keseluruhan variabel meteorologi dan polusi udara yang digunakan dalam penelitian.

Nilai akumulasi varians yang mencapai 67.15% ini menunjukkan bahwa reduksi dimensi dari enam variabel menjadi dua komponen utama sudah cukup representatif, karena telah melampaui ambang batas minimum 60% yang umum digunakan dalam analisis data lingkungan. Untuk memberikan gambaran visual mengenai dominasi PC1 dan PC2 dalam menangkap informasi dataset, hasil tersebut divisualisasikan pada Gambar 4.1 berikut:



Gambar 4.1 *Eigenvalues* dan *Explained Variance Ratio*

Gambar 4.1 menyajikan perbandingan nilai *Eigenvalue* dan rasio varians untuk masing-masing komponen. Grafik tersebut mempertegas bahwa PC1 merupakan dimensi yang paling signifikan dalam menjelaskan pola data, sementara PC2 berperan sebagai komponen pendukung yang melengkapi total informasi agar

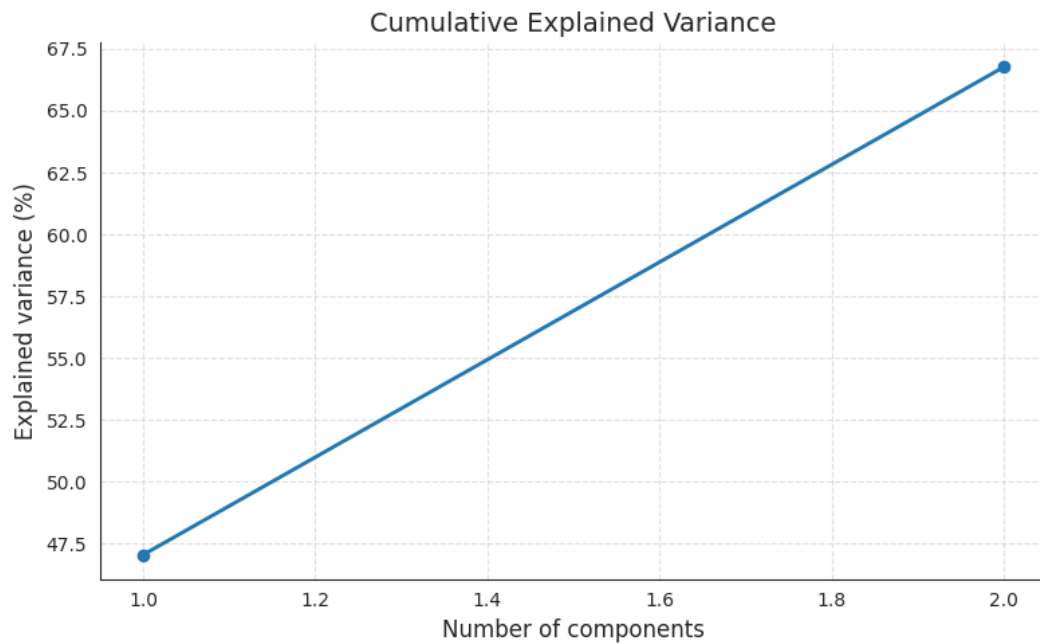
proses pengelompokan (*clustering*) pada tahap selanjutnya memiliki tingkat akurasi yang lebih tinggi.

4.3 Hasil Pengelompokan Analisis Komponen Utama

Penentuan jumlah *principal component* (PC) dalam penelitian ini dilakukan dengan mengevaluasi nilai *eigenvalue* dan *variance ratio*. Merujuk pada Tabel 4.2, komponen utama pertama (PC1) mencatatkan nilai *eigenvalue* sebesar 2.83 dengan kontribusi *variance ratio* mencapai 47.06%. Perlu ditegaskan bahwa dalam penelitian ini, nilai *eigenvalue* terbesar tidak menjadi satu-satunya parameter acuan karena fokus utama tahapan ini adalah untuk pelabelan data polusi udara, bukan sekadar mengidentifikasi fitur yang paling dominan.

Analisis ini juga melibatkan komponen utama kedua (PC2) yang memiliki nilai *eigenvalue* sebesar 1.20 dengan *variance ratio* 20.09%. Meskipun nilainya lebih rendah daripada PC1, integrasi PC2 berfungsi sebagai parameter komparatif dalam mengevaluasi akurasi pelabelan data polusi berdasarkan variabel meteorologi, yang meliputi suhu rata-rata, kelembapan, curah hujan, penyinaran matahari, dan kecepatan angin.

Distribusi nilai *variance ratio* untuk PC1 dan PC2 yang disajikan pada Gambar 4.2 menunjukkan tren akumulasi linier. Hal ini mengindikasikan bahwa PC2 memberikan kontribusi sebaran data yang signifikan dalam melengkapi cakupan informasi yang telah dipetakan oleh PC1.



Gambar 4.2 *Graph of PC1 and PC2 Variance ratio Value*

Gambar 4.2 menunjukkan hubungan antara jumlah komponen utama dengan persentase variasi data yang dapat dijelaskan. Komponen pertama (PC1) mampu menjelaskan 47.06% variasi, sedangkan penambahan komponen kedua (PC2) meningkatkan total variasi kumulatif menjadi 67.15%. Hal ini menandakan bahwa penggunaan dua komponen utama sudah cukup untuk menjelaskan sebagian besar informasi dari data polusi *Particulate Matter* ($PM_{2.5}$) dan faktor meteorologi terkait, sehingga dapat digunakan untuk reduksi dimensi tanpa kehilangan terlalu banyak informasi penting dari dataset asli.

4.4 Penentuan Label dan Evaluasi *Silhouette Score*

Penentuan label untuk mengkategorikan kondisi polusi udara dilakukan berdasarkan hasil reduksi dimensi yang telah diselesaikan melalui proses *Principal Component Analysis* (PCA). Langkah ini bertujuan untuk mengorganisir data ke

dalam kelompok-kelompok yang bersifat homogen agar dapat mengidentifikasi pola tersembunyi (*hidden patterns*) dalam data polusi.

4.4.1 Analisis Kualitas Klaster dengan *Silhouette Score*

Penentuan jumlah klaster (k) yang paling optimal dilakukan untuk menjamin validitas pengelompokan secara statistik terhadap dataset polusi udara Kota Malang. Prosedur ini melibatkan perbandingan nilai *Silhouette Score* pada beberapa skenario jumlah kelompok, yaitu dari skenario $k=2$ hingga $k=6$. Implementasi metode ini bertujuan untuk mengukur efektivitas pemisahan antar-klaster (*separation*) serta tingkat konsistensi data di dalam masing-masing kelompok (*cohesion*) melalui perhitungan koefisien jarak rata-rata antar-titik koordinat pada ruang komponen utama.

Pengujian melalui *Silhouette Score* dalam penelitian ini diposisikan sebagai instrumen validasi untuk memastikan bahwa jumlah kelompok yang dipilih mampu merepresentasikan perbedaan karakteristik atmosfer secara signifikan. Nilai koefisien yang dihasilkan memberikan indikasi numerik mengenai seberapa tepat sebuah titik data teralokasi pada klasternya dibandingkan dengan potensi tumpang tindih (*overlap*) pada klaster lain. Kedalaman analisis ini diperlukan mengingat distribusi data polusi udara dan variabel meteorologi memiliki kompleksitas tinggi dan sering kali saling beririsan pada ruang komponen utama.

Identifikasi konfigurasi klaster yang paling stabil dilakukan dengan mengevaluasi berbagai skenario tersebut untuk memastikan jarak antar-kelompok berada pada nilai maksimal, sementara variansi di dalam kelompok berada pada nilai minimal. Stabilitas struktur ini menjadi landasan statistik yang menentukan

untuk mendukung validitas pelabelan pada tahapan selanjutnya, sehingga klasifikasi antara kondisi normal dan ekstrem didasarkan pada pemisahan fitur yang nyata serta objektif. Hasil pengukuran parameter tersebut disajikan pada Tabel 4.3 berikut:

Tabel 4.3 Hasil Evaluasi Jumlah Kluster Berdasarkan *Silhouette Score*

Jumlah Kluster (k)	Silhouette Score
2	0.8683(<i>tertinggi</i>)
3	0.8232
4	0.8423
5	0.8073
6	0.8244

Perolehan skor tertinggi sebagaimana disajikan pada Tabel 4.3 dicapai oleh skenario dua kluster ($k = 2$) dengan nilai sebesar 0.8683. Merujuk pada kriteria interpretasi kekuatan kluster, seluruh nilai koefisien siluet yang dihasilkan dari rentang skenario $k = 2$ hingga $k = 6$ secara absolut berada di atas angka 0.80, yang mengindikasikan bahwa struktur pengelompokan yang terbentuk di setiap skenario dikategorikan sebagai *strong structure*. Hasil ini memberikan justifikasi teknis bahwa pemisahan data menggunakan pendekatan ambang batas jangkauan (*threshold*) memiliki tingkat reliabilitas yang sangat tinggi dan tingkat kesalahan alokasi data yang minimal di semua tingkatan konfigurasi jumlah kluster.

Kalkulasi rata-rata *Silhouette Score* secara keseluruhan pada model terpilih menunjukkan angka sebesar 0.87. Pencapaian tersebut membuktikan bahwa tingkat

pemisahan antar-kelompok dalam penelitian ini telah terdefinisi secara tegas, di mana konstituen di dalam setiap kelompok memiliki tingkat kemiripan fitur yang sangat rapat. Tingginya nilai 0.87 mengonfirmasi keberhasilan algoritma dalam mengidentifikasi pola yang konsisten tanpa adanya tumpang tindih (*overlap*) yang signifikan antar-sampel, yang merepresentasikan kondisi di mana jarak intra-klaster berada pada titik minimal dan jarak antar-klaster berada pada titik maksimal.

Analisis mendalam terhadap tren nilai tersebut menunjukkan karakteristik fisis yang unik, di mana seiring meningkatnya nilai k , super-klaster utama berkapasitas 357 data harian normal tetap terkonsolidasi secara utuh tanpa mengalami perpecahan struktur internal. Fluktuasi nilai koefisien siluet yang bergerak dari 0.8232 pada $k = 3$, naik ke 0.8423 pada $k = 4$, lalu menyentuh titik terendah 0.8073 pada $k = 5$, dan kembali naik ke 0.8244 pada $k = 6$ sepenuhnya merupakan representasi geometris dari dinamika alokasi 9 data pencilan ekstrem di sisi marginal ruang komponen utama. Pola naik-turun ini mengonfirmasi bahwa penambahan faset kelompok baru dari $k = 3$ hingga $k = 6$ tidak mengganggu stabilitas data harian utama, melainkan hanya membagi variansi fisis dari kondisi anomali cuaca penekan polutan secara lebih spesifik.

Data polusi udara dan meteorologi yang telah melalui proses reduksi dimensi PCA menunjukkan kecenderungan alami untuk terbagi ke dalam dua pola besar yang berbeda secara fundamental. Kondisi tersebut menjadi fondasi utama bagi penelitian ini untuk melanjutkan tahapan berikutnya, yaitu penetapan ambang batas (*threshold*) pada sumbu komponen utama untuk pengelompokan kondisi

polusi udara menjadi kategori Normal dan Ekstrem. Keberhasilan dalam mencapai struktur klaster yang kuat ini menjamin bahwa pelabelan yang dilakukan pada tahap selanjutnya memiliki tingkat akurasi yang dapat dipertanggungjawabkan secara ilmiah.

4.4.2 Penetapan Ambang Batas (*Threshold*) Pengelompokan

Penetapan ambang batas (*threshold*) dilakukan sebagai langkah operasional untuk memisahkan seluruh sampel data ke dalam dua kategori utama berdasarkan struktur klaster yang telah divalidasi. Prosedur ini memanfaatkan distribusi nilai pada sumbu komponen utama pertama (PC1), mengingat fitur tersebut merepresentasikan variansi informasi paling dominan dalam sistem data polusi udara. Berdasarkan observasi sebaran spasial pada ruang koordinat dua dimensi, ditemukan adanya diskontinuitas pola yang memungkinkan dilakukannya pembagian kelompok secara tegas.

Titik ambang batas yang ditetapkan dalam penelitian ini adalah sebesar 5.0 pada sumbu PC1. Angka tersebut dipilih secara objektif karena menjadi garis batas geometris yang memisahkan kumpulan data dengan kepadatan tinggi pada area pusat (*centroid*) normal dengan sekumpulan data yang mengalami pencaran signifikan pada area ekstrem. Penggunaan satu sumbu dominan sebagai basis *thresholding* bertujuan untuk menjamin konsistensi pelabelan, sehingga setiap sampel data memiliki kepastian posisi kelompok tanpa adanya ambiguitas interpretatif.

Penerapan nilai ambang batas tersebut menghasilkan pengelompokan dataset polusi udara ke dalam dua kriteria sebagai berikut:

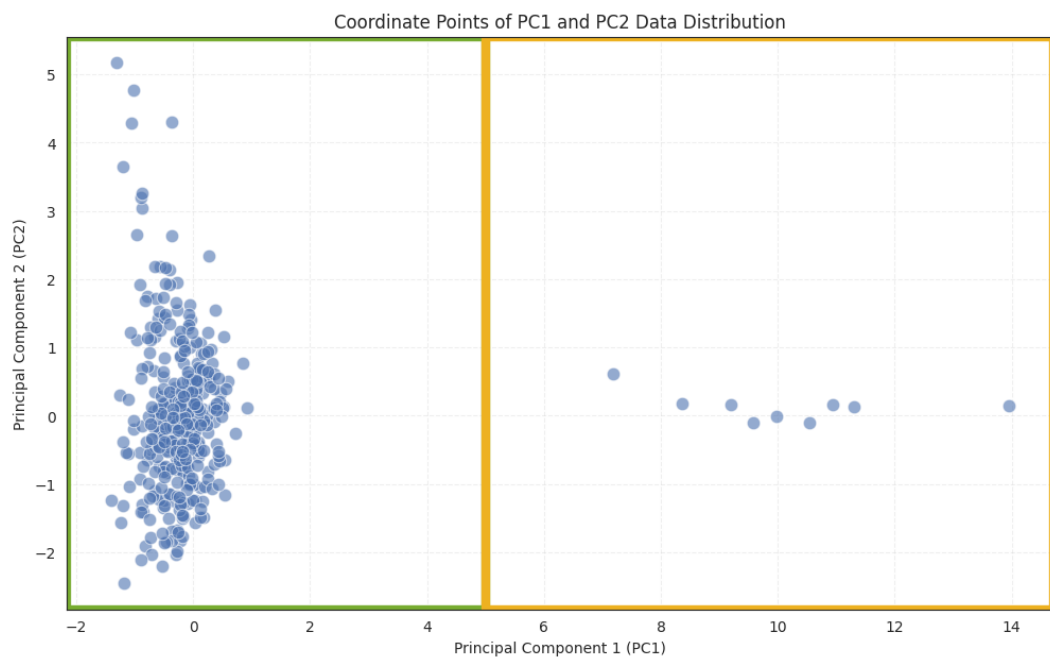
1. Kondisi Normal: Ditujukan bagi sampel data yang memiliki nilai koordinat $PC1 < 5.0$. Kelompok ini mencakup mayoritas distribusi data yang merepresentasikan stabilitas parameter atmosfer, di mana konsentrasi polutan $PM_{2.5}$ berada dalam rentang fluktuasi rata-rata harian yang aman.
2. Kondisi Ekstrem: Ditujukan bagi sampel data yang memiliki nilai koordinat $PC1 \geq 5.0$. Kelompok ini mengidentifikasi sekumpulan sampel yang menunjukkan anomali atau lonjakan nilai yang signifikan, yang dipicu oleh interaksi ekstrem antara variabel meteorologi dan sumber polutan.

Strategi pelabelan berbasis ambang batas pada PC1 ini memastikan bahwa stratifikasi data dilakukan secara konsisten dengan hasil evaluasi *silhouette score* yang telah dibahas sebelumnya. Dengan demikian, setiap label yang dihasilkan dalam tahapan ini memiliki validitas statistik yang kuat dan merepresentasikan fenomena fisik polusi udara yang terjadi di lapangan secara akurat.

4.5 Visualisasi Hasil Pengelompokan Data

Pemetaan distribusi sampel data pada ruang komponen utama dilakukan guna memberikan justifikasi visual terhadap efektivitas proses pengelompokan yang telah dilaksanakan. Melalui diagram pencar (*scatter plot*), koordinat setiap sampel pada sumbu PC1 dan PC2 diproyeksikan untuk mengamati pola sebaran spasial secara komprehensif. Penggunaan kedua sumbu ini didasarkan pada

kemampuannya dalam mengekstraksi variansi informasi paling dominan dari dataset polusi udara, sehingga posisi setiap titik data merepresentasikan karakteristik fisik dari variabel meteorologi dan polutan secara akurat. Distribusi spasial tersebut disajikan pada Gambar 4.3.



Gambar 4.3 *Coordinate Points of PC1 and PC2 Data Distribution*

Gambar 4.3 menjelaskan adanya pemisahan kelompok data melalui penerapan nilai ambang batas (*threshold*) sebesar 5.0 pada sumbu PC1. Garis kuning pada gambar tersebut merupakan batas yang membagi dataset secara fisik berdasarkan sebaran titik koordinat hasil clustering. Sampel data yang berkumpul pada area hijau (sisi kiri garis) merupakan kondisi normal yang menunjukkan stabilitas parameter meteorologi harian. Sementara itu, sampel data yang berada pada area oranye (sisi kanan garis) merupakan kondisi ekstrem, di mana terjadi

pergeseran nilai variabel polutan yang menjauh dari kumpulan data utama akibat pengaruh faktor meteorologi tertentu.

Untuk memperoleh pemahaman yang lebih mendalam mengenai karakteristik masing-masing kelompok yang terbentuk, dilakukan analisis statistik deskriptif terhadap variabel-variabel penelitian. Analisis ini bertujuan untuk mengidentifikasi perbedaan karakteristik antara klaster normal dan klaster ekstrem berdasarkan nilai rata-rata setiap variabel. Hasil perhitungan statistik deskriptif tersebut disajikan pada Tabel 4.8.

Tabel 4.8 Distribusi Data Berdasarkan Klaster

Kategori Klaster	Parameter Ambang Batas (PC1)	Jumlah Data	Keterangan Plot
Kondisi Normal	$PC1 \leq 5.0$	357 Sampel	Hijau (Stabil)
Kondisi Ekstrem	$PC1 > 5.0$	9 Sampel	Oranye (Kritis)

Tabel 4.8 menunjukkan bahwa sebanyak 357 data atau 97.54% dari keseluruhan data termasuk ke dalam klaster kondisi normal dengan nilai PC1 kurang dari atau sama dengan 5.0. Sementara itu, sebanyak 9 data atau 2.46% dari total data tergolong ke dalam klaster kondisi ekstrem dengan nilai PC1 lebih besar dari 5.0. Dominasi jumlah data pada klaster kondisi normal mengindikasikan bahwa sebagian besar hasil pengamatan memiliki karakteristik yang relatif seragam, sedangkan klaster kondisi ekstrem merepresentasikan sejumlah data yang memiliki karakteristik berbeda dari pola umum yang terbentuk.

Hasil tersebut sejalan dengan visualisasi pada Gambar 4.3 yang memperlihatkan adanya pemisahan yang jelas antara kedua kelompok pada ruang komponen utama. Kejelasan pemisahan antarklaster tersebut juga didukung oleh

nilai *Silhouette Score* sebesar 0.87 yang menunjukkan kualitas pengelompokan yang sangat baik. Dengan demikian, pendekatan PCA-*Clustering* yang diterapkan mampu menghasilkan pemisahan kelompok data yang jelas berdasarkan karakteristik meteorologi dan konsentrasi polutan yang diamati.

4.6 Perbandingan Kinerja PCA-*Clustering* dan K-Means

Perbandingan kinerja PCA-*Clustering* dan K-Means dilakukan untuk menganalisis perbedaan hasil pengelompokan yang dihasilkan oleh kedua metode pada representasi data yang sama. PCA-*Clustering* merupakan pendekatan yang memanfaatkan hasil reduksi dimensi *Principal Component Analysis* (PCA) untuk mengidentifikasi kelompok data berdasarkan distribusi komponen utama, sedangkan K-Means melakukan pengelompokan berdasarkan kedekatan jarak setiap observasi terhadap pusat klaster (*centroid*).

Penggunaan metode pembandingan diperlukan untuk memberikan gambaran mengenai perbedaan struktur klaster yang dihasilkan oleh kedua pendekatan. Dengan menggunakan data hasil reduksi dimensi yang sama, yaitu komponen utama pertama (PC1) dan komponen utama kedua (PC2), perbedaan hasil pengelompokan yang diperoleh dapat dianalisis berdasarkan karakteristik masing-masing metode tanpa dipengaruhi oleh perbedaan data masukan.

Analisis perbandingan dilakukan melalui tiga tahapan. Tahap pertama adalah implementasi algoritma K-Means pada data hasil PCA. Tahap kedua adalah analisis hasil clustering yang terbentuk berdasarkan visualisasi distribusi data dan jumlah anggota pada setiap klaster. Tahap ketiga adalah evaluasi kualitas

pengelompokan menggunakan metrik *Silhouette Score* untuk membandingkan tingkat kekompakan dan keterpisahan klaster yang dihasilkan oleh metode *PCA-Clustering* dan K-Means.

Hasil perbandingan tersebut memberikan gambaran yang lebih komprehensif mengenai efektivitas metode *PCA-Clustering* dalam mengidentifikasi pola pengelompokan data polusi udara berdasarkan faktor meteorologi dibandingkan dengan metode *clustering* konvensional.

4.6.1 Implementasi Metode K-Means

Algoritma K-Means diterapkan sebagai metode pembanding terhadap pendekatan *PCA-Clustering* yang digunakan dalam penelitian ini. Implementasi dilakukan pada data hasil reduksi dimensi *Principal Component Analysis* (PCA) yang direpresentasikan oleh dua komponen utama, yaitu komponen utama pertama (PC1) dan komponen utama kedua (PC2). Penggunaan kedua komponen utama tersebut didasarkan pada hasil analisis sebelumnya yang menunjukkan bahwa PC1 dan PC2 mampu menjelaskan sebesar 67.15% variasi total data, sehingga dianggap cukup representatif untuk menggambarkan karakteristik data polusi udara dan faktor meteorologi yang diamati.

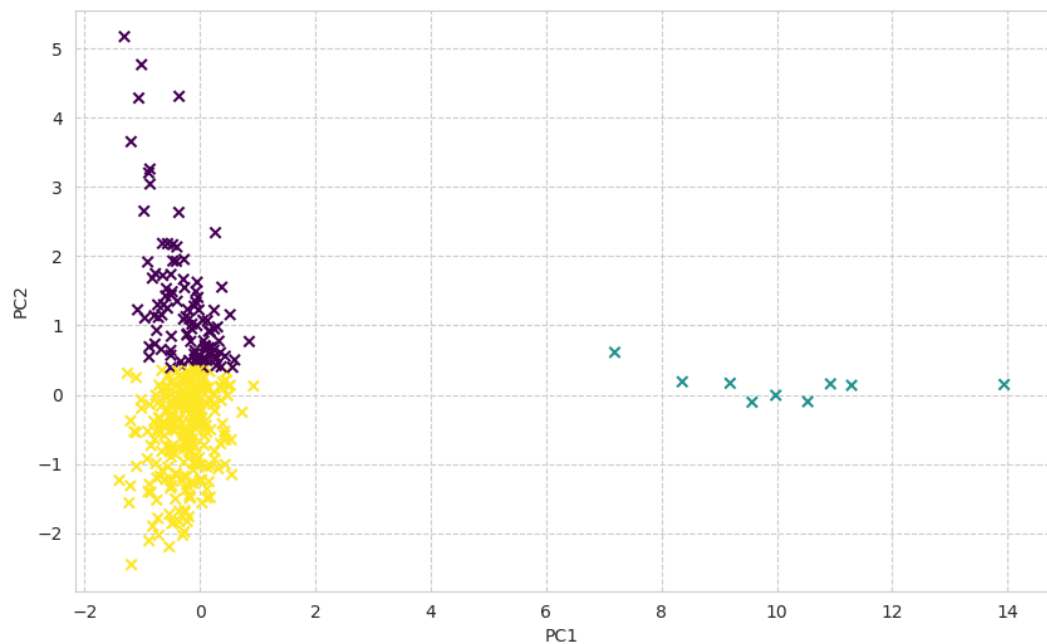
Proses pengelompokan menggunakan K-Means dilakukan dengan jumlah klaster sebanyak tiga kelompok ($k = 3$). Pemilihan jumlah klaster ini dimaksudkan untuk memperoleh representasi pengelompokan yang lebih rinci sehingga dapat digunakan sebagai pembanding terhadap hasil *PCA-Clustering* yang menghasilkan

dua kelompok utama, yaitu kondisi normal dan kondisi ekstrem. Setiap observasi kemudian dikelompokkan berdasarkan jarak terdekat terhadap *centroid* yang dibentuk oleh algoritma K-Means melalui proses iteratif hingga diperoleh kondisi konvergen.

Penerapan K-Means pada ruang komponen utama memberikan keuntungan berupa penyederhanaan struktur data tanpa menghilangkan sebagian besar informasi penting yang terkandung dalam data asli. Selain memudahkan proses komputasi, pendekatan ini juga memungkinkan visualisasi hasil pengelompokan dilakukan secara lebih jelas pada ruang dua dimensi. Hasil *clustering* yang diperoleh selanjutnya dianalisis untuk mengetahui pola distribusi data pada masing-masing klaster dan dibandingkan dengan hasil pengelompokan yang diperoleh melalui pendekatan PCA-*Clustering*.

4.6.2 Hasil *Clustering* Menggunakan K-Means

Pengelompokan data menggunakan algoritma K-Means dilakukan untuk mengidentifikasi kelompok-kelompok observasi yang memiliki karakteristik serupa berdasarkan parameter polusi udara dan faktor meteorologi. Proses pengelompokan dilakukan pada ruang komponen utama pertama (PC1) dan komponen utama kedua (PC2) yang diperoleh dari hasil *Principal Component Analysis* (PCA). Penggunaan ruang komponen utama bertujuan untuk menyederhanakan struktur data multidimensi tanpa menghilangkan informasi utama yang terkandung di dalamnya. Hasil pengelompokan data menggunakan algoritma K-Means disajikan pada Gambar 4.4.



Gambar 4.4 Hasil Pengelompokan Data Menggunakan K-Means

Gambar 4.4 menunjukkan hasil pengelompokan data menggunakan algoritma K-Means pada ruang komponen utama pertama (PC1) dan komponen utama kedua (PC2). Setiap warna merepresentasikan klaster yang berbeda berdasarkan hasil proses pengelompokan yang dilakukan oleh algoritma K-Means. Berdasarkan visualisasi tersebut, data berhasil dikelompokkan ke dalam tiga klaster yang memiliki pola distribusi berbeda pada ruang komponen utama. Sebagian besar observasi terkonsentrasi pada area pusat distribusi data, sedangkan kelompok lainnya membentuk klaster yang terpisah pada posisi tertentu sesuai karakteristik data yang dimiliki. Pola pengelompokan ini menunjukkan bahwa algoritma K-Means mampu mengidentifikasi variasi karakteristik data berdasarkan tingkat kemiripan antar observasi pada ruang hasil reduksi dimensi PCA.

Berdasarkan visualisasi pada Gambar 4.4, algoritma K-Means berhasil membentuk tiga klaster yang menunjukkan pola distribusi yang berbeda pada ruang komponen utama. Pengelompokan tersebut mengindikasikan adanya variasi karakteristik data yang dapat dibedakan berdasarkan kedekatan antar observasi pada ruang hasil reduksi dimensi PCA. Struktur klaster yang terbentuk memberikan gambaran mengenai pola sebaran data yang diperoleh dari kombinasi variabel polusi udara dan faktor meteorologi yang digunakan dalam penelitian.

Distribusi jumlah anggota pada masing-masing klaster disajikan pada Tabel 4.9 Tabel tersebut menunjukkan jumlah observasi yang termasuk ke dalam setiap kelompok hasil pengelompokan menggunakan algoritma K-Means.

Tabel 4.9 Distribusi Jumlah Data Hasil *Clustering* K-Means

Cluster	Jumlah Data
Cluster 0	104
Cluster 1	9
Cluster 2	253
Total	366

Tabel 4.9 menyajikan distribusi jumlah observasi pada masing-masing klaster hasil pengelompokan menggunakan algoritma K-Means. Distribusi tersebut menunjukkan bahwa anggota klaster tidak tersebar secara merata pada seluruh kelompok. Perbedaan jumlah observasi antar klaster mencerminkan adanya variasi karakteristik data yang terbentuk berdasarkan kemiripan pola pada ruang komponen utama. Klaster dengan jumlah anggota terbesar menggambarkan kondisi yang paling dominan selama periode pengamatan, sedangkan klaster dengan jumlah anggota yang lebih sedikit merepresentasikan kondisi yang relatif lebih jarang ditemukan dalam data penelitian.

Hasil pengelompokan yang diperoleh melalui algoritma K-Means selanjutnya digunakan sebagai dasar dalam analisis perbandingan dengan metode *PCA-Clustering*. Perbandingan dilakukan menggunakan nilai *silhouette score* sebagai ukuran kuantitatif untuk mengevaluasi kualitas kluster yang dihasilkan oleh masing-masing metode. Melalui evaluasi tersebut, tingkat kekompakan antar anggota dalam kluster serta tingkat keterpisahan antar kluster dapat dianalisis secara objektif sehingga diperoleh gambaran mengenai metode yang menghasilkan struktur pengelompokan yang lebih baik.

4.6.3 Perbandingan Kinerja *PCA-Clustering* dan K-Means

Perbandingan kinerja antara metode *PCA-Clustering* dan K-Means dilakukan menggunakan nilai *silhouette score* sebagai ukuran kualitas pengelompokan. *silhouette score* digunakan untuk mengevaluasi tingkat kekompakan anggota dalam suatu kluster serta tingkat keterpisahan antar kluster. Semakin tinggi nilai *silhouette score*, semakin baik kualitas pengelompokan yang dihasilkan karena menunjukkan bahwa anggota dalam suatu kluster memiliki tingkat kemiripan yang tinggi dan terpisah dengan baik dari kluster lainnya. Hasil perbandingan nilai *silhouette score* antara metode *PCA-Clustering* dan K-Means disajikan pada Tabel 4.10.

Tabel 4.10 Perbandingan Nilai *Silhouette Score* *PCA-Clustering* dan K-Means

Metode	Jumlah Kluster	<i>Silhouette Score</i>
<i>PCA-Clustering</i>	2	0.8683
K-Means	3	0.4722

Tabel 4.10 menyajikan hasil perbandingan nilai *silhouette score* antara metode PCA-*Clustering* dan K-Means. Berdasarkan tabel tersebut, metode PCA-*Clustering* menghasilkan nilai *silhouette score* sebesar 0.87, sedangkan metode K-Means menghasilkan nilai sebesar 0.47. Perbedaan nilai tersebut menunjukkan bahwa PCA-*Clustering* memiliki tingkat kekompakan dan keterpisahan klaster yang lebih baik dibandingkan dengan K-Means.

Nilai *silhouette score* sebesar 0.87 mengindikasikan bahwa sebagian besar observasi pada metode PCA-*Clustering* berada pada klaster yang sesuai dan memiliki pemisahan yang sangat baik terhadap klaster lainnya. Kondisi ini menunjukkan bahwa pendekatan pengelompokan berdasarkan distribusi komponen utama mampu membentuk struktur klaster yang lebih homogen dan terpisah dengan jelas. Sebaliknya, nilai *silhouette score* sebesar 0.47 pada metode K-Means menunjukkan bahwa kualitas pemisahan antar klaster relatif lebih rendah dibandingkan PCA-*Clustering*.

Perbedaan hasil tersebut dapat dijelaskan oleh karakteristik kedua metode yang berbeda. PCA-*Clustering* dalam penelitian ini memanfaatkan distribusi data pada ruang komponen utama untuk memisahkan observasi ke dalam kelompok kondisi normal dan kondisi ekstrem berdasarkan ambang batas yang telah ditetapkan. Pendekatan tersebut menghasilkan struktur pengelompokan yang lebih jelas karena mengikuti pola distribusi data setelah proses reduksi dimensi. Sementara itu, K-Means membentuk klaster berdasarkan kedekatan jarak terhadap

centroid, sehingga hasil pengelompokan sangat dipengaruhi oleh posisi pusat klaster yang terbentuk selama proses iterasi.

Hasil perbandingan menunjukkan bahwa metode *PCA-Clustering* menghasilkan kualitas pengelompokan yang lebih baik dibandingkan K-Means pada data polusi udara dan faktor meteorologi yang digunakan dalam penelitian ini. Nilai *silhouette score* yang lebih tinggi menunjukkan bahwa klaster yang terbentuk memiliki tingkat kekompakan dan keterpisahan yang lebih baik. Dengan demikian, pendekatan *PCA-Clustering* dinilai lebih efektif dalam mengidentifikasi pola kondisi kualitas udara berdasarkan karakteristik data yang diamati.

BAB V

PEMBAHASAN

5.1 Analisis Pola Pengelompokan Polusi Udara

Hasil pengolahan data menggunakan metode *PCA-Clustering* menunjukkan bahwa data polusi udara dan faktor meteorologi pada wilayah penelitian dapat dikelompokkan ke dalam dua klaster utama, yaitu klaster kondisi normal dan klaster kondisi ekstrem. Pengelompokan tersebut dilakukan berdasarkan kesamaan karakteristik data yang direpresentasikan pada ruang komponen utama, sehingga setiap klaster menggambarkan kondisi kualitas udara yang berbeda selama periode pengamatan.

Berdasarkan hasil pengelompokan, sebanyak 357 data tergolong ke dalam klaster kondisi normal, sedangkan 9 data lainnya tergolong ke dalam klaster kondisi ekstrem. Dominasi jumlah data pada klaster normal menunjukkan bahwa sebagian besar kondisi atmosfer selama periode pengamatan berada pada keadaan yang relatif stabil. Sebaliknya, keberadaan klaster ekstrem menunjukkan adanya kondisi tertentu yang memiliki karakteristik berbeda secara signifikan dibandingkan pola umum yang terbentuk dalam data.

Pemisahan kedua klaster terlihat jelas pada ruang komponen utama, khususnya pada sumbu PC1 yang merepresentasikan variansi terbesar dari keseluruhan variabel penelitian. Sebaran data menunjukkan bahwa klaster normal dan klaster ekstrem menempati wilayah yang berbeda dengan tingkat tumpang

tindih yang sangat rendah. Kondisi tersebut mengindikasikan bahwa karakteristik masing-masing klaster memiliki perbedaan yang nyata sehingga proses pengelompokan mampu mengidentifikasi pola kualitas udara yang terbentuk selama periode pengamatan.

Kualitas hasil pengelompokan ditunjukkan oleh nilai *silhouette score* sebesar 0.8683. Nilai tersebut menunjukkan bahwa anggota dalam setiap klaster memiliki tingkat kemiripan yang tinggi terhadap anggota klaster yang sama serta memiliki perbedaan yang jelas terhadap anggota pada klaster lainnya. Dengan demikian, struktur klaster yang dihasilkan memiliki tingkat kohesi (*cohesion*) dan separasi (*separation*) yang sangat baik sehingga mampu merepresentasikan pola data secara optimal.

Terbentuknya dua klaster utama menunjukkan bahwa data kualitas udara pada wilayah penelitian memiliki pola yang terstruktur dan dapat dibedakan berdasarkan karakteristik masing-masing kelompok. Perbedaan karakteristik antara klaster normal dan klaster ekstrem mengindikasikan bahwa variasi kualitas udara tidak terjadi secara acak, melainkan berkaitan dengan kondisi atmosfer yang berbeda selama periode pengamatan. Kondisi tersebut menunjukkan adanya hubungan antara faktor meteorologi dan konsentrasi polutan yang membentuk pola pengelompokan data.

Hasil pengelompokan menunjukkan bahwa pola distribusi polusi udara di wilayah penelitian dipengaruhi oleh interaksi faktor meteorologi yang terdiri atas suhu udara (*temp_avg*), kelembapan udara (*humidity_avg*), curah hujan (*rainfall*),

penyinaran matahari (*sunlight*), dan kecepatan angin (*wind_avg*). Faktor-faktor tersebut berperan dalam mengendalikan proses penyebaran, pengendapan, dan akumulasi polutan di atmosfer sehingga menghasilkan kondisi kualitas udara yang berbeda. Kondisi meteorologi yang mendukung akumulasi polutan cenderung membentuk kelompok kondisi ekstrem, sedangkan kondisi yang mendukung dispersi dan deposisi polutan cenderung membentuk kelompok kondisi normal. Temuan ini menunjukkan bahwa pola distribusi spasial-temporal polusi udara di wilayah urban tidak terbentuk secara acak, melainkan dipengaruhi oleh karakteristik meteorologi yang menentukan dinamika konsentrasi polutan selama periode pengamatan.

5.2 Pengaruh Kondisi Cuaca terhadap Kejadian Polusi Ekstrem

Pemisahan yang jelas antara kelompok kondisi normal dan kondisi ekstrem pada ruang komponen utama menunjukkan adanya perbedaan karakteristik antara kedua kelompok yang terbentuk. Penumpukan titik data pada area ekstrem ($PC1 \geq 5.0$) merupakan hasil interaksi antara konsentrasi polutan $PM_{2.5}$ dengan variabel meteorologi yang meliputi suhu udara (*temp_avg*), kelembapan udara (*humidity_avg*), curah hujan (*rainfall*), penyinaran matahari (*sunlight*), dan kecepatan angin (*wind_avg*). Secara fisik, kecepatan angin yang rendah menyebabkan berkurangnya kemampuan atmosfer dalam mendispersikan polutan secara horizontal sehingga partikel $PM_{2.5}$ cenderung terakumulasi di dekat permukaan. Kondisi tersebut dapat diperparah oleh kelembapan udara yang tinggi, karena partikel-partikel polutan mampu menyerap uap air dan mengalami

pertumbuhan higroskopis yang menyebabkan konsentrasi partikulat di lapisan bawah atmosfer semakin meningkat.

Peningkatan konsentrasi polusi pada kondisi ekstrem juga dipengaruhi oleh intensitas penyinaran matahari (*sunlight*) dan suhu udara rata-rata (*temp_avg*). Radiasi matahari yang tinggi dapat memicu berbagai reaksi fotokimia di atmosfer yang berkontribusi terhadap pembentukan partikulat sekunder, terutama ketika sirkulasi udara terhambat akibat kondisi angin yang tenang. Kondisi tersebut menjadi semakin kompleks apabila curah hujan (*rainfall*) berada pada nilai yang rendah atau bahkan nol, karena tidak terjadi mekanisme pembersihan atmosfer secara alami melalui deposisi basah (*wet deposition*). Akibatnya, partikel PM_{2.5} cenderung bertahan lebih lama di atmosfer sehingga meningkatkan potensi terjadinya kondisi polusi udara ekstrem.

Pemisahan yang jelas antara kelompok kondisi normal dan kondisi ekstrem pada hasil pengelompokan menunjukkan bahwa faktor-faktor meteorologi bekerja secara simultan dalam memengaruhi karakteristik kualitas udara. Interaksi antara kecepatan angin yang rendah, kelembapan udara yang tinggi, intensitas penyinaran matahari yang besar, serta minimnya curah hujan menciptakan kondisi atmosfer yang mendukung akumulasi polutan PM_{2.5} sehingga menghasilkan karakteristik yang berbeda dengan kondisi normal. Sebaliknya, kondisi meteorologi yang mendukung proses dispersi dan deposisi polutan cenderung menghasilkan kualitas udara yang relatif lebih baik.

Dengan demikian, hasil penelitian menunjukkan bahwa variasi polusi udara di Kota Malang dipengaruhi oleh interaksi berbagai faktor meteorologi, yaitu suhu udara (*temp_avg*), kelembapan udara (*humidity_avg*), curah hujan (*rainfall*), penyinaran matahari (*sunlight*), dan kecepatan angin (*wind_avg*). Faktor-faktor tersebut memengaruhi proses akumulasi, dispersi, dan deposisi polutan PM_{2.5} di atmosfer sehingga menghasilkan kondisi kualitas udara yang berbeda, mulai dari kondisi normal hingga kondisi ekstrem. Kecepatan angin yang rendah, kelembapan udara yang tinggi, intensitas penyinaran matahari yang besar, serta minimnya curah hujan cenderung meningkatkan akumulasi polutan dan memicu terjadinya kondisi polusi udara ekstrem. Sebaliknya, kondisi meteorologi yang mendukung proses penyebaran dan pembersihan alami polutan cenderung menghasilkan kualitas udara yang relatif normal. Temuan ini menunjukkan bahwa dinamika meteorologi memiliki peran penting dalam menentukan variasi polusi udara dari waktu ke waktu, sehingga perlu dipertimbangkan dalam upaya pemantauan dan pengendalian kualitas udara di Kota Malang.

5.3 Ketepatan Garis Batas (*Threshold*) dalam Pemberian Label

Penentuan ambang batas (*threshold*) pada nilai PC1 = 5.0 digunakan sebagai dasar untuk memisahkan data ke dalam dua kategori utama, yaitu kondisi normal dan kondisi ekstrem. Penetapan batas tersebut didasarkan pada hasil transformasi *Principal Component Analysis* (PCA) yang menunjukkan adanya pemisahan yang jelas antara kelompok data pada ruang komponen utama. Sebagaimana ditunjukkan pada Gambar 4.3, sebagian besar data terkonsentrasi pada sisi kiri nilai PC1 = 5.0,

sedangkan kelompok data lainnya berada pada sisi kanan dengan karakteristik yang berbeda. Kondisi tersebut menunjukkan bahwa nilai PC1 mampu merepresentasikan perbedaan utama yang terbentuk dari kombinasi variabel PM_{2.5} dan faktor meteorologi.

Pemilihan nilai *threshold* sebesar 5.0 dilakukan karena nilai tersebut berada pada area pemisahan (*gap*) antar kelompok data pada sumbu PC1. Posisi tersebut menunjukkan minimnya tumpang tindih antar klaster sehingga dapat digunakan sebagai batas yang representatif dalam membedakan kondisi normal dan kondisi ekstrem. Karena PC1 merupakan komponen utama yang menjelaskan proporsi variansi terbesar dalam data, penggunaan *threshold* pada sumbu tersebut memberikan dasar yang objektif dan sesuai dengan struktur data hasil PCA-*Clustering*.

Hasil penerapan *threshold* menunjukkan bahwa sebagian besar data teridentifikasi sebagai kondisi normal, sedangkan sebagian lainnya tergolong sebagai kondisi ekstrem. Distribusi tersebut konsisten dengan hasil pengelompokan yang diperoleh dan ditunjukkan pada Tabel 4.8. Kesesuaian antara hasil pelabelan dan distribusi klaster menunjukkan bahwa *threshold* yang digunakan mampu merepresentasikan karakteristik masing-masing kelompok secara baik. Dengan demikian, proses pelabelan tidak dilakukan secara subjektif, melainkan berdasarkan pola yang terbentuk dari hasil analisis data.

Selain itu, kualitas struktur klaster yang terbentuk didukung oleh nilai *silhouette score* sebesar 0.8683 sebagaimana ditunjukkan pada Tabel 4.3. Nilai

tersebut termasuk dalam kategori sangat baik karena mendekati angka 1, yang menunjukkan tingkat kohesi (*cohesion*) yang tinggi di dalam klaster serta tingkat separasi (*separation*) yang kuat antar klaster. Hal ini mengindikasikan bahwa anggota pada masing-masing klaster memiliki karakteristik yang relatif homogen dan berbeda secara jelas dari anggota pada klaster lainnya. Oleh karena itu, batas pemisah pada nilai $PC1 = 5.0$ memiliki dasar yang kuat karena didukung oleh kualitas pengelompokan yang tinggi.

Berdasarkan hasil tersebut, threshold pada nilai $PC1 = 5.0$ dapat dianggap tepat untuk digunakan dalam proses pemberian label kondisi kualitas udara. Batas tersebut mampu memisahkan kondisi normal dan kondisi ekstrem sesuai dengan struktur klaster yang terbentuk melalui metode *PCA-Clustering*. Dengan adanya kesesuaian antara distribusi data pada ruang komponen utama, hasil pengelompokan, dan nilai evaluasi klaster, threshold yang digunakan terbukti mampu merepresentasikan perbedaan karakteristik kualitas udara secara objektif dan konsisten.

5.4 Implikasi Hasil Pengelompokan terhadap Kualitas Udara

Hasil pengelompokan menggunakan metode *PCA-Clustering* menunjukkan bahwa data kualitas udara dan faktor meteorologi pada wilayah penelitian terbagi ke dalam dua klaster utama, yaitu klaster kondisi normal yang terdiri atas 357 data dan klaster kondisi ekstrem yang terdiri atas 9 data. Pemisahan kedua klaster tersebut didukung oleh nilai *silhouette score* sebesar 0.8683 yang mengindikasikan bahwa struktur klaster yang terbentuk memiliki tingkat kohesi dan separasi yang

baik. Temuan ini menunjukkan bahwa kondisi kualitas udara selama periode pengamatan dapat dibedakan berdasarkan pola yang terbentuk dari kombinasi konsentrasi $PM_{2.5}$ dan faktor meteorologi yang diamati.

Terbentuknya dua klaster utama menunjukkan adanya perbedaan kondisi kualitas udara yang berkaitan dengan karakteristik meteorologi selama periode pengamatan. Kondisi tersebut mengindikasikan bahwa analisis kualitas udara tidak hanya dapat dilakukan berdasarkan konsentrasi polutan, tetapi juga perlu mempertimbangkan faktor-faktor meteorologi yang berperan dalam proses dispersi, deposisi, dan akumulasi polutan di atmosfer. Dengan demikian, hasil segmentasi yang diperoleh memberikan informasi yang lebih komprehensif mengenai karakteristik kualitas udara berdasarkan interaksi antara polutan dan dinamika meteorologi.

Karakteristik yang dimiliki oleh masing-masing klaster dapat dimanfaatkan untuk mendukung kegiatan pemantauan kualitas udara. Keberadaan kelompok kondisi ekstrem memberikan gambaran mengenai kondisi yang berkaitan dengan peningkatan konsentrasi $PM_{2.5}$ sehingga dapat menjadi dasar dalam mengidentifikasi kondisi kualitas udara yang memerlukan perhatian lebih lanjut. Informasi tersebut dapat mendukung proses pemantauan berdasarkan pola kondisi yang teridentifikasi dari data, sehingga pemantauan kualitas udara dapat dilakukan secara lebih adaptif terhadap perubahan kondisi meteorologi.

Hasil penelitian ini juga memiliki implikasi terhadap upaya pengendalian kualitas udara dan perlindungan kesehatan masyarakat. Karakteristik kondisi yang

tergolong ke dalam kelompok ekstrem dapat dimanfaatkan oleh pemerintah daerah dan instansi terkait sebagai dasar dalam pengembangan sistem peringatan dini (*early warning system*) kualitas udara berbasis faktor meteorologi. Sistem tersebut dapat digunakan untuk meningkatkan kewaspadaan masyarakat ketika kondisi atmosfer menunjukkan potensi peningkatan konsentrasi $PM_{2.5}$.

Selain itu, hasil segmentasi yang diperoleh dapat digunakan sebagai bahan pendukung dalam penyusunan strategi pengendalian kualitas udara yang lebih efektif dan berkelanjutan. Dinas Lingkungan Hidup dapat meningkatkan intensitas pemantauan kualitas udara pada kondisi yang ditandai oleh kecepatan angin yang rendah, kelembapan udara yang tinggi, intensitas penyinaran matahari yang besar, serta curah hujan yang rendah, karena kondisi tersebut cenderung mendukung akumulasi polutan di atmosfer. Di sisi lain, Dinas Kesehatan dapat memanfaatkan informasi tersebut dalam penyusunan program edukasi dan perlindungan bagi kelompok masyarakat rentan, seperti anak-anak, lansia, dan penderita penyakit pernapasan, guna meminimalkan dampak kesehatan akibat paparan polusi udara.

Dengan demikian, hasil pengelompokan yang diperoleh melalui metode *PCA-Clustering* tidak hanya memberikan gambaran mengenai kondisi kualitas udara selama periode pengamatan, tetapi juga menghasilkan rekomendasi strategis berbasis data yang dapat dimanfaatkan oleh berbagai pemangku kepentingan dalam mendukung kegiatan pemantauan kualitas udara, pengelolaan lingkungan, serta upaya perlindungan kesehatan masyarakat secara lebih efektif dan berkelanjutan.

5.5 Perspektif Islam Terhadap Hasil Pengelompokan Polusi Udara

Hasil pemodelan *PCA-Clustering* yang membagi data polusi udara dan faktor meteorologi Kota Malang ke dalam dua kelompok utama, yaitu kondisi normal harian yang mencakup 357 data dan kondisi ekstrem yang mencakup 9 data, tidak hanya memberikan gambaran statistik mengenai karakteristik kualitas udara, tetapi juga menunjukkan adanya keteraturan pola yang terbentuk dari interaksi antara konsentrasi $PM_{2.5}$ dan faktor meteorologi. Nilai *silhouette score* sebesar 0.8683 menunjukkan bahwa struktur klaster yang terbentuk memiliki tingkat kohesi dan separasi yang baik, sehingga variasi kualitas udara yang terjadi tidak berlangsung secara acak, melainkan mengikuti pola tertentu yang dapat dikenali secara ilmiah. Dalam perspektif sains Islam, keteraturan tersebut dapat dipandang selaras dengan konsep *sunnatullah*, yaitu hukum-hukum alam yang diciptakan Allah SWT dan berlangsung sesuai dengan ketetapan-Nya. Dengan demikian, hasil pengelompokan yang diperoleh tidak hanya memiliki makna ilmiah, tetapi juga dapat ditinjau melalui tiga hubungan fundamental dalam islam, yaitu *Hablum Min Allah*, *Hablum Min An-nas*, dan *Hablum Min Al-alam* yang saling terintegrasi dalam mewujudkan kemaslahatan dan menjaga keberlangsungan kehidupan.

Hubungan antara hasil pengelompokan dan nilai-nilai Islam dapat ditinjau melalui konsep *Hablum Min Allah*, yaitu hubungan manusia dengan Allah SWT. Keteraturan pola yang terbentuk melalui metode *PCA-Clustering* menunjukkan bahwa variasi kualitas udara tidak terjadi secara acak, melainkan mengikuti mekanisme tertentu yang dapat dikenali melalui pendekatan ilmiah. Dalam

perspektif Islam, keteraturan tersebut dapat dipandang selaras dengan konsep Sunnatullah, yaitu hukum-hukum alam yang berlangsung sesuai dengan ketetapan Allah SWT. Allah SWT berfirman dalam Surah Al-Qamar ayat 49:

إِنَّا كُلَّ شَيْءٍ خَلَقْنَاهُ بِقَدَرٍ ﴿٤٩﴾

Artinya: “*Sesungguhnya Kami menciptakan segala sesuatu menurut ukuran.*”

Menurut Al-Razi, (1981) dalam *Tafsir Al-Kabir (Mafatih al-Ghaib)*, frasa *bi qadar* menunjukkan bahwa setiap unsur di alam semesta diciptakan dengan ukuran, batasan, dan ketetapan tertentu yang sangat presisi. Selaras dengan hal tersebut, Ibn Kathir, (1999) dalam *Tafsir al-Qur'an al-'Azhim* menjelaskan bahwa seluruh makhluk telah ditetapkan menurut ukuran tertentu yang menjadi dasar keteraturan alam. Dalam konteks penelitian ini, kemampuan metode PCA-*Clustering* dalam mengidentifikasi pola kualitas udara berdasarkan interaksi antara konsentrasi PM_{2.5} dan faktor meteorologi memberikan gambaran bahwa fenomena atmosfer berlangsung mengikuti keteraturan tertentu yang dapat dipahami melalui ilmu pengetahuan. Oleh karena itu, hasil segmentasi yang diperoleh dapat dipandang selaras dengan konsep penciptaan segala sesuatu menurut ukuran (*bi qadar*) dan menjadi sarana untuk memahami tanda-tanda kebesaran Allah SWT yang termanifestasi melalui keteraturan fenomena alam.

Selain berkaitan dengan hubungan manusia dengan Allah SWT, hasil pengelompokan yang diperoleh juga memiliki implikasi sosial yang berkaitan dengan konsep *Hablum Min An-nas*. Identifikasi terhadap kondisi polusi udara ekstrem tidak hanya memiliki makna statistik, tetapi juga menunjukkan adanya

kondisi yang berpotensi meningkatkan risiko gangguan kesehatan masyarakat serta menurunkan kualitas hidup. Informasi mengenai karakteristik kondisi ekstrem tersebut dapat dimanfaatkan sebagai dasar dalam pengembangan sistem peringatan dini, penyusunan strategi mitigasi, serta perlindungan terhadap kelompok masyarakat yang rentan.

Landasan kepedulian terhadap sesama manusia ditegaskan oleh Allah SWT dalam Surah Al-Hujurat ayat 10:

إِنَّمَا الْمُؤْمِنُونَ إِخْوَةٌ فَأَصْلَحُوا بَيْنَ أَخَوَيْكُمْ وَاتَّقُوا اللَّهَ لَعَلَّكُمْ تُرْحَمُونَ ﴿١٠﴾

Artinya: “*Sesungguhnya orang-orang mukmin itu bersaudara, karena itu damaikanlah antara kedua saudaramu dan bertakwalah kepada Allah agar kamu mendapat rahmat.*”

Imam Al-Qurtubi dalam *Al-Jami' li Ahkam al-Qur'an* menjelaskan bahwa konsep ishlah pada hakikatnya merupakan upaya mewujudkan kemaslahatan serta mencegah berbagai bentuk kerusakan yang membahayakan manusia (Al-Qurtubi, 2006). Sejalan dengan hal tersebut, Ibn Kathir, (1999) menjelaskan bahwa ukhuwah antarsesama mukmin mengandung kewajiban untuk saling melindungi dari berbagai bentuk kemudharatan. Dalam konteks penelitian ini, kemampuan PCA-*Clustering* dalam mengidentifikasi kondisi polusi ekstrem memiliki nilai kemaslahatan sosial karena informasi yang dihasilkan dapat dimanfaatkan dalam pengembangan sistem peringatan dini, peningkatan kewaspadaan masyarakat, serta perlindungan kelompok rentan. Dengan demikian, hasil segmentasi yang diperoleh memiliki relevansi dengan prinsip *Hifdzun Nafs* dalam maqashid syariah, yaitu menjaga keselamatan dan keberlangsungan hidup manusia. Hubungan ini

menunjukkan bahwa pemanfaatan ilmu pengetahuan tidak hanya bertujuan menghasilkan informasi, tetapi juga mendukung terwujudnya kemaslahatan bagi masyarakat.

Perspektif Islam terhadap hasil pengelompokan tersebut juga dapat dipahami melalui konsep *Hablum Min Al-alam*, yaitu hubungan manusia dengan lingkungan dan alam semesta. Keberadaan kondisi normal dan kondisi ekstrem memberikan gambaran bahwa kualitas lingkungan dipengaruhi oleh keseimbangan berbagai komponen yang menyusun sistem atmosfer. Kondisi meteorologi yang mendukung proses dispersi polutan cenderung menghasilkan kualitas udara yang lebih baik, sedangkan kondisi yang kurang mendukung proses tersebut dapat meningkatkan akumulasi polutan di atmosfer. Hal ini menunjukkan bahwa keseimbangan lingkungan merupakan faktor penting dalam menjaga keberlangsungan kehidupan.

Prinsip menjaga keseimbangan lingkungan tersebut sejalan dengan firman Allah SWT dalam Surah Al-A'raf ayat 56:

وَلَا تُفْسِدُوا فِي الْأَرْضِ بَعْدَ إِصْلَاحِهَا وَادْعُوهُ خَوْفًا وَطَمَعًا إِنَّ رَحْمَتَ اللَّهِ قَرِيبٌ مِّنَ الْمُحْسِنِينَ ﴿٥٦﴾

Artinya: “Dan janganlah kamu membuat kerusakan di muka bumi setelah Allah memperbaikinya.”

Menurut Al-Zamakhshari, (1998) dalam *Al-Kasysyaf, frasa ba'da ishlahiha* menunjukkan bahwa Allah SWT telah menciptakan bumi beserta sistemnya dalam keadaan yang teratur dan seimbang sehingga mampu menopang kehidupan makhluk-Nya. Didukung oleh pandangan Al-Mawardi, (1992) *An-Nukat wa al-*

'*Uyun*, larangan melakukan kerusakan mencakup segala bentuk tindakan yang dapat mengganggu kemaslahatan dan keseimbangan alam semesta. Dalam kaitannya dengan penelitian ini, hasil pengelompokan menunjukkan pentingnya menjaga keseimbangan lingkungan agar kemampuan atmosfer dalam mendispersikan polutan tetap terjaga. Oleh karena itu, hasil penelitian ini memperlihatkan pentingnya menjaga hubungan manusia dengan lingkungan sebagai bentuk implementasi *Hablum Min al-Alam* serta tanggung jawab manusia sebagai khalifah di muka bumi.

Dengan demikian, hasil pengelompokan berbasis *PCA-Clustering* dalam penelitian ini tidak hanya memberikan pemahaman mengenai karakteristik kualitas udara berdasarkan interaksi antara polutan dan faktor meteorologi, tetapi juga dapat dipandang selaras dengan nilai-nilai Islam. Melalui *Hablum Min Allah*, keteraturan pola yang terbentuk merefleksikan adanya *sunnatullah* yang mengatur fenomena alam. Melalui *Hablum Minannas*, identifikasi kondisi polusi ekstrem memberikan landasan bagi upaya perlindungan kesehatan masyarakat sebagai implementasi prinsip *Hifdzun Nafs*. Sementara itu, melalui *Hablum Minal 'Alam*, hasil penelitian ini menegaskan pentingnya menjaga keseimbangan lingkungan dan mencegah berbagai bentuk kerusakan yang dapat mengganggu keberlangsungan kehidupan. Ketiga hubungan tersebut tidak berdiri secara terpisah, melainkan saling melengkapi dalam mewujudkan kemaslahatan manusia dan keberlanjutan lingkungan sesuai dengan nilai-nilai yang diajarkan dalam Islam.

BAB VI

KESIMPULAN DAN SARAN

6.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Faktor meteorologi memiliki keterkaitan dengan variasi kondisi polusi udara di Kota Malang. Hasil *Principal Component Analysis* (PCA) menunjukkan bahwa dua komponen utama pertama mampu menjelaskan 67.15% variasi total data. Hasil tersebut menunjukkan bahwa parameter meteorologi berupa suhu udara, kelembapan udara, curah hujan, penyinaran matahari, dan kecepatan angin berkontribusi dalam membentuk pola variasi kondisi polusi udara.
2. Metode *PCA-Clustering* efektif dalam mengidentifikasi kluster polusi udara berdasarkan interaksi antara faktor meteorologi dan kualitas udara. Hasil pengelompokan menghasilkan dua kluster utama, yaitu kluster kondisi normal sebanyak 357 data dan kluster kondisi ekstrem sebanyak 9 data. Validasi menggunakan *silhouette score* sebesar 0.87 menunjukkan bahwa kluster yang terbentuk memiliki tingkat kekompakan dan keterpisahan yang sangat baik. Selain itu, hasil perbandingan dengan metode K-Means yang menghasilkan *silhouette score* sebesar 0.47 menunjukkan bahwa *PCA-Clustering* memberikan kualitas pengelompokan yang lebih baik.
3. Hasil segmentasi polusi udara memberikan implikasi terhadap pengelolaan kualitas udara dan perlindungan kesehatan masyarakat. Identifikasi kluster kondisi normal dan kondisi ekstrem dapat dimanfaatkan sebagai informasi

pendukung dalam kegiatan pemantauan kualitas udara, pengembangan sistem peringatan dini, serta penyusunan kebijakan pengendalian pencemaran udara yang lebih berbasis data. Selain itu, hasil pengelompokan dapat membantu proses identifikasi kondisi meteorologi yang berasosiasi dengan peningkatan risiko paparan polusi udara pada masyarakat.

6.2 Saran

Berdasarkan keterbatasan penelitian yang telah dilakukan, beberapa saran yang dapat menjadi acuan bagi penelitian selanjutnya adalah sebagai berikut:

1. Menambahkan parameter kualitas udara lainnya, seperti NO₂, SO₂, dan CO, serta mempertimbangkan faktor eksternal seperti volume lalu lintas harian untuk memperoleh gambaran kondisi kualitas udara yang lebih komprehensif.
2. Menggunakan data dengan periode pengamatan yang lebih panjang agar pola variasi kualitas udara dan faktor meteorologi yang diperoleh dapat menggambarkan kondisi yang lebih representatif.
3. Mengembangkan dataset yang telah dilabeli melalui metode PCA-*Clustering* sebagai data latih (*training data*) dalam membangun model pembelajaran terawasi (*supervised learning*), seperti *Long Short-Term Memory* (LSTM), untuk mendukung prediksi kualitas udara pada periode mendatang.

DAFTAR PUSTAKA

- Al-Mawardi, A. al-H. (1992). *An-Nukat wa al-'Uyun (Tafsir Al-Mawardi)*. Dar al-Kutub al-'Ilmiyyah.
<https://shamela.ws/book/23602>](<https://shamela.ws/book/23602>
- Al-Qurtubi, A. 'Abdullah. (2006). *Al-Jami' li Ahkam al-Qur'an*. Muassasatur Risalah. <https://shamela.ws/book/23586>](<https://shamela.ws/book/23586>
- Al-Razi, F. al-D. (1981). *Tafsir Al-Kabir (Mafatih al-Ghaib)*. Dar al-Fikr.
<https://shamela.ws/book/23615>](<https://shamela.ws/book/23615>
- Al-Zamakhshari, A. al-Q. (1998). *Al-Kasysyaf 'an Haqa'iq Ghawamidh al-Tanzil*. Dar al-Kutub al-'Ilmiyyah.
<https://shamela.ws/book/23585>](<https://shamela.ws/book/23585>
- Almais, A. T. W., Susilo, A., Naba, A., Sarosa, M., Crysdian, C., Tazi, I., Hariyadi, M. A., Muslim, M. A., Basid, P. M. N. S. A., Arif, Y. M., Purwanto, M. S., Parwatiningsy, D., Supriyono, & Wicaksono, H. (2023). Principal Component Analysis-Based Data Clustering for Labeling of Level Damage Sector in Post-Natural Disasters. *IEEE Access*, 11(May), 74590–74601. <https://doi.org/10.1109/ACCESS.2023.3275852>
- Badan Pusat Statistik Kota Malang. (2023). *Pengamatan Kelembaban Udara di Kota Malang (Persen (%))*, 2023. BPS Kota Malang.
<https://malangkota.bps.go.id>
- Friedman, H. & T. (2009). Elements of Statistical Learning - Data Mining, Interface, and Prediction. In *Springer* (Vol. 2nd).
- Huang, Y., Cai, Y., Jiao, J., Pan, C., Wang, G., Li, C., Jia, Z., Chen, Z., Zhou, Y., & Zhou, G. (2025). The Impact of Meteorological Factors and Canopy Structure on PM2.5 Dynamics Under Different Urban Functional Zones in a Subtropical City. *Forests*, 16(3). <https://doi.org/10.3390/f16030479>
- Ibn Kathir, I. bin U. (1999). *Tafsir al-Qur'an al-'Azhim*. Dar Thayyibah.
<https://shamela.ws/book/23578>](<https://shamela.ws/book/23578>
- Indonesia, K. A. R. (2019). *Al-Qur'an dan Terjemahannya (Edisi Penyempurnaan 2019)* (Edisi Peny). Kementerian Agama Republik Indonesia.
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).
<https://doi.org/10.1098/rsta.2015.0202>
- Kalajdzieski, J., Trivodaliev, K., Mirceva, G., Kalajdziski, S., & Gievska, S. (2023). A Complete Air Pollution Monitoring and Prediction Framework. *IEEE Access*, 11(January), 88730–88744.
<https://doi.org/10.1109/ACCESS.2023.3251346>

- Liu, W., Chen, F., & Chen, Y. (2022). PM2.5 Concentration Prediction Based on Pollutant Pattern Recognition Using PCA-clustering Method and CS Algorithm Optimized SVR. *Nature Environment and Pollution Technology*, 21(1), 393–403. <https://doi.org/10.46488/NEPT.2022.v21i01.047>
- Montazeri, A., Lilienthal, A. J., & Albertson, J. D. (2021). A spatial land use clustering framework for investigating the role of land use in mediating the effect of meteorology on urban air quality. *Atmospheric Environment: X*, 12(August), 100126. <https://doi.org/10.1016/j.aeaoa.2021.100126>
- Nor Asri, L., Eka Sari, K., & Christia Meidiana. (2022). Emisi CO Kendaraan Bermotor Pada Ruas Jalan Dengan Tingkat Pelayanan Rendah Di Kota Malang. *Planning for Urban Region and Environment*, 11(1), 31–38.
- Rivera-Campoverde, N. D., Muñoz-Sanz, J. L., & Arenas-Ramirez, B. D. V. (2021). Estimation of pollutant emissions in real driving conditions based on data from OBD and machine learning. *Sensors*, 21(19), 1–16. <https://doi.org/10.3390/s21196344>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C), 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Roy, M., Balachandran, S., & Brokamp, C. (2022). Clustering and Regression-Based Analysis of PM2.5 Sensitivity to Meteorology in Cincinnati, Ohio. *Atmosphere*, 13(4), 1–20. <https://doi.org/10.3390/atmos13040545>
- Shafi, J., & Waheed, A. (2020). K-Means Clustering Analysing Abrupt Changes in Air Quality. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 26–30. <https://doi.org/10.1109/ICECA49313.2020.9297493>
- Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson. Pearson.
- Wang, J., Geng, Y., Zhao, Q., Zhang, Y., Miao, Y., Yuan, X., Jin, Y., & Zhang, W. (2021). Water Quality Prediction of Water Sources Based on Meteorological Factors using the CA-NARX Approach. *Environmental Modeling and Assessment*, 26(4), 529–541. <https://doi.org/10.1007/s10666-021-09759-5>
- World Health Organization. (2021). *Ambient (outdoor) air pollution*. World Health Organization.
- World Health Organization. (2023). *World Health Statistics 2023*. World Health Organization. <https://www.who.int/publications/b/69040>
- Xu, H., Hu, P., Wang, H., Croot, P., Li, Z., Li, C., Xie, S., Zhou, H., & Zhang, C. (2025). Identification of the pollution sources and hidden clustering patterns for potentially toxic elements in typical peri-urban agricultural soils in southern China. *Environmental Pollution*, 370(January), 125904.

<https://doi.org/10.1016/j.envpol.2025.125904>

Zareba, M., Dlugosz, H., Danek, T., & Weglinska, E. (2023). Big-Data-Driven Machine Learning for Enhancing Spatiotemporal Air Pollution Pattern Analysis. *Atmosphere*, 14(4). <https://doi.org/10.3390/atmos14040760>