CLUSTERING CURAH HUJAN DENGAN *PRINCIPAL COMPONENT ANALYSIS* MENGGUNAKAN DATA KLIMATOLOGI

TESIS

Oleh:
SYAHRENI
NIM.240605210002

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS ILMU SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

CLUSTERING CURAH HUJAN DENGAN *PRINCIPAL COMPONENT ANALYSIS* MENGGUNAKAN DATA KLIMATOLOGI

TESIS

Diajukan kepada:

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

Oleh:
SYAHRENI
NIM. 240605210002

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

CLUSTERING CURAH HUJAN DENGAN *PRINCIPAL COMPONENT ANALYSIS* MENGGUNAKAN DATA KLIMATOLOGI

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:
SYAHRENI
NIM. 240605210002

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

***CLUSTERING* CURAH HUJAN DENGAN *PRINCIPAL COMPONENT ANALYSIS* MENGGUNAKAN DATA KLIMATOLOGI**

TESIS

Oleh :
SYAHRENI
NIM. 240605210002

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 12 November 2025

Pembimbing I,	Pembimbing II,
Dr. Agung Teguh Wibowo Almais, S.kom., M.T	Dr. Ir. Fresy Nugroho, S.T., M.T., I.P.M., ASEAN Eng
NIPPPK. 198603012023211016	NIP. 19710722 201101 1 001

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* CURAH HUJAN DENGAN *PRINCIPAL COMPONENT ANALYSIS* MENGGUNAKAN DATA KLIMATOLOGI**

TESIS

Oleh :
SYAHRENI
NIM. 240605210002

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 12 November 2025

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Prof. Sri Harini., M.Si NIP. 19731014 200112 2 0002	( )
Penguji II	:	Dr. Cahyo Crydian, M.Cs NIP. 19740424 2000901 1 008	( )
Pembimbing I	:	Dr. Agung Teguh Wibowo Almais, S.kom., M.T NIPPPK. 198603012023211016	( )
Pembimbing II	:	Dr. Ir. Fresy Nugroho, S.T., M.T., I.P.M., ASEAN Eng NIP. 19710722 201101 1 001	( )

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

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama : Syahreni
NIM : 240605210002
Program Studi : Magister Informatika
Fakultas : Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Thesis 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 Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 01 Desember 2025
Yang Membuat Pernyataan,

Syahreni
240605210002

MOTTO

"Ciptakan kesempatanmu sendiri."

HALAMAN PERSEMBAHAN

Kupersembahkan karyaku ini:

1. Kepada diri sendiri, yang telah berjuang tanpa henti hingga sampai pada titik ini, meskipun harus berulang kali keluar masuk rumah sakit selama proses penyusunan tesis berlangsung. Segala ujian tersebut menjadi bagian dari perjalanan berharga yang semakin menguatkan tekad untuk menyelesaikan studi ini dengan sebaik-baiknya.
2. Kedua orang tua tercinta, H. Syamsuddin, S.K.M., dan Hj. Refmawati, atas segala doa, kasih sayang, serta dukungan moral dan spiritual yang tiada henti diberikan selama proses pendidikan hingga terselesaikannya tesis ini.
3. Saudara dan keluarga besar yang selalu menjadi roda-roda ketika aku jatuh dan memberi semangat yang besar.
4. Segenap sivitas akademika Fakultas Sains dan Teknologi Universitas Ibrahimy.
5. Teman - Teman Magister Inforamatika semua yang angkatan X khusunya, mereka teman perjuanganku.

KATA PENGANTAR

Assalamu'alaikum Wr. Wb.

Syukur Alhamdulillah penulis panjatkan kehadirat Allah SWT atas limpahan rahmat, taufik, dan hidayah-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. Tesis ini merupakan salah satu syarat dalam menyelesaikan studi di Program Magister Informatika, dengan penuh perjuangan, doa, dan dukungan dari berbagai pihak. Dengan segala keterbatasan dan tantangan, penulis bersyukur dapat melalui proses ini hingga tahap akhir sidang tesis.

Pada kesempatan ini, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Allah SWT, atas segala rahmat, karunia, dan kekuatan yang telah diberikan, sehingga penulis dapat menyelesaikan penelitian dan penulisan tesis ini.
2. Prof. Dr. Hj. Ilfi Nur Diana, M.Pd.I elaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang
3. Dr. H. Agus Mulyono, M.Si 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. Bapak Dr. Agung Teguh Wibowo Almais, M.T., selaku Dosen Pembimbing I, yang telah dengan sabar memberikan bimbingan, arahan, dan motivasi selama proses penelitian dan penyusunan tesis.
6. Bapak Dr. Ir. Fresy Nugroho, S.T., M.T., IPM., selaku Dosen Pembimbing II, yang telah memberikan banyak masukan berharga, pandangan ilmiah, dan bimbingan yang mendalam selama proses penyusunan tesis ini.
7. Prof. Sri Harini., M.Si selaku dosen penguji I ang telah menguji serta memberikan masukan sehingga penulis dapat menuntaskan Thesis dengan baik.

8. Dr. Cahyo Crydian, M.Cs selaku dosen penguji II ang telah menguji serta memberikan masukan sehingga penulis dapat menuntaskan Thesis dengan baik.
9. Bapak Ahmad Zarkoni, S.Kom., rekan seperjuangan tesis sekaligus pembimbing lapangan dari BMKG banyak membantu penulis dalam pengumpulan dan pemahaman data penelitian. dan Saudara Tomy Ivan Sugiharto dan Fanny Brawijaya rekan seperjuangan dalam penyusunan tesis di Program Magister Informatika, yang selalu memberikan semangat, dukungan, dan kebersamaan selama proses penelitian.
10. Seluruh rekan-rekan mahasiswa Magister Informatika UIN Maulana Malik Ibrahim Malang, atas kebersamaan, diskusi, serta dukungan yang turut memberikan semangat selama masa studi.
11. Seluruh pihak dan orang-orang terdekat penulis yang tidak dapat disebutkan satu per satu, namun selalu hadir memberikan doa, perhatian, dan dukungan dalam bentuk apa pun selama proses penyusunan tesis ini.

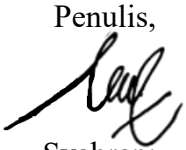
Penulis menyadari bahwa tesis ini masih jauh dari sempurna, baik dari segi isi maupun penyajiannya. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan untuk perbaikan di masa mendatang. Semoga karya sederhana ini dapat memberikan manfaat bagi pembaca, serta menjadi kontribusi ilmiah dalam pengembangan ilmu pengetahuan di bidang informatika dan klimatologi.

Akhir kata, semoga segala bantuan, bimbingan, dan doa yang diberikan kepada penulis mendapat balasan pahala dan keberkahan dari Allah SWT. Amin Ya Rabbal ‘Alamin.

Wassalamu’alaikum Wr. Wb

Malang, 01 Desember 2025

Penulis,



Syahreni

DAFTAR ISI

HALAMAN PENGAJUAN i
HALAMAN PERSETUJUAN iii
HALAMAN PENGESAHAN iv
PERNYATAAN KEASLIAN TULISAN v
MOTTO vi
HALAMAN PERSEMBAHAN vii
KATA PENGANTAR viii
DAFTAR ISI x
DAFTAR TABEL xii
DAFTAR GAMBAR xiii
Abstrak xiv
Abstract ..... **Error! Bookmark not defined.**
مستخلص البحث xvi
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Pernyataan Masalah 6
1.3 Tujuan Penelitian 6
1.4 Hipotesis 7
1.5 Manfaat Penelitian 7
1.6 Batasan Penelitian 7
1.7 Sistematika Penulisan 8
BAB II STUDI PUSTAKA 10
2.1 *Clustering* Rainfall 10
2.2 Curah Hujan 15

2.3 Hari Hujan 16
2.4 Curah Tidak Hujan 16
2.5 Kerangka Teori 17
BAB III METODOLOGI PENELITIAN 20
3.1 Kerangka Konsep 20
3.2 Desain Sistem 22
3.3 Data Preparation 25
3.4 *Principal Component Analysis Clustering* (PCA - *Clustering*) 28
3.5 Skenario Evaluasi 35
BAB IV HASIL DAN PEMBAHASAN 39
4.1 Normalisasi Data 39
4.2 Pemformatan Jumlah (n) Principal Component (PC) dan Penentuan Rasio Varians. 40
4.3 Hasil Pengelompokan Analisis Komponen Utama 43
4.4 Label Berdasarkan Hasil Grafis 44
4.5 Proses *Clustering* 44
4.6 Hasil *Clustering* 46
4.7 Validasi Hasil *Clustering* 49
BAB V KESIMPULAN 57
5.1 Kesimpulan 57
5.2 Saran 58
DAFTAR PUSTAKA 59

DAFTAR TABEL

Tabel 2.1 Peneliti terdahulu 14
Tabel 3.1 Variabel Data klimatologi(Arisandi et al., 2021)..... 25
Tabel 3.2 Data curah hujan 26
Tabel 4.1 Hasil Normalisasi Data 39
Tabel 4.2 Eigen Value And Variance ratio 41
Tabel 4.3 Point and coordinate *value* PC1 dan PC2 45
Tabel 4.4 Hasil klaster..... 48
Tabel 4.5 Hasil Evaluasi Kualitas *Clustering* dengan *Silhouette Score*..... 50
Tabel 4.6 Label Data Hasil Perbandingan PC1 Berdasarkan Target Asli..... 55

DAFTAR GAMBAR

Gambar 2.1 Kerangka Teori 17
Gambar 3.1 Kerangka Konsep 20
Gambar 3.2 Desain Sistem 22
Gambar 4.1. Grafik *Eigenvalue* dan *Varian Ratio* 42
Gambar 4.2. Graph of PC1 and PC2 *Variance ratio*n *Value* 43
Gambar 4.3 Coordinate Points of PC1 and PC2 Data Distribution 45
Gambar 4.4 Titik koordinat PC1 dan PC2 46
Gambar 4.5 Hasil dari *Clustering* data Titik koordinat PC1 dan PC2 47
Gambar 4.6 Hasil 3 Klaster 47
Gambar 4.7 Titik koordinat PC1 dan PC2 51
Gambar 4.8 Hasil dari *Clustering* data Titik koordinat PC1 dan PC2 53
Gambar 4.9 Hasil dari *Clustering* data target 54
Gambar 4.10 Hasil Pengelompokan data PC1 berdasarkan target data asli 55

Abstrak

Syahreni 2025. ***Clustering* Curah Hujan Dengan Principal Component Analysis Menggunakan Data Klimatologi.** Tesis. Program Study Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing (1) Dr. Agung Teguh Wibowo Almais, S.kom., M.T Pembimbing (II) Dr. Ir. Fresy Nugroho, S.T., M.T., I.P.M., ASEAN Eng.

Kata Kunci: PCA, *Clustering*, Curah Hujan, Hari Tidak Hujan, Hari Hujan, silhouette score

Curah hujan merupakan parameter klimatologi yang sangat penting dalam analisis cuaca dan mitigasi bencana hidrometeorologi, terutama di wilayah tropis seperti Indonesia yang memiliki variabilitas curah hujan tinggi. Penelitian ini bertujuan untuk mengidentifikasi pola hari hujan dan tidak hujan menggunakan metode Principal Component Analysis–*Clustering* (PCA–*Clustering*). Data klimatologi diperoleh dari Stasiun Klimatologi Sumberpucung dan Karangploso dengan enam variabel utama, yaitu temperatur rata-rata, curah hujan, crah hujan jam 07.00, penyinaran matahari, penguapan dan tekjanan udara. Proses normalisasi dilakukan menggunakan StandardScaler untuk menyeragamkan skala data sebelum diterapkan PCA. PCA kemudian mereduksi keenam variabel menjadi dua komponen utama (PC1 dan PC2) yang mampu menjelaskan 73,33% variasi data. Berdasarkan rentang nilai PC1, data terbagi menjadi dua kelompok: Cluster 1 ($n < 0$) yang mewakili hari hujan, dan Cluster 2 ($0 \leq n \leq 3$) yang mewakili hari tidak hujan. Hasil pengelompokan ini menunjukkan konsistensi dengan kategori intensitas curah hujan resmi BMKG. Validasi internal menggunakan Silhouette Score menghasilkan nilai 0,55, yang mengindikasikan kualitas pengelompokan yang cukup baik, dengan pemisahan antarcluster yang jelas. Validasi eksternal melalui perbandingan dengan data target BMKG juga menunjukkan kesesuaian pola sebaran cluster. Temuan ini menegaskan bahwa PCA bukan hanya teknik reduksi dimensi, tetapi juga dapat digunakan sebagai dasar pembentukan label dan pengelompokan data curah hujan secara efektif. Penelitian selanjutnya disarankan untuk memperluas cakupan data spasial dan temporal serta mengintegrasikan metode ini dengan model prediktif berbasis machine learning untuk analisis pola curah hujan yang lebih mendalam.

Abstract

Syahreni 2025. ***Rainfall Clustering*** **with Principal Component Analysis Using Climatological Data.** Thesis. Master Program of Computer Science, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Supervisor (1) Dr. Agung Teguh Wibowo Almais, S.kom., M.T Supervisor (II) Dr. Ir. Fresy Nugroho, S.T., M.T., I.P.M., ASEAN Eng.

Keywords: PCA, *Clustering*, Rainfall, No Rainy Day, Rainy Day, silhouette score

Rainfall is a very important climatological parameter in weather analysis and hydrometeorological disaster mitigation, especially in tropical regions such as Indonesia which have high rainfall variability. This study aims to identify the pattern of rainy and non-rainy days using the Principal Component Analysis–*Clustering* (PCA–*Clustering*) method. Climatological data was obtained from the Sumberpucung and Karangploso Climatology Stations with six main variables, namely average temperature, rainfall, rainfall at 07.00, solar irradiation, evaporation and air pressure. The normalization process is carried out using StandardScaler to standardize the data scale before PCA is applied. PCA then reduced the six variables to two main components (PC1 and PC2) which were able to explain 73.33% of the data variation. Based on the PC1 value range, the data is divided into two groups: Cluster 1 ($n < 0$) which represents rainy days, and Cluster 2 ($0 \leq n \leq 3$) which represents non-rainy days. The results of this grouping show consistency with the official BMKG rainfall intensity category. Internal validation using the Silhouette Score yielded a value of 0.55, which indicates a fairly good quality of clustering, with clear separation between clusters. External validation through comparison with BMKG target data also showed the suitability of the cluster distribution pattern. These findings confirm that PCA is not only a dimension reduction technique, but can also be used as a basis for labeling and grouping rainfall data effectively. Further research is suggested to expand the scope of spatial and temporal data and integrate this method with machine learning-based predictive models for more in-depth analysis of rainfall patterns.

مستخلص البحث

شهريني. 2025. تجميع كمية الأمطار باستخدام تحليل المكون الرئيسي باستخدام بيانات المناخ. رسالة الماجستير. قسم المعلومات، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: د. أجونج تيغوه وييووو ألمايس، الماجستير؛ المشرف الثاني: د. فريشي نوجروهو، الماجستير.

الكلمات الرئيسية: تحليل مكون رئيسي، تجميع، كمية أمطار، أيام غير ممطرة، أيام ممطرة، درجة تماثل.

هطول الأمطار هو أحد مؤشرات المناخ المهمة جدًا في تحليل الطقس والحد من مخاطر الكوارث الهيدرومناخية، خاصة في المناطق الاستوائية مثل إندونيسيا التي تتمتع بتغيرات كبيرة في هطول الأمطار. تهدف هذه الرسالة إلى تحديد نمط الأيام الممطرة والأيام غير الممطرة باستخدام طريقة تحليل المكون الرئيسي– التجميع–*PCA*) (*Clustering*. تم الحصول على البيانات المناخية من محطة مناخ سومبير بوجونج وكارانغ بلوسو باستخدام ستة متغيرات رئيسية، وهي متوسط درجة الحرارة، وهطول الأمطار، وهطول الأمطار الساعة 07:00، والإشعاع الشمسي، والتبخر، والضغط الجوي. تم تنفيذ عملية التطبيع باستخدام *StandardScaler* لتوحيد مقياس البيانات قبل تطبيق تحليل المكون الرئيسي. ثم قام PCA بتقليص المتغيرات الستة إلى مكونين رئيسيين (*PC1* و*PC2*) قادرين على شرح 73.33% من تباين البيانات. بناءً على نطاق قيمة *PC1* ، تم تقسيم البيانات إلى مجموعتين: المجموعة 1 ($n < 0$) التي تمثل الأيام الممطارة، والمجموعة 2 ($0 \leq n \leq 3$) التي تمثل الأيام غير الممطرة . أظهرت نتائج التجميع هذه توافقًا مع فئات شدة هطول الأمطار الرسمية التابعة للهيئة العامة للأرصاد الجوية الإندونيسية (BMKG). وأسفرت عملية التحقق الداخلية باستخدام درجة السيلويت (*Silhouette Score*) عن قيمة 0.55، مما أشار إلى جودة تجميع جيدة إلى حد ما، مع فصل واضح بين المجموعات. كما أظهرت عملية التحقق الخارجية من خلال مقارنة البيانات مع بيانات الهدف لـ BMKG توافق أنماط توزيع المجموعات. تؤكد هذه النتائج أن تحليل المكون الرئيسي (*PCA*) ليس مجرد تقنية لتقليل الأبعاد، بل يمكن استخدامه أيضًا كأساس لإنشاء العلامات وتجميع بيانات هطول الأمطار بشكل فعال. يُنصح في الدراسات المستقبلية بتوسيع نطاق البيانات المكانية والزمنية ودمج هذه الطريقة مع نماذج التنبؤ المعتمدة على التعلم الآلي لتحليل أنماط هطول الأمطار بشكل أعمق.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Indonesia merupakan negara beriklim tropis karena terletak di garis khatulistiwa yang membuatnya memiliki variasi curah hujan yang tinggi. Letak geografis ini menyebabkan Indonesia mengalami pola cuaca yang beragam sepanjang tahun, dengan musim hujan dan musim kemarau yang jelas. Curah hujan yang bervariasi ini dipengaruhi oleh berbagai faktor, seperti pergerakan angin monsun dan kondisi lautan yang dinamis Fakta ini menunjukkan bahwa pola curah hujan di Indonesia sangat kompleks dan memiliki variasi yang tinggi, sehingga diperlukan analisis pengelompokan untuk memahami karakteristik pola cuacanya dengan lebih jelas. Dampaknya, Indonesia rentan terhadap berbagai bencana *hidrometeorologi*, seperti banjir, angin kencang, dan tanah longsor yang sering kali terjadi akibat peningkatan atau penurunan curah hujan ekstrem (Rachmawati, 2021). Variabilitas curah hujan di Indonesia, yang sangat dipengaruhi oleh faktor geografis dan cuaca global, menuntut pemahaman yang lebih mendalam mengenai distribusi dan pola curah hujan untuk mengurangi risiko bencana (Pansera et al., 2013).

Wilayah Malang memiliki karakteristik curah hujan yang dipengaruhi oleh iklim tropis dengan pola musim yang jelas, yakni musim hujan yang berlangsung pada periode November hingga April serta musim kemarau pada bulan-bulan berikutnya (Irawan et al., 2022). *Topografi* wilayah yang bervariasi, mulai dari dataran rendah hingga kawasan pegunungan, turut memengaruhi distribusi dan

intensitas curah hujan di daerah ini (Suharyanto Maulana, Suprayogo, Devia, & Kurniawan, 2023). Kondisi tersebut mengindikasikan adanya perbedaan intensitas curah hujan antarwilayah, yang menunjukkan bahwa pola hujan di Indonesia tidak bersifat seragam. Variasi ini terutama terlihat pada daerah pegunungan yang cenderung memiliki curah hujan lebih tinggi dibandingkan kawasan lain, sehingga analisis pengelompokan diperlukan untuk memahami karakteristik pola hujan secara lebih mendalam, seperti banjir dan tanah longsor (Darmawan et al., 2022). Oleh karena itu, analisis pola curah hujan yang mendalam menjadi penting untuk mendukung pengelolaan sumber daya air dan memahami dinamika iklim lokal di wilayah Malang, sehingga informasi klimatologis dapat digunakan secara lebih efektif oleh berbagai sektor (Karamma, 2025).

Perubahan pola curah hujan akibat perubahan iklim global semakin memperburuk situasi, sehingga menimbulkan tantangan besar dalam mitigasi risiko dan pengelolaan sumber daya air. Sebagai contoh, dalam sektor pertanian, ketidakpastian curah hujan dapat menyebabkan gagal panen dan mengancam ketahanan pangan nasional (Hirvonen et al., 2022). Hal ini menunjukkan bahwa ketidakteraturan curah hujan tidak hanya berdampak pada aspek lingkungan, tetapi juga pada aspek sosial dan ekonomi. Oleh karena itu, diperlukan analisis mendalam terhadap variasi curah hujan, baik secara temporal (yang berkaitan dengan perubahan curah hujan dari waktu ke waktu, seperti musiman atau tahunan) maupun spasial (yang berkaitan dengan distribusi curah hujan di berbagai lokasi atau wilayah geografis), guna mendukung perencanaan strategis di sektor pengelolaan sumber daya air dan mitigasi bencana *hidrometeorologi* (Worku et al., 2022).

Beberapa penelitian sebelumnya telah memfokuskan kajiannya pada analisis curah hujan serta pengelompokannya berdasarkan dampaknya terhadap lingkungan, sektor pertanian, dan kehidupan sehari-hari. Namun, permasalahan yang masih tersisa adalah pengembangan lebih lanjut terhadap variasi musiman masih sangat penting karena memiliki pengaruh signifikan terhadap berbagai sektor (Darlan et al., 2020). Seiring dengan kemajuan teknologi, metode analisis data juga mengalami perkembangan pesat. Salah satu pendekatan yang digunakan adalah teknik *Clustering*, yang dapat membantu memahami pola dan distribusi curah hujan dengan mengelompokkan wilayah berdasarkan karakteristik yang serupa. Metode tradisional seperti regresi dan analisis deret waktu sering kali kurang efektif dalam menangkap keragaman spasial dan temporal yang kompleks, terutama dalam konteks perubahan iklim yang dinamis (Alaziz et al., 2023). Oleh karena itu, metode *Clustering* modern seperti DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) dan HDBSCAN (*Hierarchical DBSCAN*) mulai banyak digunakan karena kemampuannya dalam mengenali struktur data yang kompleks dan mengidentifikasi *outlier*, yang sangat relevan dalam analisis curah hujan yang sering kali tidak teratur (Chen et al., 2024). Selain itu, pendekatan *Clustering* temporal memungkinkan analisis yang lebih mendalam terhadap pola musiman dan tren jangka panjang dalam data curah hujan.

Hujan merupakan fenomena alam yang lazim terjadi di bumi. Dengan sebab munculnya akan menjadi sumber kehidupan sebagaimana telah dijelaskan di dalam al-qur’an surah Al-Furqan ayat 48 – 50 (Kementrian Agama Republik Indonesia, 2019).

وَهُوَ الَّذِيْٓ اَرْسَلَ الرِّيٰحَ بُشْرًاۢ بَيْنَ يَدَيْ رَحْمَتِهٖۚ وَاَنْزَلْنَا مِنَ السَّمَاۤءِ مَاۤءً طَهُوْرًاۙ لِّنُحْيِۦَ بِهٖ

بَلْدَةً مَّيْتًا وَّنُسْقِيَهٗ مِمَّا خَلَقْنَآ اَنْعَامًا وَّاَنَاسِيَّ كَثِيْرًا وَلَقَدْ صَرَّفْنٰهُ بَيْنَهُمْ لِيَذَّكَّرُوْاۖ فَاَبٰٓى اَكْثَرُ

النَّاسِ اِلَّا كُفُوْرًا

Artinya : Dialah yang meniupkan angin (sebagai) pembawa kabar gembira sebelum kedatangan rahmat-Nya (hujan). Kami turunkan dari langit air yang sangat suci.Agar dengannya (air itu) Kami menghidupkan negeri yang mati (tandus) dan memberi minum kepada sebagian apa yang telah Kami ciptakan, (berupa) hewan-hewan ternak dan manusia yang banyak .Sungguh, Kami benar-benar telah mempergilirkannya (hujan itu) di antara mereka agar mereka mengambil pelajaran. Akan tetapi, kebanyakan manusia tidak mau (bersyukur), bahkan mereka mengingkari (nikmat) *(QS. AL-Furqan : 48-50).*

Berdasarkan hal tersebut, Penelitian ini bertujuan untuk memanfaatkan metode terbaru *PCA-Clustering* dalam mengelompokkan pola hari hujan dan tidak hujan di dua stasiun klimatologi Jawa Timur, yaitu di daerah Sumberpucung dan Karangploso. Melalui pendekatan ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai pola distribusi kondisi cuaca berdasarkan karakteristik klimatologi, sehingga hasilnya dapat mendukung upaya mitigasi risiko bencana hidrometeorologi serta meningkatkan efektivitas pengelolaan sumber daya air, khususnya di wilayah dengan keragaman geografis yang tinggi seperti di Indonesia.

Penelitian yang dilakukan oleh Pamuji et al., membahas analisis dan pengelompokan data curah hujan di Jakarta dengan menggunakan dua algoritma kluster yang berbeda, yaitu K-Means dan DBScan. Dengan memanfaatkan data TRMM selama periode 1998–2007, penelitian tersebut menunjukkan bahwa K-Means lebih efisien dalam waktu pemrosesan, sedangkan DBScan memiliki kemampuan yang lebih baik dalam mengidentifikasi struktur data yang kompleks (Pamuji & Rongtao, 2020). Hasilnya menunjukkan bahwa K-mseans berhasil mengelompokkan data menjadi tiga kluster dalam waktu yang sangat cepat, hanya 2 detik, sementara DBScan membutuhkan waktu yang jauh lebih lama, yaitu 13 menit 32 detik, dan hanya menghasilkan dua kluster. PCA, atau *Principal Component Analysis*, adalah salah satu metode statistik yang banyak digunakan untuk mereduksi data dengan dimensi tinggi. Penelitian sebelumnya menunjukkan bahwa PCA digunakan untuk menganalisis tingkat kerusakan bangunan pasca bencana alam, dengan hasil yang menjanjikan menggunakan metode *Clustering* baru yang disebut *PCA-Clustering* (Almais et al. 2023). Di sisi lain, analisis *Clustering* dalam penelitian curah hujan bertujuan suntuk mengidentifikasi kelompok wilayah dengan karakteristik curah hujan serupa, yang dapat memberikan wawasan lebih dalam mengenai pola distribusi curah hujan (Sofro et al., 2024). Ini juga dapat mendukung pengembangan strategi adaptasi yang lebih efisien dalam menghadapi musim hujan ekstrem dan kekeringan berkepanjangan, serta memperkuat perencanaan tata ruang wilayah (Urgilés et al., 2024).

Dengan pendekatan yang sama, penelitian ini diharapkan dapat mengidentifikasi pola hari hujan dan tidak hujan secara lebih akurat, serta mendukung pengambilan keputusan berbasis data klimatologi di wilayah Malang

dan sekitarnya.PCA memiliki kelebihan dalam mereduksi dimensi data, sehingga mempermudah visualisasi dan interpretasi hasil *Clustering*. Namun, metode ini juga memiliki kekurangan, seperti ketergantungan pada asumsi linearitas dan kemungkinan kehilangan informasi selama proses reduksi dimensi. Dengan demikian, PCA lebih cocok digunakan ketika data memiliki dimensi tinggi dan kompleks. Pemilihan metode analisis yang tepat tetap bergantung pada karakteristik data yang dianalisis dan tujuan dari penelitian itu sendiri.

1.2 Pernyataan Masalah

1. Bagaimana mengimplementasikan *PCA-Clustering* untuk mengelompokkan curah hujan?
2. Bagaimana menentukan tingkat efesiensi *PCA-Clustering* dalam mengelompokkan curah hujan?

1.3 Tujuan Penelitian

1. Mengimplementasikan metode *PCA-Clustering* dalam proses pengelompokan data curah hujan guna melihat pola dan karakteristik hujan yang serupa pada wilayah atau waktu tertentu.
2. Menganalisis tingkat efisiensi metode *PCA-Clustering* dalam mengelompokkan curah hujan berdasarkan hasil evaluasi kualitas cluster yang dihasilkan, sehingga dapat diketahui sejauh mana metode ini efektif digunakan dalam analisis data klimatologi.

1.4 Hipotesis

1. H1: Penerapan metode PCA–*Clustering* dapat mengelompokkan data curah hujan berdasarkan karakteristik klimatologi menjadi kelompok dengan pola serupa.
2. H2: PCA–*Clustering* memiliki tingkat efisiensi yang cukup baik dalam mengelompokkan curah hujan, yang ditunjukkan oleh nilai Silhouette Score positif (>0,5).

1.5 Manfaat Penelitian

1. Memberikan informasi pola hari hujan dan tidak hujan untuk perencanaan aktivitas sehari-hari masyarakat.
2. Mendukung pengelolaan sumber daya air secara lebih efisien, termasuk irigasi dan penyimpanan air.
3. Membantu mitigasi risiko bencana hidrometeorologi, seperti banjir dan tanah longsor.
4. Menjadi dasar pengembangan sistem peringatan dini berbasis data.
5. Mendukung pengambilan keputusan masyarakat dan pemerintah untuk mengurangi dampak negatif terhadap lingkungan

1.6 Batasan Penelitian

1. Menggunakan data curah hujan dari Stasiun klimatologi Jawa Timur di pos hujan Sumberpucung dan Karangploso, Kabupaten Malang.
2. Periode data menggunakan data periode tahun 2023.

3. Penelitian ini memanfaatkan metode PCA untuk reduksi dimensi dan *Clustering* untuk mengelompokkan hari hujan dan tidak hujan berdasarkan data curah hujan.
4. Penelitian tidak membahas faktor sosial-ekonomi yang mungkin memengaruhi dampak curah hujan ekstrem.

1.7 Sistematika Penulisan

BAB I PENDAHULUAN

Menjelaskan latar belakang pentingnya analisis curah hujan di Malang, perumusan masalah, tujuan, manfaat, serta batasan penelitian.

BAB II STUDI PUSTAKA

Menguraikan teori tentang *Clustering* curah hujan, metode PCA–*Clustering*, penelitian terdahulu, serta kerangka teori sebagai dasar penelitian.

BAB III METODOLOGI PENELITIAN

Membahas desain sistem, sumber data BMKG 2023, tahap pembersihan dan standarisasi data, penerapan PCA, *Clustering*, serta metode evaluasi (Silhouette Score dan validasi ahli).

BAB IV PEMBAHASAN

Menyajikan hasil standarisasi, penentuan *Eigenvalue* dan variance ratio, pembentukan komponen utama (PC1 & PC2), proses *Clustering*, hasil label curah hujan, serta validasi dengan nilai Silhouette Score.

BAB V KESIMPULAN

Merangkum bahwa PCA–*Clustering* mampu mengelompokkan curah hujan menjadi dua kategori (hari hujan & hari tidak hujan) dengan kualitas cukup baik,

serta memberikan rekomendasi penelitian lanjutan dengan data lebih panjang dan variabel tambahan.

BAB II

STUDI PUSTAKA

2.1 *Clustering* Rainfall

PCA merupakan salah satu teknik untuk mereduksi data yang memiliki dimensi tinggi (Yan et al., 2024). Algoritma ini termasuk dalam metode *unsupervised* learning, yang artinya tidak membutuhkan label data. Proses PCA dimulai dengan melakukan normalisasi atau standarisasi data agar setiap fitur memiliki rata-rata nol dan standar deviasi satu, mengingat PCA sensitif terhadap skala variabel. Setelah itu, dihitung matriks kovarians untuk memahami hubungan antar variabel dan menentukan arah variansi terbesar dalam data. Dari matriks kovarians ini, PCA menghasilkan *Eigenvalue s* (nilai *Eigen*) dan *Eigenvectors* (*vektor Eigen*) yang masing-masing merepresentasikan besaran variansi dan arah komponen utama. PCA kemudian memilih beberapa komponen utama dengan nilai *Eigen* terbesar yang mencakup proporsi variansi total tertinggi, sering kali berdasarkan ambang batas tertentu, misalnya 90% dari variansi total. Data asli kemudian diproyeksikan ke ruang baru yang terbentuk oleh komponen utama terpilih, menghasilkan data berdimensi lebih rendah namun tetap mengandung informasi penting dari data awal. Parameter utama yang memengaruhi hasil PCA adalah jumlah komponen utama yang dipilih serta skala data. PCA biasanya digunakan sebagai langkah pra-pemrosesan dalam algoritma pembelajaran mesin, terutama pada data berdimensi tinggi, untuk mengurangi kompleksitas tanpa kehilangan terlalu banyak informasi kritis Klaster terbentuk pada berbagai skala spasial dan muncul di beragam disiplin ilmu. Fenomena ini tidak hanya menarik

dari segi ilmiah, tetapi juga memiliki implikasi signifikan dalam konteks teknologi (Uykan, 2023). *Clustering* merepresentasikan manifestasi konkret dari data yang tergolong dalam salah satu kelompok populasi tertentu. Beragam pendekatan seperti teknik campuran, algoritma *k-means*, *single linkage*, *complete linkage*, serta metode umum lainnya dianalisis untuk mengevaluasi validitas dan representativitas klaster yang terbentuk.

Penelitian Lima et al. Berfokus pada karakterisasi kejadian curah hujan ekstrem di Negara Bagian Rio de Janeiro, Brasil, memanfaatkan fungsi distribusi probabilitas (PDF) dan analisis klustering. Tujuan dari studi ini adalah untuk menentukan distribusi probabilitas yang paling sesuai dalam merepresentasikan distribusi curah hujan harian maksimum tahunan, serta mengevaluasi pola distribusi spasial dari curah hujan ekstrem tersebut. Metodologi yang diterapkan mencakup *Time-Series Analysis, Statistical Analysis, Clustering Analysis* untuk mengidentifikasi daerah homogen yang memiliki pola kejadian ekstrem yang serupa. Hasil penelitian mengungkap bahwa distribusi Gumbel, GEV, dan log-normal merupakan distribusi yang sangat cocok untuk mendeskripsikan curah hujan maksimum tahunan, sementara pola distribusi spasial dari kejadian-kejadian curah hujan ekstrem tersebut dipengaruhi oleh faktor topografi dan kedekatannya dengan pantai(Lima et al., 2021).

Penelitian oleh Sofro et al., analisis pola hujan di Indonesia dengan menggunakan pendekatan pengelompokan berbasis deret waktu, bertujuan untuk mengidentifikasi variasi pola hujan yang ada di berbagai wilayah dan bagaimana pola-pola tersebut berubah seiring waktu. Metode *Time-Series Analysis* dan *Clustering Analysis*, berhasil mengungkap beberapa kelompok yang masing-

masing menunjukkan pola hujan yang berbeda(Pamuji & Rongtao, 2020). Hasil penelitian ini sangat berharga bagi perencanaan dan manajemen sumber daya air di Indonesia, karena memungkinkan pengelolaan yang lebih efisien dan berkelanjutan berdasarkan karakteristik pola hujan yang teridentifikasi (Sofro et al., 2024).

Menurut Jayasekara et al,. pemantauan kualitas air di lahan basah Kotagala, Nuwara Eliya, Sri Lanka, dengan tujuan untuk memahami parameter kualitas air yang kompleks. Tantangan utama yang dihadapi adalah bagaimana menginterpretasikan data multidimensi yang rumit melalui metode statistik multivariat. Hasil penelitian menunjukkan bahwa analisis cluster berhasil mengelompokkan tujuh lokasi pengambilan sampel menjadi dua kategori, yaitu area tercemar dan kurang tercemar, dengan Stasiun 2 teridentifikasi sebagai lokasi paling tercemar akibat pembuangan limbah pertanian yang tidak terkendali. Parameter kunci seperti konduktivitas listrik (EC), pH, dan konsentrasi nitrat berperan penting dalam pengelompokan ini. Selain itu, analisis diskriminan mengungkapkan bahwa EC, Total *Dissolved Solids* (TDS), Total *Suspended Solids* (TSS), dan konsentrasi nitrat adalah parameter yang paling berpengaruh dalam membedakan kondisi musim kering dan basah. Temuan ini memberikan wawasan mendalam tentang dinamika kualitas air di lahan basah Kotagala, membantu pemangku kepentingan dalam upaya perlindungan dan pengelolaan sumber daya air yang berkelanjutan (Jayasekara et al., 2024).

Menurut Handayani et al., analisis perubahan iklim di Sumatera dan sekitarnya dengan memanfaatkan data meteorologi harian. Tantangan utama yang dihadapi adalah bagaimana mengelompokkan kota-kota berdasarkan kondisi meteorologi mereka untuk memahami pola perubahan iklim yang terjadi. Hasil

penelitian menunjukkan bahwa metode *Agglomerative Clustering* berhasil mengelompokkan 17 kota menjadi dua kluster utama. Cluster 1, yang terdiri dari Aceh, Sabang, Pekanbaru, Padang, dan Padang Lawas, memiliki karakteristik suhu rata-rata yang lebih tinggi, suhu minimum yang lebih rendah, serta durasi sinar matahari dan kelembaban yang lebih rendah dibandingkan dengan Cluster 2, yang mencakup kota-kota seperti Nagan Raya, Batam, dan Palembang. Selain itu, analisis tren tahunan mengungkapkan adanya peningkatan konsisten pada suhu minimum, peningkatan durasi sinar matahari, dan penurunan kecepatan angin, yang semuanya menunjukkan dampak nyata dari perubahan iklim di wilayah tersebut. Temuan ini memberikan wawasan penting bagi pengambil kebijakan dan masyarakat untuk lebih memahami dan merespons tantangan perubahan iklim yang semakin mendesak (Jayasekara et al., 2024).

Menurut Hendrawati et al 2024, model ARIMA yang tepat untuk mengelompokkan data deret waktu curah hujan di Jawa Barat, Indonesia. Masalah utama yang dihadapi adalah ketidakpastian dalam pemilihan model, di mana berbagai kriteria seleksi seperti AIC dan BIC dapat menghasilkan model yang berbeda-beda. Hasil penelitian menunjukkan bahwa metode ensemble distance secara konsisten memberikan persentase keanggotaan cluster yang lebih tinggi dibandingkan dengan metode Piccolo. Dalam periode observasi t=50, metode ini meningkatkan akurasi sebesar 15,16% berdasarkan kriteria RMSE, dan pada periode t yang lebih panjang, peningkatan akurasi tetap signifikan, dengan nilai yang bervariasi antara 11,46% hingga 12,54%. Dengan menerapkan metode ini pada data curah hujan, peneliti berhasil mengelompokkan 26 stasiun pemantau menjadi tiga cluster berbeda, masing-masing mencerminkan karakteristik curah

hujan yang unik. Temuan ini tidak hanya memberikan wawasan yang lebih baik tentang pola curah hujan di Jawa Barat, tetapi juga menawarkan pendekatan yang lebih handal untuk analisis data iklim di masa depan (Hendrawati et al., 2024).

Tabel 2.1 Peneliti terdahulu

Reference	Topik	Metode	Subject
Pamuji & Rongtao, 2020	Perbandingan algoritma *Clustering* curah hujan di Jakarta	K-Means vs DBScan	Data curah hujan TRMM 1998–2007
Almais et al., 2023a	Labeling tingkat kerusakan pasca bencana	*PCA-Clustering*	Data pascabencana multidimensi
Sofro et al., 2024	*Clustering* pola curah hujan berbasis waktu	Time Series *Clustering*	Curah hujan Indonesia
Lima et al., 2021	Karakterisasi hujan ekstrem di Brasil	PDF, *Clustering*	Hujan ekstrem Rio de Janeiro
Jayasekara et al., 2024	Kualitas air dan klasifikasi pencemaran	Multivariate *Clustering*, Discriminant Analysis	Data kualitas air Kotagala Wetland, Sri Lanka
Handhayani & Lewenusa, 2024	Pola iklim di Sumatra	Agglomerative *Clustering*	Data meteorologi harian di 17 kota
Hendrawati et al., 2024	*Clustering* curah hujan dengan ARIMA dan ensemble distance	Model-Based *Clustering*	Data curah hujan deret waktu Jawa Barat
Urgilés et al., 2024	Pola spatio-temporal hujan ekstrem di Andes	*Clustering* Analysis	Hujan ekstrem di pegunungan Andes
Bhattacharyya & Saha, 2023	Disaggregasi hujan dengan DL	ANN + K-Means	Data harian → per jam (India)
Wu et al., 2024	Regionalisasi hujan & analisis multiskala	PCA + Multivariate SOM + Wavelet	Data klimatologi ERA5 (Shanxi, China)

Reference	Topik	Metode	Subject
Wolski et al., 2020	Hubungan pola sirkulasi & hujan	PCA + SOM + regresi	Data atmosfer & curah hujan (Afrika Selatan)
Bhattacharyya et al., 2024	Disaggregasi hujan untuk banjir kota	MMRC, ANN-K, MMRC-K	Data 1 jam & harian dari 4 kota India
Penelitian ini	*Clustering*	PCA *Clustering*	*Clustering* Pola Hujan Dengan *PCA-Clustering* Menggunakan Data Klimatologi

2.2 Curah Hujan

Curah hujan, yang umumnya dinyatakan dalam satuan milimeter (mm), didefinisikan sebagai ketebalan lapisan air hujan yang terkumpul di atas permukaan datar apabila air tersebut tidak mengalami penguapan, tidak meresap ke tanah, dan tidak mengalir. Dengan kata lain, angka curah hujan menggambarkan volume air hujan per satuan luas dalam periode waktu tertentu. Instrumen yang digunakan untuk mengukur curah hujan adalah *ombrometer* atau *rain gauge*. Beberapa penelitian klimatologi menyebutkan bahwa nilai curah hujan tidak hanya menunjukkan intensitas presipitasi, tetapi juga menjadi indikator penting dalam analisis karakteristik iklim, pengelolaan sumber daya air, mitigasi bencana hidrometeorologi, serta perencanaan sektor pertanian dan infrastruktur. Dalam kajian ilmiah, curah hujan sering dianalisis sebagai variabel utama untuk memahami pola cuaca dan variasi iklim baik harian, bulanan, maupun tahunan.

2.3 Hari Hujan

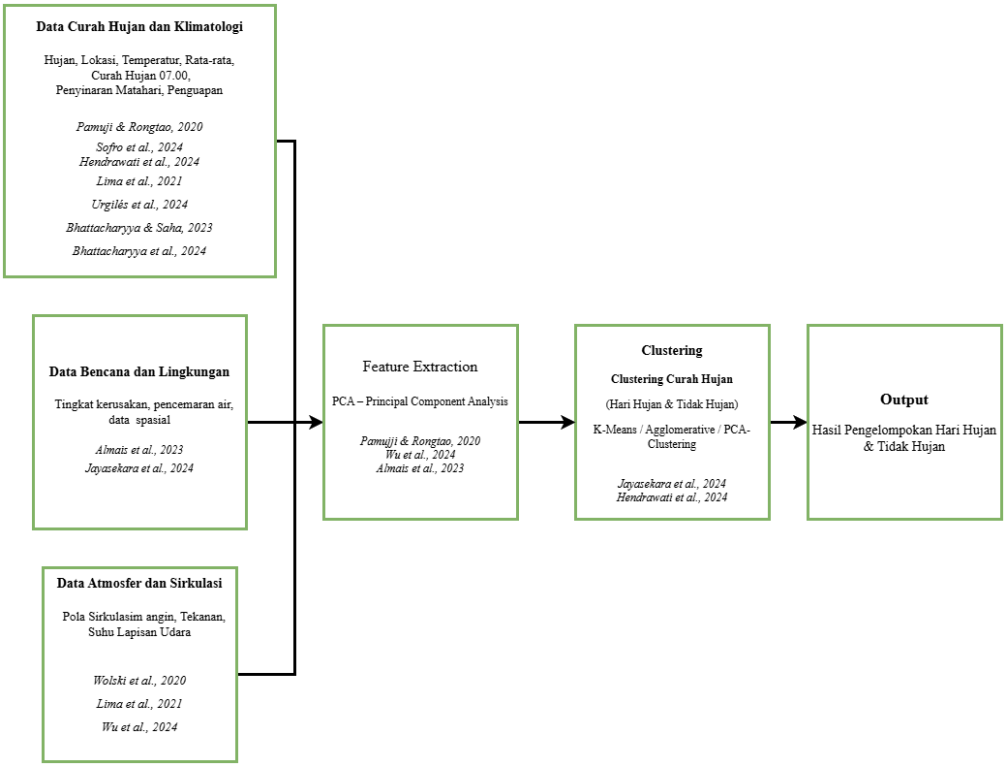
Hari hujan adalah hari ketika jumlah curah hujan yang tercatat dalam periode pengamatan 24 jam mencapai atau melebihi 0,5 mm. Ambang batas ini merupakan standar operasional yang digunakan oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) untuk membedakan hari dengan kejadian presipitasi dari hari tanpa hujan. Nilai ≥0,5 mm dianggap cukup untuk menunjukkan adanya hujan yang terukur secara meteorologis dan bukan hanya kabut, embun, atau gerimis sangat ringan yang tidak signifikan secara klimatologis. Dalam kajian klimatologi, frekuensi hari hujan menjadi indikator penting untuk memahami dinamika musim, variabilitas iklim, serta perubahan intensitas presipitasi harian yang berpengaruh pada analisis pola cuaca, manajemen sumber daya air, dan mitigasi risiko hidrometeorologi.

2.4 Curah Tidak Hujan

Hari tidak hujan adalah hari ketika jumlah curah hujan yang tercatat dalam periode pengamatan 24 jam berada pada kisaran kurang dari 0,5 mm. Pada kondisi ini, tidak terdapat presipitasi yang cukup signifikan untuk dikategorikan sebagai hujan menurut standar pengamatan BMKG. Nilai curah hujan <0,5 mm umumnya disebabkan oleh fenomena meteorologis seperti embun, kabut, atau tetesan sangat ringan yang tidak terukur secara konsisten oleh alat. Dalam konteks klimatologi, hari tidak hujan menjadi indikator penting untuk mengidentifikasi durasi periode kering, menganalisis karakteristik musim kemarau, serta memahami dinamika variabilitas atmosfer harian. Frekuensi hari tidak hujan juga berperan dalam analisis indeks kekeringan, evaluasi ketahanan air tanah, dan pengembangan strategi mitigasi risiko kekeringan meteorologis.

2.5 Kerangka Teori

Dasar penelitian ini dengan kerangka teori pada beberapa jurnal dan telah diilustrasikan dalam Gambar 2.1. Kerangka Teori disajikan di bawah ini.



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 data klimatologi seperti curah hujan, suhu udara, tekanan, kelembapan, dan penyinaran matahari, serta data atmosferik dan lingkungan yang berkaitan dengan kondisi pascabencana dan pola sirkulasi udara. Jenis data tersebut menjadi dasar utama dalam memahami variabilitas iklim dan distribusi curah hujan di suatu wilayah, terutama di daerah beriklim tropis seperti Indonesia.

Selanjutnya, pada bagian *proses*, Gambar 2.1 menampilkan dua kategori utama pendekatan analisis data, yaitu metode tanpa reduksi dan metode dengan reduksi dimensi. Metode tanpa reduksi, seperti K-Means, DBSCAN, *Time Series Clustering*, dan *Discriminant Analysis*, berfokus langsung pada pengelompokan data berdasarkan jarak dan kepadatan tanpa mengurangi jumlah dimensi variabel. Pendekatan ini efektif untuk dataset dengan dimensi rendah, namun sering mengalami penurunan performa ketika diterapkan pada data klimatologi yang bersifat multidimensi dan kompleks. Sebaliknya, metode dengan reduksi dimensi, seperti PCA (*Principal Component Analysis*), SOM (*Self-Organizing Map*), dan *Wavelet Regression*, digunakan untuk mengekstraksi variabel-variabel utama yang paling berpengaruh terhadap pola curah hujan. Dengan cara ini, data yang semula kompleks dapat disederhanakan tanpa kehilangan informasi penting, sehingga memudahkan proses visualisasi dan pengelompokan.

Posisi penelitian ini terletak pada pendekatan reduksi menggunakan PCA yang kemudian diintegrasikan dengan *Clustering*, membentuk metode *PCA-Clustering*. Pendekatan ini dipilih karena mampu mengoptimalkan hasil analisis dengan cara mengekstraksi komponen utama dari data klimatologi, kemudian memanfaatkannya untuk mengelompokkan pola hari hujan dan tidak hujan secara lebih efisien dan akurat. Penggunaan data klimatologi sebagai input utama menjadikan penelitian ini relevan dalam konteks analisis hidrometeorologi, sebab variabel seperti suhu, tekanan udara, dan penyinaran matahari memiliki hubungan langsung terhadap proses pembentukan hujan. Hasil akhir dari proses ini berupa *Clustering* curah hujan yang menggambarkan distribusi spasial dan temporal pola hujan, sehingga dapat dimanfaatkan untuk mendukung mitigasi bencana,

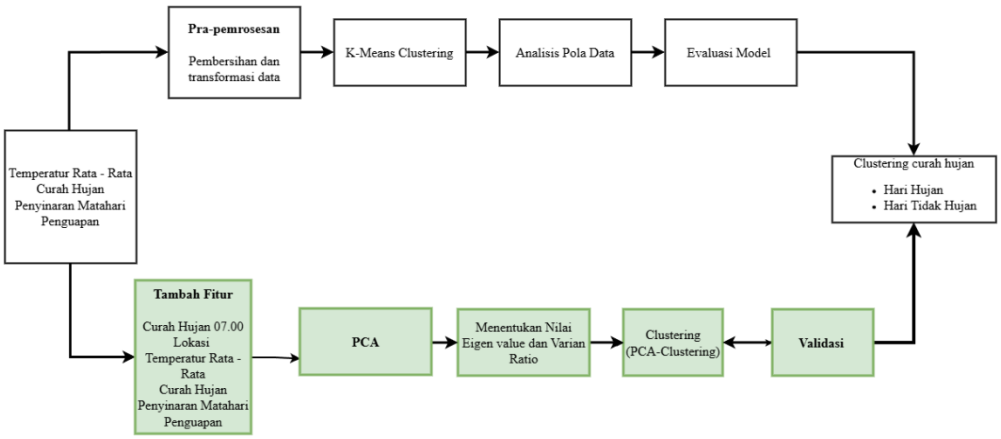
pengelolaan sumber daya air, serta pengambilan keputusan berbasis data klimatologi.

BAB III

METODOLOGI PENELITIAN

3.1 Kerangka Konsep

Kerangka konsep penelitian ini dirancang untuk menguraikan antara konsep data curah hujan, metode yang digunakan, dan hasil yang diharapkan yang akan diukur dalam sebuah penelitian dan telah diilustrasikan dalam Gambar 3.1.



Gambar 3.1 Kerangka Konsep

Gambar 3.1 menjelaskan secara terperinci alur analisis pola curah hujan 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 hari hujan dan tidak hujan yang representatif terhadap kondisi klimatologis di wilayah penelitian.

Pada tahap Input, data klimatologi yang digunakan mencakup beberapa variabel utama, yaitu suhu udara rata-rata, curah hujan harian (termasuk curah hujan pukul 07.00), lama penyinaran matahari, dan penguapan. Variabel-variabel tersebut diperoleh dari Stasiun Klimatologi BMKG yang mewakili karakteristik cuaca harian di wilayah penelitian. Sebelum dilakukan analisis, data melalui tahap pra-pemrosesan meliputi pembersihan data untuk mengatasi nilai hilang dan data

anomali, 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.

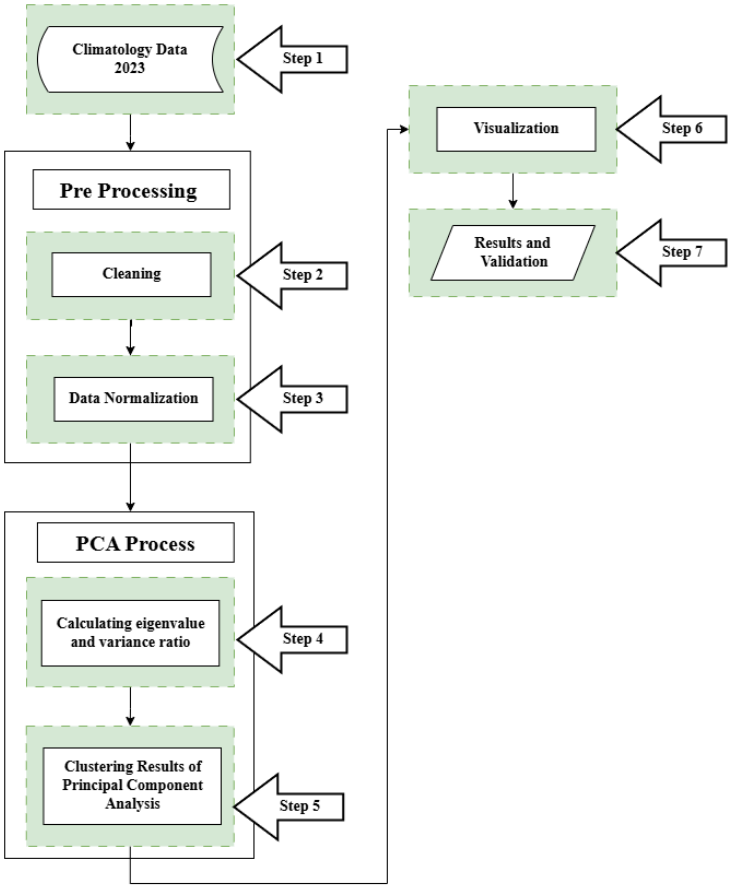
Tahap Proses merupakan inti dari analisis ini. Setelah data dinormalisasi, dilakukan perhitungan nilai eigenvalue dan variance ratio untuk menentukan komponen utama dengan menggunakan metode Principal Component Analysis (PCA). Proses ini bertujuan mereduksi dimensi data dengan tetap mempertahankan sebagian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

Tahap Output menghasilkan pengelompokan hari hujan dan tidak hujan yang merepresentasikan variasi pola curah hujan di wilayah penelitian. Melalui hasil pengelompokan ini, dapat diidentifikasi periode atau kelompok hari dengan karakteristik cuaca yang serupa, yang sangat berguna untuk analisis pola iklim dan perencanaan mitigasi bencana hidrometeorologi. 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 hari hujan dan tidak hujan yang lebih akurat, efisien, dan interpretatif.

3.2 Desain Sistem

Penelitian ini bertujuan untuk melakukan *Clustering* pada curah hujan. Diagram alur *Clustering* yang akan diusulkan pada studi penelitian ini tergambar pada Gambar 3.1 Desain Sistem.



Gambar 3.2 Desain Sistem

Principal Component Analysis digunakan untuk melakukan *Clustering* curah hujan. Langkah – Langkah *Prinsipal Component Analysis* yang digunakan adalah sebagai berikut :

Langkah 1: Data

Data diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Jawa Timur, khususnya dari pos hujan Sumberpucung dan karangploso https://dataonline.bmkg.go.id. Dataset mencakup periode 1 Januari sampai 31 Desember 2023.

Langkah 2: Pembersihan data

Proses pembersihan data dilakukan dengan mengeliminasi kesalahan, inkonsistensi, serta nilai-nilai yang hilang. Tahapan ini krusial untuk menjamin keakuratan dan keandalan data sebelum digunakan dalam analisis lebih lanjut (Rahman et al., 2019).

Langkah 3: Normalisasi data

Proses normalisasi data bertujuan memastikan kesesuaian dengan standar metode PCA. Tahapan ini menjadikan data sebanding serta berada pada skala seragam. Dalam penelitian mengenai PCA untuk *Clustering* mengkluster curah hujan, normalisasi dilakukan dengan metode Standard Scaler (SS) (Khan & Maity, 2020).

$$X_{Standart} = \frac{X - mean\ (x)}{\text{standartdeviation}\ (\chi)} \tag{3-1}$$

Standar deviasi menggunakan persamaan (1-2) sebagai berikut :

$$X \frac{\sum_{i=1}^{n} Xi}{n} \tag{3-2}$$

Langkah 4: Tentukan agenvalue dan ratio

Setelah data dinormalisasi, tahap selanjutnya adalah menentukan rasio varians serta nilai *Eigen* (*Eigenvalue*). Kedua parameter ini berperan penting untuk

menggambarkan seberapa besar informasi yang dapat ditangkap oleh setiap *principal component* (PC). Rasio varians dan nilai *Eigen* digunakan sebagai representasi distribusi data dalam suatu dimensi. Semakin besar nilainya, semakin tinggi pula proporsi varians yang mampu dijelaskan oleh PC tersebut. Dengan demikian, komponen utama yang memiliki nilai rasio varians dan *Eigen* tertinggi akan membawa lebih banyak informasi signifikan terkait data. (Almais et al. 2023).

Selanjutnya, jumlah *principal component* (PC) ditetapkan berdasarkan rasio varians kumulatif signifikan, sekitar 85%–95% dari total varians. Sejumlah nilai *Eigen* terbesar dipilih untuk membentuk PC. Penetapan PC dalam konteks ini difokuskan pada pelabelan data, bukan untuk identifikasi fitur dominan.

$$s^2 = \sum_{i=\frac{n}{1}} \frac{(x_i - x^2)}{(n-1)} \tag{3-3}$$

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. Hari Hujan : 0-20 mm/hari
b. Hari Tidak Hujan : 20-50 mm/hari

Langkah 6: Visualisasi

Hasil *Clustering* kemudian dianalisis berdasarkan komponen utama yang diperoleh dari proses PCA (J.M.A.U. Jayasekara, 2024). Melalui analisis ini, pola

pengelompokan dapat dikenali dengan lebih mudah karena data yang telah direduksi dimensinya menyajikan struktur yang lebih sederhana.

Langkah 7 : Hasil dan Validasi

Interpretasi hasil dan validasi data merupakan tahap akhir dalam proses PCA untuk *Clustering* tingkat kerusakan bangunan pasca gempabumi. Interpretasi mengacu pada grafik hasil *Clustering* berdasarkan rentang nilai yang telah ditetapkan sebagai patokan. Validasi memiliki peran penting untuk memastikan kesesuaian hasil dengan persyaratan. Proses validasi dilakukan dengan membandingkan pengelompokan data PC terhadap data target asli melalui korelasi, sehingga diperoleh nilai kesalahan terkecil. Metode ini meliputi *silhouette score* dan perbandingan hasil pengelompokan data PC1 dengan data target asli.

3.3 Data Preparation

Data yang digunakan dalam penelitian ini diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Jawa Timur, khususnya dari Stasiun Klimatologi Sumberpucung dan Karangkates. Dataset mencakup periode 1 Januari 31 Desember 2023 dengan resolusi harian. Terdapat 6 variabel seperti pada Tabel 3.1 dengan total data 713.

Tabel 3.1 Variabel Data klimatologi(Arisandi et al., 2021)

No.	Nama Variabel	Keterangan
1.	Lokasi	Lokasi pengambilan data
2.	Temperatur Rata-rata	Rata-rata suhu udara harian, dihitung dari nilai suhu maksimum dan minimum dalam 24 jam.
3.	Curah Hujan	Jumlah hujan yang turun dalam 24 jam pengamatan.
4.	Curah hujan (07.00)	Nilai curah hujan tertinggi yang tercatat dalam satu hari selama periode pengamatan.
5.	Penyinaran Matahari	Total waktu matahari bersinar di atas ambang intensitas tertentu dalam satu hari.
6.	Penguapan	Jumlah air yang menguap dari permukaan (biasanya dari panci evaporasi) dalam 24 jam.

Tabel 3.1 di atas 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 curah hujan

location	sp	sp	sp	…	kp	kp	kp
temp_avg	26,60	25,70	25,75	…	25,15	25,05	25,58
rh_07	72,75	84,25	23,00	…	91,19	93,09	93,05
rr	0,01	4,30	24,00	…	2,90	4,70	6,30
sunlight	43,75	56,25	21,25	…	93,75	40,00	68,75
pressure	978.70	979,00	979,90	…	946,90	945,90	946,30

Tabel 3.2 menjelasksn Variabel utama yang menjadi fokus dalam pembentukan label curah hujan adalah lokasi, temperature rata-rata, hari hujan jam 07.00, curah hujan, penyinaran matahari dan tekanan. Berdasarkan keeenam variavel, dilakukan klasifikasi tingkat intensitas hujan menjadi tiga kategori, yaitu hari hujan 00-20 mm/hari, dan hari tidak hujan 20-50mm/hari. Kategorisasi ini mengacu pada standar yang digunakan oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG).

Missing value penghapusan duplikasi, dan identifikasi serta perlakuan terhadap *outlier*. Untuk mengatasi *missing value*, digunakan metode *interpolasi linier*, yaitu dengan memperkirakan nilai hilang berdasarkan kecenderungan data sebelumnya dan sesudahnya(Libasin et al., 2021). Rumus interpolasi linier persamaan 4:

$$x=x1+\frac{x2-x1}{t2-t1}x(t-t1) \quad (3\text{-}4)$$

Keterangan :

x= nilai yang ingin dicari atau nilai hasil estimasi (nilai interpolasi pada waktu t)

x_1= nilai data yang diketahui sebelum data hilang (pada waktu t_1)

x_2= nilai data yang diketahui setelah data hilang (pada waktu t_2)

t_1= waktu (atau indeks) dari data ke-1 yang diketahui

t_2= waktu (atau indeks) dari data ke-2 yang diketahui

t= waktu (atau indeks) posisi data yang hilang (*missing value*)

Untuk normalisasi menggunakan normalisasi skalar adalah proses mengubah data numerik ke skala tertentu agar seragam dan bisa dibandingkan, (Shantal et al., 2023) pers 5:

$$X' = \frac{x-\mu}{\sigma} \quad (3\text{-}5)$$

Keterangan:

X : nilai asli

μ : rata-rata seluruh data

σ : standar deviasi data

x′ : nilai setelah distandarkan

Dalam proses *Clustering* untuk melihat apakah kelompok yang terbentuk secara tidak terawasi (*unsupervised*) dapat merepresentasikan pola hujan yang sesuai. Apabila jumlah variabel yang digunakan terlalu tinggi dan menunjukkan korelasi antar variabel yang signifikan, maka dilakukan reduksi dimensi menggunakan metode *Principal Component Analysis* (PCA) (Almais et al., 2023). PCA bertujuan untuk menyederhanakan kompleksitas data dengan memproyeksikannya ke dalam ruang berdimensi lebih rendah tanpa kehilangan terlalu banyak informasi.

3.4 *Principal Component Analysis Clustering* (PCA - *Clustering*)

Menurut (Almais et al. 2023), *Principal Component Analysis* (PCA) merupakan pendekatan statistik yang sering diterapkan pada analisis data berdimensi tinggi, meliputi reduksi dimensi, penyaringan derau, serta seleksi fitur. PCA juga berfungsi dalam penentuan label data, di mana label tersebut kemudian dimanfaatkan untuk proses klasifikasi melalui skala nilai hasil normalisasi yang telah diproyeksikan ke dalam sejumlah principal component (PC). Setiap komponen memiliki variasi rasio serta nilai *Eigen* yang berbeda. (Almais et al., 2024) menunjukkan bahwa PCA relevan diterapkan dalam proses pembentukan klaster, sehingga tingkat kerusakan pascabencana alam dapat diidentifikasi melalui label yang dihasilkan. Proses *Clustering* dengan PCA dilakukan melalui beberapa tahapan, mulai dari persiapan data, normalisasi, perhitungan nilai *Eigen* serta rasio varians, penentuan jumlah komponen utama, visualisasi dalam bentuk grafik, interpretasi hasil, hingga tahap validasi.

3.4.1 Input Data

Menurut (Almais et al. 2023) proses *Clustering* dengan metode *Principal Component Analysis* (PCA) diawali tahap persiapan data (*set up data*). Tahap ini berfokus pada penyiapan data yang akan diproses dalam analisis *Clustering*. *Clustering.* Persiapan data dilakukan dengan mengumpulkan data pada sejumlah frekuensi dan mengaturnya ke dalam table.

3.4.2 Normalisasi

Normalisasi merupakan tahap penting untuk menyeragamkan data agar sesuai dengan standar pada PCA. Menurut (Almais et al. 2023) standar data PCA adalah data yang digunakan harus memiliki derajat atau nilai yang sama dan

seimbang untuk setiap datanya. Proses normalisasi data pada PCA untuk *Clustering* curah hujan ini mengacu pada normalisasi menggunakan metode *Standardscale* (SS), yakni mengonversi data sehingga memiliki rata-rata sama dengan nol dengan standar deviasi sebesar satu. Persamaan yang digunakan pada proses normalisasi data *Standardscale* (SS) ditunjukkan pada persamaan 3.6 dan 3.7**.**

$$Xstandard = \frac{X - mean(x)}{standard\ deviation(x)} \quad (3.6)$$

$$x = \sum_{i=1}^{n} \frac{x_i}{n} \quad (3.7)$$

Simbol x̄ (Xbar) tersebut merepresentasikan nilai rata-rata dari himpunan x

3.4.3 Penentuan Rasio Varians, Nilai *Eigen* dan Jumlah (n) Principal Component

Dikutip dari penelitian (Almais et al. 2023) setelah tahap normalisasi data, langkah berikutnya yaitu menentukan rasio varians dan nilai *Eigen* (*Eigenvalue*). Kedua komponen tersebut berperan penting dalam memahami jumlah informasi yang tersimpan pada setiap *principal component* (PC). Rasio varians menunjukkan proporsi variabilitas data yang dijelaskan masing-masing PC, sehingga pada PCA rasio tersebut mencerminkan distribusi informasi dalam komponen utama.

$$s^2 = \sum_{i=1}^{n} \frac{(Xi - \bar{x})^2}{(n-1)} \quad (3.8)$$

Keterangan :

s^2= varians (variance), yaitu ukuran seberapa jauh penyebaran data dari nilai rata-ratanya.

x_i= nilai data ke-i pada suatu variabel.

$\bar{x}$= rata-rata (mean) dari seluruh nilai data pada variabel tersebut.

n= jumlah data dalam satu variabel.

$\sum_{i=1}^{n}$ = simbol penjumlahan dari data ke-1 hingga ke-n.

Persamaan 2.3 mempunyai pengertian bahwa distribusi data memiliki ukuran tertentu yang dapat digunakan untuk menentukan besarnya distribusi data.

Nilai *Eigen* merupakan skalar terkait setiap vektor *Eigen* (*Eigenvector*) dari matriks kovarians atau korelasi data. Dalam PCA, nilai ini merepresentasikan besarnya varians pada setiap *principal component* (PC). Sama seperti rasio varians, semakin tinggi nilai *Eigen* maka semakin besar pula varians yang dijelaskan PC tersebut. Kondisi ini menunjukkan bahwa komponen membawa lebih banyak informasi penting tentang data.

Setelah rasio varians dan nilai *Eigenvalue* ditentukan, langkah berikutnya yaitu penetapan jumlah *principal component* (PC). Tahap ini krusial karena berpengaruh terhadap tingkat akurasi data. Jumlah PC diperoleh dari nilai *Eigen* serta rasio varians sebelumnya, umumnya dengan mengambil nilai *Eigen* terbesar. Namun, nilai tertinggi tidak selalu menjadi acuan pada *PCA-Clustering* karena penetapan PC diarahkan untuk pelabelan data, bukan untuk identifikasi fitur dominan. Dalam hal ini, nilai *Eigen* berikutnya dapat dijadikan pembanding guna memastikan

ketepatan koordinat antar-PC terhadap data target asli, sehingga koordinat tersebut menghasilkan bentuk pelabelan data (Almais et al. 2023).

3.4.4 Visualisasi data

Dikutip dari Almais, dkk. (2023), visualisasi dalam bentuk grafik pada proses Principal Component Analysis (PCA) memiliki peranan yang signifikan dalam menyajikan hasil analisis, khususnya pada tahap keluaran dari proses reduksi dimensi. Visualisasi tersebut memungkinkan peneliti untuk memahami struktur data secara lebih komprehensif melalui representasi grafis yang menggambarkan hubungan antar komponen utama. Salah satu bentuk penyajian yang umum digunakan adalah visualisasi tiga dimensi (3D) yang mampu menampilkan rentang nilai setiap komponen secara lebih rinci. Nilai-nilai tersebut kemudian dapat diproyeksikan ke dalam grafik dua dimensi (2D) untuk memperoleh tampilan yang lebih sederhana, namun tetap informatif, sehingga hasil analisis dapat diinterpretasikan secara lebih efektif dan akurat.

Lebih lanjut, visualisasi PCA berfungsi tidak hanya sebagai sarana representasi hasil, tetapi juga sebagai instrumen analitis dalam mengidentifikasi pola keterkaitan antar variabel dan sebaran data pada ruang komponen utama. Melalui proyeksi dua dimensi, setiap titik data direpresentasikan berdasarkan kombinasi linear dari komponen utama yang memiliki kontribusi paling besar terhadap variansi total. Representasi ini memberikan gambaran mengenai bagaimana variabel klimatologi—seperti suhu udara, tekanan, kelembapan, penyinaran matahari, dan curah hujan—berkontribusi terhadap pembentukan pola atau kelompok data. Dengan demikian, visualisasi ini menjadi dasar untuk

mengamati kecenderungan data dan potensi pembentukan klaster sebelum dilakukan pengelompokan secara komputasional.

Selain itu, visualisasi PCA juga memiliki fungsi evaluatif dalam menilai efektivitas proses reduksi dimensi. Apabila sebaran data yang telah direduksi menunjukkan pola pemisahan yang jelas antara kelompok hari hujan dan hari tidak hujan, maka dapat diindikasikan bahwa proses PCA telah berhasil mempertahankan proporsi informasi utama dari data asli. Sebaliknya, apabila distribusi data menunjukkan tumpang tindih antar kelompok, hal tersebut mengindikasikan perlunya penyesuaian jumlah komponen utama atau metode standardisasi yang digunakan. Dalam konteks penelitian ini, visualisasi hasil PCA digunakan sebagai alat bantu analisis untuk mengevaluasi dan memperkuat tahap *Clustering*, sehingga hasil pengelompokan yang diperoleh melalui metode PCA–*Clustering* dapat dipertanggungjawabkan secara analitis maupun empiris.

3.4.5 Pembuatan Rentang

Dikutip dari Lambers, dkk. dalam penelitian Almais, dkk. (2023), pembuatan rentang nilai dilakukan dengan memaksimalkan nilai tertinggi pada suatu himpunan data serta meminimalkan kehilangan informasi akibat proses reduksi dimensi. Pendekatan ini bertujuan agar hasil analisis tetap mempertahankan proporsi variansi yang signifikan dari data asli. Dalam konteks Principal Component Analysis (PCA), rentang nilai berfungsi untuk menentukan batas atas dan batas bawah dari nilai komponen utama (principal components) yang dihasilkan, sehingga distribusi data dapat dikaji secara lebih objektif.

Lebih lanjut, Almais, dkk. (2023) menjelaskan bahwa rentang nilai dalam penelitiannya diturunkan dari hasil komponen utama (PC) yang telah diperoleh setelah proses transformasi PCA. Setiap komponen utama memiliki nilai variansi tersendiri yang menggambarkan sejauh mana komponen tersebut berkontribusi terhadap total keragaman data. Dengan menggunakan nilai variansi dan skor komponen utama tersebut, dapat dibentuk rentang numerik yang mencerminkan sebaran atau penyebaran nilai dari masing-masing PC. Rentang inilah yang kemudian menjadi dasar dalam analisis lanjutan, seperti pembagian interval atau penentuan batas klasifikasi data.

Pembuatan rentang nilai yang tepat berperan penting dalam menjaga keseimbangan antara tingkat representasi informasi dan efisiensi model. Jika rentang nilai terlalu sempit, maka variasi data yang signifikan dapat tereduksi secara berlebihan, sedangkan jika terlalu lebar, maka akan muncul redundansi informasi yang dapat mengganggu interpretasi. Oleh karena itu, penentuan rentang nilai yang optimal berdasarkan hasil komponen utama PCA menjadi tahap krusial untuk memastikan bahwa analisis berikutnya—seperti pengelompokan atau klasifikasi—dapat dilakukan dengan akurasi yang tinggi dan tetap menggambarkan karakteristik asli dari data klimatologi yang dianalisis.

3.4.6 *Clustering*

Dalam penelitian ini, proses pengelompokan dilakukan dengan membagi data curah hujan harian ke dalam dua kategori utama, yaitu hari hujan dan hari tidak hujan. Klasifikasi ini didasarkan pada besaran curah hujan (rr) yang tercatat dalam satuan milimeter per hari. Hari hujan didefinisikan sebagai hari dengan curah hujan berkisar antara 0 hingga 20 mm/hari, yang menunjukkan adanya

presipitasi atau turunnya hujan dengan intensitas ringan hingga sedang. Sedangkan hari tidak hujan didefinisikan sebagai hari dengan curah hujan antara 20 hingga 50 mm/hari, yang pada penelitian ini diartikan sebagai kondisi di mana tidak terjadi hujan dalam kategori yang sama dengan hari hujan. Pembagian kategori ini digunakan untuk menyederhanakan data curah hujan menjadi dua kelompok utama agar proses analisis dan pengelompokan dengan metode *Principal Component Analysis* (PCA) dan *Clustering* dapat dilakukan dengan lebih efektif. Melalui pendekatan ini, data curah hujan dapat diolah untuk mengidentifikasi pola hari hujan dan tidak hujan secara lebih terstruktur berdasarkan variasi nilai curah hujan di masing-masing lokasi penelitian.

Tahap pengelompokan data atau *Clustering* selanjutnya dilakukan untuk membagi data hasil reduksi PCA ke dalam kelompok tertentu berdasarkan kemiripan karakteristik nilai komponen utamanya. Metode ini tidak memerlukan syarat homogenitas penuh antar kelompok, melainkan berfokus pada pengelompokan titik data yang memiliki kedekatan nilai pada ruang komponen utama. Penelitian (Almais et al., 2023) menjelaskan bahwa proses *Clustering* data dapat dilakukan dengan memanfaatkan titik koordinat *Principal Component* (PC) berdasarkan rentang nilai tertentu. Namun demikian, hasil pengelompokan masih memerlukan tahap validasi tambahan agar tingkat akurasi dan kebenarannya dapat dipastikan.

Setelah proses pengelompokan selesai, tahap akhir penelitian berupa interpretasi hasil dan validasi dilakukan untuk menilai kualitas serta kesesuaian hasil yang diperoleh. Validasi memiliki peran penting dalam memastikan bahwa hasil *Clustering* yang dihasilkan benar-benar mewakili pola hari hujan dan tidak

hujan sesuai dengan data observasi. Dalam penelitian ini digunakan dua jenis validasi, yaitu validasi internal menggunakan nilai *Silhouette Score* untuk mengukur konsistensi antar anggota dalam klaster, serta validasi eksternal dengan membandingkan hasil pengelompokan data PCA terhadap data target asli dari BMKG. Melalui perbandingan tersebut, dapat diketahui tingkat kesesuaian hasil pengelompokan dengan data curah hujan aktual sehingga diperoleh nilai kesalahan terkecil dan hasil yang lebih representatif (Almais et al., 2024).

3.5 Skenario Evaluasi

3.5.1 Silhouette Score (Internal)

Silhouette Score merupakan metrik yang digunakan untuk mengevaluasi kualitas *Clustering* dengan mengukur seberapa baik data dalam sebuah cluster serupa satu sama lain (koherensi) dan seberapa baik cluster tersebut terpisah dari *cluster* lainnya (separasi) (Punhani et al., 2022).

Nilai untuk *Silhouette Score* mengevaluasi sejauh mana data dalam cluster serupa dan terpisah dengan *cluster* lain. Nilai berkisar antara -1 (buruk) hingga +1 (baik).

a. +1: Indikasi *Clustering* sangat baik. Data berada di *cluster* yang benar dan jauh dari cluster lainnya.
b. 0: Data berada di batas dua *cluster*, menunjukkan *ambiguitas*.
c. -1: *Clustering* buruk. Data lebih dekat ke cluster yang salah daripada ke *cluster* yang benar.

Langkah-langkah perhitungan *Silhouette Score*:

1. Menghitung *Cohesiveness* $a(i)$: Untuk setiap data i dalam cluster C, hitung rata-rata jarak *Euclidean* ke semua data lain dalam cluster C.

$a(i)\frac{1}{|C|-1}\sum_{j\in C, j\neq i} d(i,j)$ (3.2)

Keterangan:

$d(i,j)$: Jarak *Euclidean* antara data i dan j.

$|C|$: Jumlah data dalam cluster C.

2. Menghitung *Separability* $b(i)$: Untuk setiap data i, identifikasi cluster lain C' yang terdekat. Hitung rata-rata jarak data iii ke semua data dalam cluster C'.
3. Menghitung *Silhouette Score* untuk Data i: Substitusi nilai $a(i)$ dan $b(i)$ ke dalam rumus.

$S(i)\frac{b(i)-a(i)}{\max(a(i),b(i))}$ (3.3)

Keterangan :

a. S(i) : Nilai *Silhouette Score* untuk data ke-*i*. Nilai ini menunjukkan seberapa baik data tersebut berada di dalam klasternya masing-masing. Nilai S(i) berada pada rentang –1 ≤ S(i) ≤ 1.

b. a(i) : Rata-rata jarak antara data ke-*i* dengan seluruh data lain yang berada dalam klaster yang sama. Nilai ini menunjukkan tingkat kedekatan (cohesion) suatu data terhadap klasternya sendiri. Semakin kecil nilai *a(i)*, semakin baik data tersebut terkelompok dengan anggota klaster lainnya.

c. b(i) : Rata-rata jarak antara data ke-*i* dengan seluruh data pada klaster terdekat (klaster tetangga) yang berbeda.Nilai ini menggambarkan tingkat keterpisahan (separation) antar klaster. Semakin besar nilai *b(i)*, semakin jauh data tersebut dari klaster lain, yang berarti klaster tersebut lebih terpisah dengan baik.

d. max(a(i), b(i)) : Digunakan sebagai pembagi untuk menormalkan perbandingan antara *a(i)* dan *b(i)* agar hasil *S(i)* berada pada skala antara –1 dan 1.

4. Menghitung *Silhouette Score* Rata-rata untuk semua data: Setelah menghitung $S(i)$ untuk setiap data, hitung rata-rata semua $S(i)$ untuk mendapatkan skor keseluruhan.

$S_{avg}\ \frac{1}{N}\sum_{i=1}^{N} S(i)$ (3.4)

Keterangan :

a. S_{avg} : Nilai rata-rata Silhouette Score dari seluruh data yang digunakan untuk menilai kualitas keseluruhan pengelompokan (*Clustering*).

   Nilai ini menjadi indikator seberapa baik seluruh data terkelompok ke dalam klasternya masing-masing.

b. N : Jumlah total data yang digunakan dalam proses pengelompokan.

c. S(i) : Nilai Silhouette Score untuk setiap data ke-i, yang dihitung menggunakan rumus

5. Interpretasi Hasil:

a. $S_{avg} > 0.5$: *Clustering* sangat baik. Data dalam *cluster* sangat kohesif dan terpisah dengan jelas.

b. $0.25 < S_{avg} \leq 0.5$: *Clustering* cukup baik tetapi mungkin ada *overlap* antar-*cluster*.

c. $0.25 < S_{avg}$: *Clustering* kurang baik. Data tidak terbagi dengan baik ke dalam *cluster*.

3.5.2 Validasi Ahli (Eksternal)

Validasi berperan penting dalam menentukan apakah hasil yang diperoleh telah sesuai dengan persyaratan yang ada (Thabet et al., 2021) (Huang et al., 2022). Dalam penelitian ini, untuk melakukan validasi hasil, digunakan perbandingan hasil pengelompokan data PC1 dengan data target asli (Basile & Ferrara, 2023).

BAB IV

HASIL DAN PEMBAHASAN

4.1 Normalisasi Data

Proses standarisasi data pada *Principal Component Analysis* (PCA) bertujuan menyesuaikan data masukan agar berada pada skala seragam dan sebanding. S mencegah fitur dengan rentang nilai besar mendominasi fitur dengan rentang kecil dalam perhitungan. Perbedaan rentang nilai antarfitur, misalnya satu fitur berkisar 1–1000 sedangkan fitur lain 0,1–1, dapat menghasilkan varians tidak seimbang. Tanpa normalisasi, PCA berpotensi memberikan bobot tidak proporsional. Pada penelitian ini, proses normalisasi dilakukan dengan metode *Standard Scaler* (SS).

Standarisasi data dilakukan melalui pemrograman menggunakan skrip *Python di Google Colab* dengan memasukkan seluruh unsur-unsur klimatologi. Semua fitur numerik (kuantitatif), berupa variabel lokasi, temperatur rata-rata, jam hujan 07.00, curah hujan, penyinaran matahari, dan tekanan, yang akan melalui proses *Clustering* harus dinormalisasi. Tabel 4.1 menampilkan hasil standarisasi menggunkan *StandardScaler*.

Tabel 4.1 Hasil Normalisasi Data

0	1	2	3	4	5
1.0	1.099829	-0.7092	-0.39135	-0.95569	0.936419
1.0	0.533462	0.848647	0.006039	-0.52209	0.957005
1.0	0.564927	0.984112	1.826623	-1.73615	1.018762
1.0	0.866989	0.272921	1.27213	-0.95569	0.936419
1.0	0.973969	0.64545	-0.29893	-0.04514	0.936419
1.0	0.546048	0.64545	-0.39135	0.561889	0.936419

Tabel 4.1 menjelaskan setiap baris merepresentasikan satu sampel dari dataset asli. Dataset berjumlah 713 data curah hujan, sehingga setelah standarisasi dengan

metode StandardScaler, tetap diperoleh 713 baris yang kemudian menjadi input untuk proses reduksi dimensi oleh PCA. Setiap kolom hasil standarisasi merepresentasikan fitur asli dari dataset, namun telah diseragamkan skalanya. Pada tahap ini, dimensi data berhasil distandarisasi dan direduksi menjadi dua principal component (PC), meskipun sebelumnya mungkin terdapat lebih banyak PC. PCA berhasil mengompresi data sehingga diperoleh representasi yang lebih ringkas, terlihat pada terbentuknya tiga kolom hasil reduksi.

Nilai hasil standarisasi data menunjukkan angka-angka yang telah diskalakan dalam ruang fitur asli. Selanjutnya, nilai tersebut ditransformasikan ke ruang principal component (PC). Angka-angka ini berperan sebagai koordinat baru untuk merepresentasikan variasi terbesar dalam data. Dengan demikian, ruang PC membentuk dimensi baru dua arah, terdiri dari PC1 dan PC2.

4.2 Pemformatan Jumlah (n) Principal Component (PC) dan Penentuan Rasio Varians.

Tentukan rasio varians serta nilai *Eigen* (*Self-esteem*). Kedua parameter ini memainkan peran penting dalam menggambarkan berapa banyak informasi yang masing-masing dapat menangkap *komponen utama* (PC). Rasio varians dan *Eigenvalue* digunakan sebagai representasi distribusi data dalam dimensi(Tripathi & Garg, 2021). Semakin besar nilainya, semakin tinggi proporsi varians yang dapat menjelaskan penyebaran PC. Dengan demikian, komponen utama yang memiliki nilai varians dan *Eigen*ratio tertinggi akan membawa informasi yang lebih signifikan terkait data. (Almais et al. 2023).

Menggunakan library yang ada pada python, untuk menentukan nilai *Eigenvalue* dan varians ratio pada masing-masing PC sehingga dapat

menghasilkan suatu nilai yang menggambarkan letak titik koordinat pada sumbu x (PC1) dan sumbu y (PC2). Untuk nilai *Eigenvalue* dan varians ratio terdapat pada table 4.1 yang sudah tervisualisasi pada gambar 4.1.

Pada persamaan 2.3 menunjukkan rumus untuk mencari *Eigen Value*

1. Nilai *Eigen*

   PC1 = X_i = 2,53

   PC2 = $\sum\lambda = 1.87$

2. Hitung total Nilai *Eigen*

$$\sum\lambda = \lambda_1 + \lambda_2 = 2.53 + 1.87 = 4.40$$

3. Hitung total Nilai *Eigen*

$$\text{Variance Ratio}_1 = \frac{2.53}{4.40} = 0.574 \approx 57.5\%$$

$$\text{Variance Ratio}_2 = \frac{1.87}{4.40} = 0.425 \approx 42.5\%$$

4. Hitung total variansi yang dijelaskan

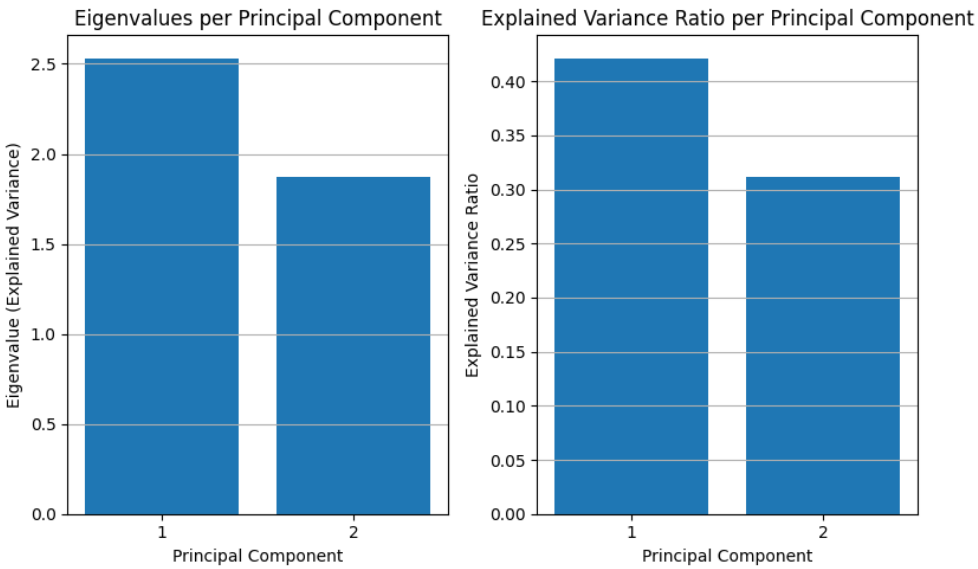
$$\text{Total Variansi Dijelaskan} = 42.14\% + 31.19\% = 73.33\%$$

PC1 dan PC2 bersama-sama menjelaskan 73.33% variansi data asli, cukup tinggi sehingga bisa dianggap representasi utama data.

Tabel 4.2 Eigen Value And Variance ratio

Number of components	*Eigenvalue*	*Variance ratio* (%)
1.0	2.53	42.14
2.0	1.87	31.19
Total		73.33

Berikut gambar grafik Dari nilai Grafik *Eigenvalue* dan *Varian Ratio*



Gambar 4.1. Grafik *Eigenvalue* dan *Varian Ratio*

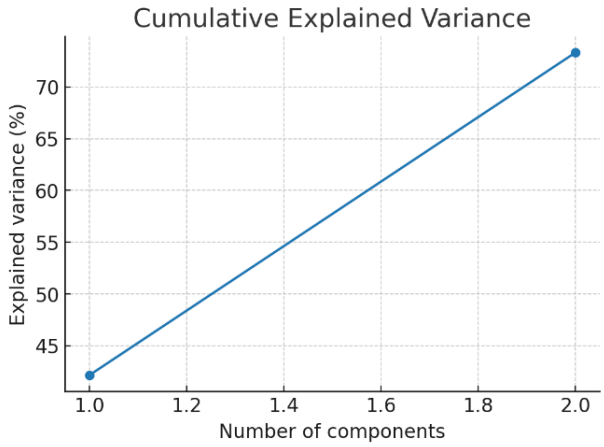
Gambar 4.1 memperlihatkan hasil analisis nilai *Eigenvalue* dan *Explained Variance Ratio* dari proses Principal Component Analysis (PCA) yang digunakan untuk mereduksi dimensi data klimatologi. Berdasarkan grafik pada sisi kiri, komponen utama pertama (PC1) memiliki nilai *Eigenvalue* sebesar sekitar 2.5, sedangkan komponen utama kedua (PC2) memiliki nilai sekitar 1.8. Nilai *Eigenvalue* ini menunjukkan besarnya variansi data yang mampu dijelaskan oleh masing-masing komponen. Semakin besar nilai *Eigenvalue*, semakin besar pula informasi atau keragaman data yang berhasil ditangkap oleh komponen tersebut. Dengan demikian, PC1 menjadi komponen dominan karena mampu menjelaskan sebagian besar variasi dari data klimatologi, seperti curah hujan, suhu, tekanan udara, dan penyinaran matahari.

4.3 Hasil Pengelompokan Analisis Komponen Utama

Untuk menentukan jumlah *principal component* (PC), perlu melihat *Eigenvalue* dan *variance ratio*. Pada table 4.2, nilai *Eigenvalue* terbesar terdapat pada komponen utama pertama (PC1) dengan *Eigenvalue* 2.53 serta *variance ratio* 42.14%. Namun, nilai *Eigenvalue* terbesar tidak dapat dijadikan acuan penelitian ini karena tujuan penelitian adalah pelabelan data, bukan penentuan fitur paling dominan.

Selain penggunaan PC1 dengan *Eigenvalue* tertinggi, analisis juga melibatkan principal component kedua (PC2) dengan nilai *Eigenvalue* lebih kecil dibandingkan PC1, yaitu 1.85 serta *variance ratio* 31.19%. Penggunaan PC2 bertujuan memberikan perbandingan saat mengukur tingkat akurasi proses pengujian pada pelabelan data.

Nilai *variance ratio* untuk PC1 serta PC2 terlihat pada grafik, memperlihatkan bahwa nilai *variance ratio* PC1 serta PC2 membentuk garis lurus. Hal tersebut menandakan data pada PC2 memiliki sebaran lebih luas jika dibandingkan dengan sebaran data pada PC1.



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

Gambar 4.2. Grafik menunjukkan hubungan jumlah komponen utama dengan persentase variasi data dapat dijelaskan. Komponen pertama mampu menjelaskan 42.14% variasi, sedangkan penambahan komponen kedua meningkatkan total variasi menjadi 73.33%. Hal ini menandakan dua komponen sudah cukup menjelaskan sebagian besar informasi data sehingga dapat digunakan untuk reduksi dimensi tanpa kehilangan terlalu banyak informasi penting

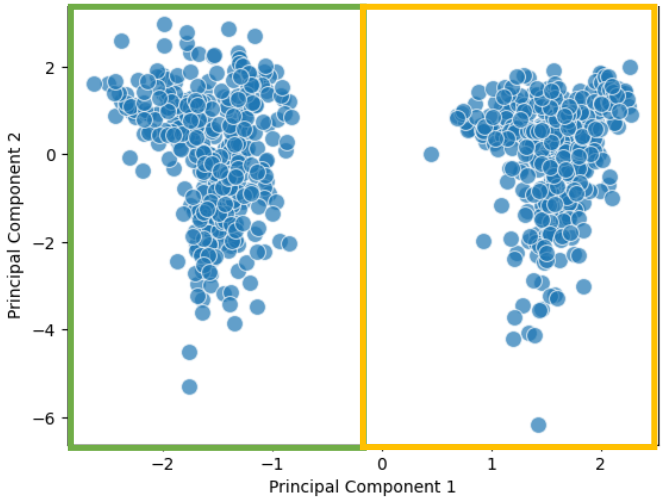
4.4 Label Berdasarkan Hasil Grafis

Setelah komponen utama diperoleh melalui proses PCA (Almais et al., 2023b), *Clustering* merupakan teknik untuk mengelompokkan data ke dalam beberapa kelompok homogen(Pansera et al., 2013). 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. Hari Hujan: 0-20 mm/hari
b. Hari Tidak Hujan: 20-50 mm/hari

4.5 Proses *Clustering*

Proses *Clustering* dilakukan dengan memanfaatkan hasil transformasi PCA(Punhani et al., 2022)(Pansera et al., 2013). Data yang telah direduksi ke dalam ruang komponen utama (PC) divisualisasikan dalam koordinat PC1 dan PC2(Almais et al., 2023b). Distribusi titik data kemudian dipisahkan ke dalam kelompok berdasarkan rentang nilai koordinat seperti pada gambar 4.



Gambar 4.3 Coordinate Points of PC1 and PC2 Data Distribution

Tabel 4.4 menjelaskan titik koordinat pada setiap kelompok berdasarkan nilai PC1 dan PC2.

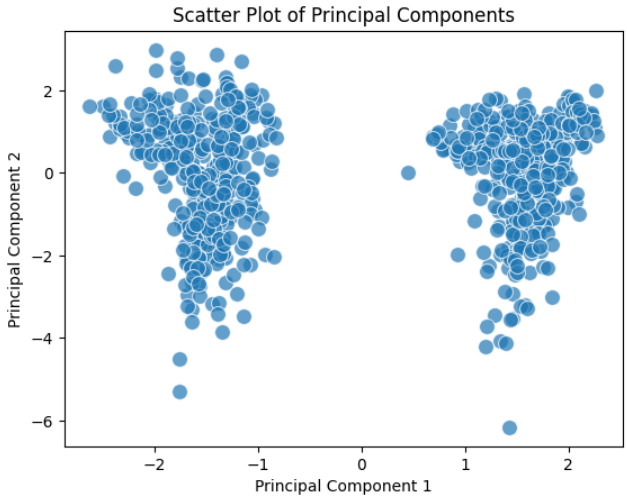
Tabel 4.3 Point and coordinate *value* PC1 dan PC2

Coordinate Point		**Coordinate *Value* Range** (n)
PC1	**PC2**	
-1.50 -0.96 -1.72	-0.39 -0.04 1.44	$n < 0$
0.12 0.40 2.00 2.55 2.81	-0.13 -1.05 -0.50 -1.01 1.45	$0 \leq n \leq 3$

Tabel 7 menyajikan hasil pengelompokan titik koordinat berdasarkan nilai PC1 dan PC2. Kelompok hari hujan tercatat memiliki nilai PC1 negatif dengan tiga titik koordinat berbeda. Kelompok hari tidak hujan mencakup titik dengan nilai PC1 antara 0 hingga 3, menghasilkan lima titik koordinat. Tidak terdapat titik dengan nilai PC1 lebih besar dari 3 sehingga kategori Rendah Lebat tidak muncul pada hasil

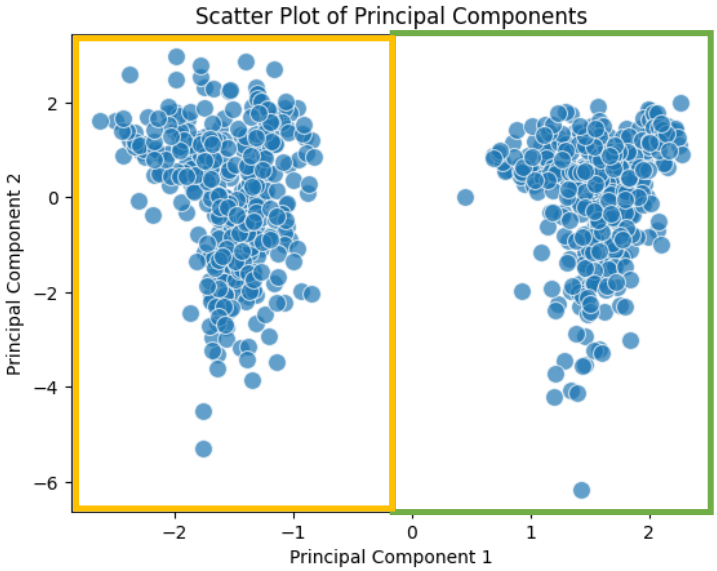
4.6 Hasil *Clustering*

Hasil distribusi PCA menghasilkan 2 komponen utama, yaitu PC1 dan PC2.



Gambar 4.4 Titik koordinat PC1 dan PC2

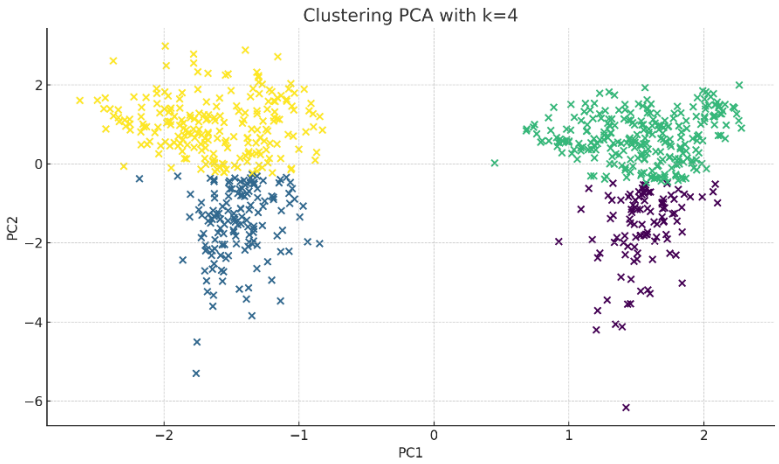
Gambar 4.4 memperlihatkan sebaran data hasil transformasi PCA ke dalam dua komponen utama, yaitu PC1 dan PC2. Titik-titik pada grafik menunjukkan distribusi curah hujan berdasarkan nilai komponen utama yang diperoleh dari data BMKG. Terlihat adanya dua kelompok sebaran data yang berbeda posisi pada sumbu PC1, menandakan potensi pembentukan hari hujan dan tidak hujan. Visualisasi ini digunakan sebagai tahap awal validasi, dengan cara membandingkan distribusi hasil PCA terhadap data target asli untuk menilai konsistensi pengelompokan.



Gambar 4.5 Hasil dari *Clustering* data Titik koordinat PC1 dan PC2

Gambar 4.5 menunjukkan hasil *Clustering* data curah hujan yang menunjukkan 2 cluster. Untuk menjelaskan data *Clustering* pada Gambar 4.5 dapat dilihat pada Tabel 4.4.

Gambar 4.6 Hasil 3 Klaster



Gambar 4.7 Hasil 4 Klaster

Tabel 4.4 Hasil klaster

Jumlah Klaster (k)	Silhouette Score
2 klaster	0.5538
3 klaster	0.5622 ***(tertinggi)***
4 klaster	0.5610

Pemilihan jumlah klaster ditentukan berdasarkan perbandingan nilai Silhouette Score untuk k = 2, 3, dan 4. Hasil evaluasi menunjukkan bahwa k = 3 0.5622, diikuti oleh k = 4 sebesar 0.5610, dan k = 2 sebesar 0.5538. Secara matematis, k=3 memiliki Silhouette sedikit lebih tinggi, tetapi selisihnya sangat kecil dan tidak signifikan secara struktur cluster. Meskipun demikian, selisih nilai antar skenario sangat kecil (< 0.01) sehingga peningkatan jumlah klaster tidak memberikan pemisahan yang jauh lebih baik. Namun, Secara klimatologi hanya ada dua kondisi utama (hujan & tidak hujan), sehingga k=2 tetap menjadi pilihan paling logis.

Secara klimatologis, data curah hujan harian hanya memiliki dua kategori fenomena utama, yaitu hari hujan dan hari tidak hujan. Oleh karena itu, pemilihan k = 2 lebih tepat secara makna fisik dan interpretasi, meskipun nilai Silhouette tidak tertinggi. Dengan demikian, penelitian ini menetapkan penggunaan dua klaster sebagai hasil akhir yang paling representatif.

4.7 Validasi Hasil *Clustering*

Validasi hasil pengelompokan dilakukan melalui dua pendekatan, yaitu validasi internal dan eksternal. Validasi internal menggunakan Silhouette Score untuk mengukur kualitas dan konsistensi klaster, sedangkan validasi eksternal dilakukan dengan membandingkan hasil pengelompokan terhadap data target asli menggunakan distribusi PCA berdasarkan dua komponen utama (PC1 dan PC2).

4.7.1 Silhouette Score

Selain validasi visual dan perbandingan dengan target data asli, evaluasi kuantitatif juga dilakukan menggunakan Silhouette Score.

a. Hitung *Cohesiveness* $a(i)$

Nilai $a(i)$menunjukkan seberapa dekat data $i$terhadap data lain dalam kluster yang sama seperti ditunjukkan persamaan (3.2).

Nilai komponen PCA-nya:

$$(PC1_1, PC2_1) = (-1.50, -0.39)$$

Anggota lain dalam kluster yang sama:

$$(-0.96, -0.04) \text{ dan } (-1.72, 1.44)$$

Hitung jarak Euclidean antar titik (hanya dalam cluster sama):

$$d(i,j) = \sqrt{(PC1_i - PC1_j)^2 + (PC2_i - PC2_j)^2}$$

Maka*:*

$$d(1,2) = \sqrt{(-1.50 + 0.96)^2 + (-0.39 + 0.04)^2} = 0.56$$
$$d(1,3) = \sqrt{(-1.50 + 1.72)^2 + (-0.39 - 1.44)^2} = 1.86$$

Rata-rata jarak dalam cluster :

$$a(1) = \frac{0.56 + 1.86}{2} = 1.21$$

b. Hitung *Separability* $b(i)$

Nilai $b(i)$menunjukkan seberapa dekat data $i$terhadap kluster lain yang paling dekat seperti ditunjukkan pada persamaan (3.2). Kluster terdekat adalah kluster curah hujan tidak hujan dengan empat titik:

$$(0.12, -0.13), (0.40, -1.05), (2.00, -0.50), (2.55, -1.01)$$

Maka jarak rata-rata ke anggota kluster tersebut:

$$b(1) = \frac{1.63 + 2.06 + 3.50 + 4.08}{4} = 2.82$$

c. Hitung Silhouette Coefficient $s(i)$

Subtitusi hasil pada Persamaan (3.4)

$$s(1) = \frac{2.82 - 1.21}{2.82} = 0.57$$

Rata-rata hasil keseluruhan data:

$$S = \frac{1}{7}\sum s(i) = 0.55$$

Tabel 4.5 Hasil Evaluasi Kualitas *Clustering* dengan *Silhouette Score*

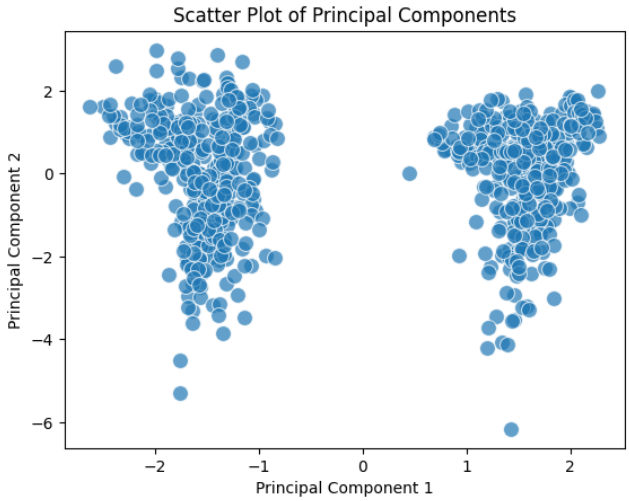
Metode	Silhouette Score
PCA–*Clustering*	0,55

Berdasarkan Tabel 4.5 nilai *Silhouette Score* sebesar 0,55 menunjukkan bahwa hasil PCA–*Clustering* memiliki kualitas pengelompokan yang cukup baik.

Nilai ini mengindikasikan bahwa data lebih dekat dengan cluster masing-masing dibandingkan dengan cluster lain, sehingga pemisahan antara cluster hari hujan dan hari tidak hujan dapat dianggap valid. Dengan demikian, hasil evaluasi kuantitatif ini mendukung kesesuaian pengelompokan PCA–*Clustering* dengan klasifikasi curah hujan resmi BMKG.

4.7.2 Validasi dengan Target Asli

Pengujian validasi hasil pengelompokan data dilakukan dengan membandingkan hasil data target asli dari 713 data yang digunakan dengan hasil distribusi data yang dihasilkan menggunakan PCA. Hasil distribusi PCA menghasilkan 2 komponen utama, yaitu PC1 dan PC2.



Gambar 4.7 Titik koordinat PC1 dan PC2

Gambar 4.7 memperlihatkan sebaran data hasil transformasi PCA ke dalam dua komponen utama, yaitu PC1 dan PC2. Titik-titik pada grafik menunjukkan distribusi curah hujan berdasarkan nilai komponen utama yang diperoleh dari data BMKG. Terlihat adanya dua kelompok sebaran data yang berbeda posisi pada sumbu PC1, menandakan potensi pembentukan klaster hari hujan dan tidak hujan. Visualisasi ini digunakan sebagai tahap awal validasi, dengan cara membandingkan

distribusi hasil PCA terhadap data target asli untuk menilai konsistensi pengelompokan.

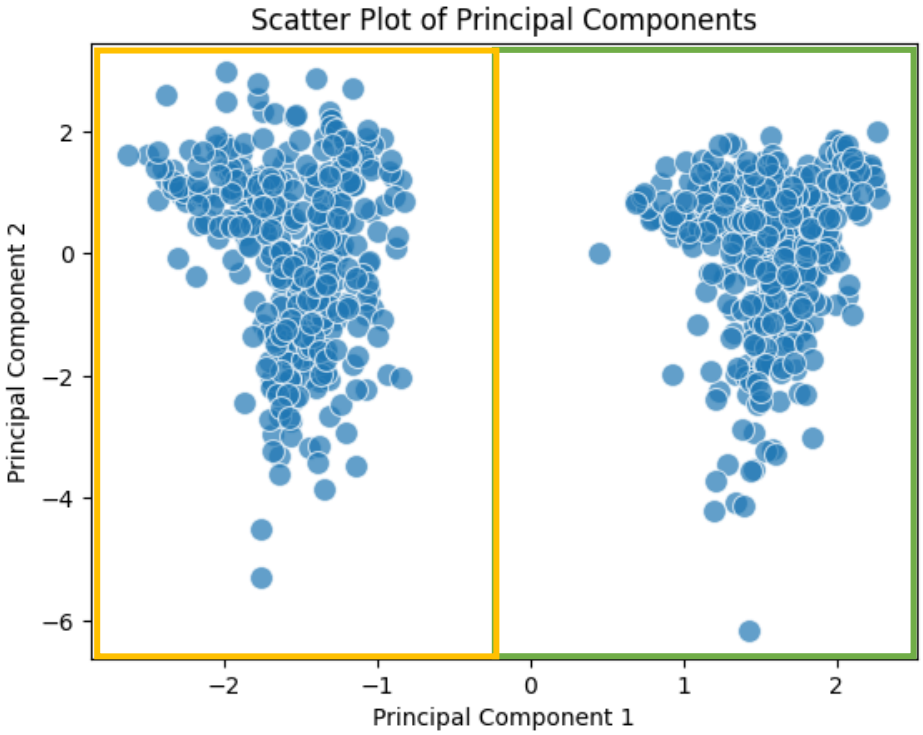
Tabel 4.4 Data Ground Truth Intensitas Curah Hujan

No	Curah Hujan Harian	Intesitas
1.	< 20 mm	Hari Hujan
2.	20 – 50 mm	Hari Tidak Hujan

Tabel 4.4 di atas menunjukkan kategori data ground truth yang digunakan sebagai acuan dalam proses validasi eksternal hasil pengelompokan. Data ground truth ini disusun berdasarkan klasifikasi intensitas curah hujan yang merujuk pada pedoman Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) serta World Meteorological Organization (WMO). Dalam penelitian ini, curah hujan harian dengan nilai kurang dari 20 mm dikategorikan sebagai hari hujan, sedangkan curah hujan dengan nilai antara 20 hingga 50 mm dikategorikan sebagai hari tidak hujan. Penentuan kategori tersebut bertujuan untuk memberikan batas kuantitatif yang jelas dalam membedakan kondisi hujan dan tidak hujan berdasarkan intensitas curah hujan harian. BMKG (2020) mengelompokkan curah hujan ke dalam beberapa kategori, antara lain: sangat ringan (<5 mm), ringan (5–20 mm), sedang (21–50 mm), lebat (51–100 mm), dan sangat lebat (>100 mm). Dengan merujuk pada batasan ini, penelitian ini menetapkan nilai <20 mm sebagai kondisi hari hujan (karena masih dalam kategori hujan ringan), sedangkan nilai antara 20–50 mm dikategorikan sebagai hari tidak hujan untuk membedakan kondisi yang tidak termasuk intensitas tinggi dalam periode pengamatan.

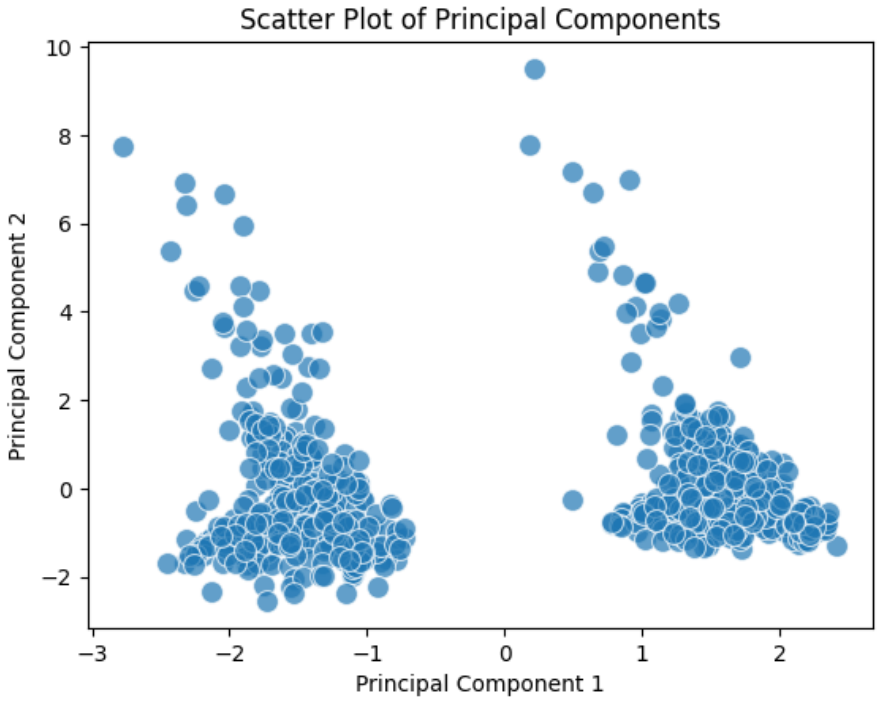
Kriteria ini digunakan sebagai label referensi (*ground truth*) dalam proses validasi eksternal, yang membandingkan hasil pengelompokan PCA–*Clustering* dengan data aktual dari BMKG. Dengan adanya data acuan ini, keakuratan hasil

Clustering dapat dievaluasi secara objektif untuk memastikan bahwa kelompok hari hujan dan tidak hujan yang dihasilkan model sesuai dengan kondisi klimatologis sebenarnya. Selain itu, penggunaan klasifikasi berbasis standar BMKG dan WMO memastikan bahwa hasil penelitian memiliki dasar ilmiah yang kuat dan dapat dibandingkan dengan kajian klimatologi lainnya pada skala nasional maupun internasiona



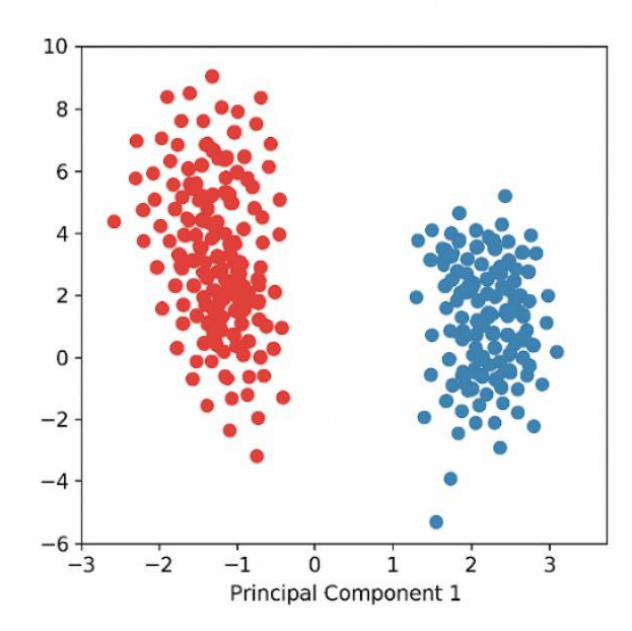
Gambar 4.8 Hasil dari *Clustering* data Titik koordinat PC1 dan PC2

Gambar 4.8 menunjukkan hasil *Clustering* data curah hujan yang menunjukkan 2 cluster. Untuk menjelaskan data *Clustering* pada Gambar 4.7 dapat dilihat pada Tabel 4.6.



Gambar 4.9 Hasil dari *Clustering* data target

Gambar 4.9 Visualisasi data target dari BMKG menunjukkan dua kelompok utama yang terpisah dengan jelas pada bidang *Principal Component* 1 (PC1) dan *Principal Component* 2 (PC2). Pola ini menggambarkan distribusi alami antara hari hujan dan hari tidak hujan berdasarkan karakteristik iklim seperti suhu, kelembapan, tekanan udara, dan penyinaran matahari. Pemisahan yang jelas tersebut menunjukkan bahwa variabel-variabel iklim memiliki pengaruh kuat dalam membedakan kondisi cuaca pada wilayah pengamatan BMKG.



Gambar 4.10 Hasil Pengelompokan data PC1 berdasarkan target data asli

Gambar 4.10 Hasil overlay antara data target (merah) dan hasil PCA–*Clustering* (biru) menunjukkan bahwa kedua kelompok memiliki pola sebaran yang serupa dan terpisah jelas di sepanjang sumbu komponen utama (PC1). Hal ini menandakan bahwa proses PCA berhasil mereduksi dimensi data dengan baik, sementara hasil *Clustering* mampu mengelompokkan data secara konsisten dengan struktur alami pada data target.

Tabel 4.6 Label Data Hasil Perbandingan PC1 Berdasarkan Target Asli

Comparison Result		**Coordinate *Value* Range n**	**Label**
PC1 *Value*	**Target Data Original**		
-1.50	0-20	n < 0	Hari Hujan
-0.96			
-1.72			
0.12	20-50	0 ≤ n ≤ 3	Hari Tidak Hujan
0.40			
2.00			
2.55			
2.81			

Tabel 4.6 memperlihatkan bahwa penentuan label tingkat curah hujan mengacu pada target data asli BMKG. Berdasarkan ambang batas yang diterapkan

pada sumbu PC1, titik data dikategorikan ke dalam dua cluster: Cluster 1 ($n < 0$) dan Cluster 2 ($0 \leq n \leq 3$), sesuai dengan hasil pengelompokan hari hujan dan tidak hujan.

Temuan ini menegaskan bahwa proses pelabelan menggunakan PCA mampu memberikan klasifikasi yang konsisten dengan kategori BMKG. Keunggulan PCA terletak pada kemampuannya untuk melakukan pengelompokan data secara unsupervised, sehingga tidak memerlukan data berlabel secara eksplisit pada tahap awal, berbeda dengan metode *supervised* yang bergantung pada data latih berlabel.

Distribusi hasil pelabelan berdasarkan PCA dapat dirumuskan sebagai berikut:

1. Target data asli 0–20 mm/hari terkelompok pada koordinat PC1 dengan rentang $n < 0$, kemudian memperoleh label hari hujan.
2. Target data asli 20–50 mm/hari terkelompok pada koordinat PC1 dengan rentang $0 \leq n \leq 3$, kemudian memperoleh label hari tidak hujan.

BAB V

KESIMPULAN

5.1 Kesimpulan

Penelitian ini membuktikan bahwa integrasi Analisis Komponen Utama (PCA) dengan *metode Clustering* merupakan pendekatan yang efektif dalam analisis pola curah hujan, di mana enam variabel klimatologi berhasil direduksi menjadi dua komponen utama (PC1 dan PC2), yang mampu menjelaskan 73,33% variasi data. Berdasarkan ambang batas yang diterapkan pada sumbu PC1, titik data dikategorikan ke dalam dua cluster: Cluster 1 ($n < 0$) dan Cluster 2 ($0 \leq n \leq 3$), sesuai dengan hasil pengelompokan hari hujan dan tidak hujan. Proses pelabelan berbasis PCA terbukti konsisten dengan kategori resmi BMKG, dengan distribusi data menunjukkan pemisahan kluster yang jelas di ruang PC1 dan PC2. Validasi kuantitatif melalui *Skor Siluet* 0,55 menunjukkan kualitas pengelompokan yang cukup baik, dengan jarak intra-cluster dan antar-cluster yang terpisah. Temuan ini menegaskan bahwa PCA tidak hanya berfungsi sebagai teknik pengurangan dimensi, tetapi juga dapat digunakan sebagai dasar pembentukan label dan pengelompokan curah hujan, sehingga memberikan kontribusi metodologis terhadap analisis data klimatologis. Namun, penelitian ini masih terbatas pada jumlah data dari dua Lokasi stasiun klimatologi pos hujan dalam satu tahun pengamatan, sehingga penelitian lebih lanjut dapat memperluas cakupan data spasial dan temporal dengan melibatkan lebih banyak stasiun dan periode pengamatan yang lebih lama, mengintegrasikan *PCA-Clustering* dengan model prediktif berbasis *machine learning* dan *deep learning* (misalnya LSTM atau CNN) untuk analisis pola curah

hujan yang lebih dinamis, mengeksplorasi variabel klimatologis tambahan seperti kelembaban, arah angin, dan indeks iklim global (ENSO, IOD) untuk mendapatkan representasi yang lebih komprehensif, dan mengembangkan visualisasi interaktif berdasarkan *Augmented Reality* (AR) atau sistem informasi geografis (GIS) untuk membuat hasil analisis lebih mudah digunakan dalam pengambilan keputusan mitigasi bencana hidrometeorologi.

5.2 Saran

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

1. Memperluas cakupan data secara spasial dan temporal dengan melibatkan lebih banyak stasiun pengamatan serta periode observasi yang lebih panjang.
2. Mengintegrasikan metode *PCA-Clustering* dengan model prediktif berbasis machine learning atau deep learning, seperti LSTM maupun CNN, untuk analisis pola curah hujan yang lebih dinamis.
3. Menambahkan variabel klimatologi lain, misalnya kelembapan udara, arah angin, serta indeks iklim global (ENSO, IOD), guna memperoleh representasi yang lebih komprehensif.
4. Mengembangkan media visualisasi interaktif berbasis Augmented Reality (AR) atau Sistem Informasi Geografis (SIG) agar hasil analisis lebih mudah dimanfaatkan dalam pengambilan keputusan mitigasi bencana hidrometeorologi.

DAFTAR PUSTAKA

Alaziz, S. N., Alshowiman, A. A., Albayati, B., El-Bagoury, A. al A. H., & Shafik, W. (2023). *Clustering* of COVID-19 Multi-Time Series-Based K-Means and PCA With Forecasting. *International Journal of Data Warehousing and Mining*, *19*(3). https://doi.org/10.4018/IJDWM.317374

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., Parwatiningtyas, D., Supriyono, & Wicaksono, H. (2023a). Principal Component Analysis-Based Data *Clustering* for Labeling of Level Damage Sector in Post-Natural Disasters. *IEEE Access*, *11*, 74590–74601. https://doi.org/10.1109/ACCESS.2023.3275852

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., Parwatiningtyas, D., Supriyono, & Wicaksono, H. (2023b). Principal Component Analysis-Based Data *Clustering* for Labeling of Level Damage Sector in Post-Natural Disasters. *IEEE Access*, *11*(March), 74590–74601. https://doi.org/10.1109/ACCESS.2023.3275852

Almais, A. T. W., Susilo, A., Naba, A., Sarosa, M., Juwono, A. M., Crysdian, C., Muslim, M. A., & Wicaksono, H. (2024). Characterization of Structural Building Damage in Post-Disaster using GLCM-PCA Analysis Integration. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2024.3469637

Arisandi, R., Ruhiat, D., & Marlina, E. (2021). Implementasi ridge regression untuk mengatasi gejala multikolinearitas pada pemodelan curah hujan berbasis data

time series klimatologi. *JRMST| Jurnal Riset ...*, *1*(November), 1–11. https://ejournal.unibba.ac.id/index.php/jrmst/article/view/735%0Ahttps://ejournal.unibba.ac.id/index.php/jrmst/article/download/735/666

Basile, F., & Ferrara, L. (2023). Validation of Industrial Automation Systems Using a Timed Model of System Requirements. *IEEE Transactions on Control Systems Technology*, *31*(1), 130–143. https://doi.org/10.1109/TCST.2022.3173890

Chen, Y., Tan, P., Li, M., Yin, H., & Tang, R. (2024). K-means *Clustering* method based on nearest-neighbor density matrix for customer electricity behavior analysis. *International Journal of Electrical Power and Energy Systems*, *161*(January). https://doi.org/10.1016/j.ijepes.2024.110165

Darlan, N. H., Arif, S. S., Sudira, P., & Nugroho, B. D. A. (2020). Spatial and Temporal Analysis of Seasonal Rainfall on the East Coast of North Sumatra, Indonesia. *Indonesian Journal of Geography*, *52*(3), 360. https://doi.org/10.22146/ijg.56724

Darmawan, R. A., Priyono, I., & Taufiq, A. (2022). *Proceedings pit iagi 51*.

Hendrawati, T., Wigena, A. H., Sumertajaya, I. M., Sartono, B., Pravitasari, A. A., & Asnawi, M. H. (2024). The ensemble distance on model-based *Clustering* for regions *Clustering* based on rainfall: The case of rainfall in West Java Indonesia. *International Journal of Data and Network Science*, *8*(2), 1187–1196. https://doi.org/10.5267/j.ijdns.2023.11.015

Hirvonen, K., Machado, E. A., Simons, A. M., & Taraz, V. (2022). More than a safety net: Ethiopia's flagship public works program increases tree cover. *Global Environmental Change*, *75*(June), 102549.

https://doi.org/10.1016/j.gloenvcha.2022.102549

Huang, F., Xia, J., Yin, C., Zhai, X., Xu, N., Yang, G., Bai, W., Sun, Y., Du, Q., Liao, M., Hu, X., Zhang, P., Duan, L., & Liu, Y. (2022). Assessment of FY-3E GNOS-II GNSS-R Global Wind Product. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 7899–7912. https://doi.org/10.1109/JSTARS.2022.3205331

Irawan, B., Wirawan, Ikawanty, B. A., & Takwim, A. (2022). Analysis of the Season Effect on Energy Generated From Hybrid Pv/Wt in Malang Indonesia. *Eastern-European Journal of Enterprise Technologies*, *5*(8–119), 70–78. https://doi.org/10.15587/1729-4061.2022.266082

J.M.A.U. Jayasekara, M. I. M. M. and N. D. K. D. (2024). *Analysis of Spatial and Temporal Variation of Water Quality : A Case Study of Kotagala Wetland , Nuwara Eliya , Sri Lanka*. *35*, 343–354.

Jayasekara, J. M. A. U., Mowjood, M. I. M., & Dayawansa, N. D. K. (2024). Analysis of Spatial and Temporal Variation of Water Quality : A Case Study of Kotagala Wetland , Nuwara Eliya , Sri Lanka. *Sri Lanka Journals Online*, *35*, 343–354.

Karamma, R. (2025). *Flood Risk Assessment and Mitigation Strategies for the Sinjai and Tangka River Catchments in Indonesia using Hydraulic Modeling and Spatial Analysis*. *15*(2), 20623–20634.

Khan, M. I., & Maity, R. (2020). Hybrid Deep Learning Approach for Multi-Step-Ahead Daily Rainfall Prediction Using GCM Simulations. *IEEE Access*, *8*(Ml), 52774–52784. https://doi.org/10.1109/ACCESS.2020.2980977

Libasin, Z., Fauzi, W. S. W. M., Ul-Saufie, A. Z., Idris, N. A., & Mazeni, N. A.

(2021). Evaluation of single missing value imputation techniques for incomplete air particulates matter (Pm10) data in Malaysia. *Pertanika Journal of Science and Technology*, *29*(4), 3099–3112. https://doi.org/10.47836/PJST.29.4.46

Lima, A. O., Lyra, G. B., Abreu, M. C., Oliveira-Júnior, J. F., Zeri, M., & Cunha-Zeri, G. (2021). Extreme rainfall events over Rio de Janeiro State, Brazil: Characterization using probability distribution functions and *Clustering* analysis. *Atmospheric Research*, *247*. https://doi.org/10.1016/j.atmosres.2020.105221

Pamuji, G. C., & Rongtao, H. (2020). A Comparison study of DBScan and K-Means *Clustering* in Jakarta rainfall based on the Tropical Rainfall Measuring Mission (TRMM) 1998-2007. *IOP Conference Series: Materials Science and Engineering*, *879*(1). https://doi.org/10.1088/1757-899X/879/1/012057

Pansera, W. A., Gomes, B. M., Vilas Boas, M. A., & de Mello, E. L. (2013). *Clustering* rainfall stations aiming regional frequency analysis. *Journal of Food, Agriculture and Environment*, *11*(2), 877–885.

Punhani, A., Faujdar, N., Mishra, K. K., & Subramanian, M. (2022). Binning-Based Silhouette Approach to Find the Optimal Cluster Using K-Means. *IEEE Access*, *10*(November), 115025–115032. https://doi.org/10.1109/ACCESS.2022.3215568

Rachmawati, R. N. (2021). Estimation of Extreme Rainfall Patterns Using Generalized Linear Mixed Model for Spatio-temporal data in West Java, Indonesia. *Procedia Computer Science*, *179*(2020), 330–336. https://doi.org/10.1016/j.procs.2021.01.013

Rahman, F. A., Kassim, R., Baharum, Z., Noor, H. A. M., & Haris, N. A. (2019). Data Cleaning in Knowledge Discovery Database-Data Mining (KDD-DM). *International Journal of Engineering and Advanced Technology*, *8*(6s3), 2196–2199. https://doi.org/10.35940/ijeat.f1100.0986s319

Shantal, M., Othman, Z., & Bakar, A. A. (2023). A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. *Symmetry*, *15*(12). https://doi.org/10.3390/sym15122185

Sofro, A., Riani, R. A., Khikmah, K. N., Romadhonia, R. W., & Ariyanto, D. (2024). Analysis of Rainfall in Indonesia Using a Time Series-Based *Clustering* Approach. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *18*(2), 0837–0848. https://doi.org/10.30598/barekengvol18iss2pp0837-0848

Suharyanto, A., Maulana, A., Suprayogo, D., Devia, Y. P., & Kurniawan, S. (2023). Land surface temperature changes caused by land cover/ land use properties and their impact on rainfall characteristics. *Global Journal of Environmental Science and Management*, *9*(3), 353–372. https://doi.org/10.22035/gjesm.2023.03.01

Thabet, A., Gasmi, N., Frej, G. B. H., & Boutayeb, M. (2021). Sliding Mode Control for Lipschitz Nonlinear Systems in Reciprocal State Space: Synthesis and Experimental Validation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *68*(3), 948–952. https://doi.org/10.1109/TCSII.2020.3018016

Tripathi, P., & Garg, R. D. (2021). Comparative Analysis of Singular Value Decomposition and Eigen Value Decomposition Based Principal Component Analysis for Earth and Lunar Hyperspectral Image. *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing*

(WHISPERS), 1–5. https://doi.org/10.1109/WHISPERS52202.2021.9483978

Urgilés, G., Célleri, R., Bendix, J., & Orellana-Alvear, J. (2024). Identification of spatio-temporal patterns in extreme rainfall events in the Tropical Andes: A *Clustering* analysis approach. *Meteorological Applications*, *31*(5). https://doi.org/10.1002/met.70005

Uykan, Z. (2023). Fusion of Centroid-Based *Clustering* With Graph *Clustering*: An Expectation-Maximization-Based Hybrid *Clustering*. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 4068–4082. https://doi.org/10.1109/TNNLS.2021.3121224

Worku, M. A., Feyisa, G. L., Beketie, K. T., & Garbolino, E. (2022). Rainfall variability and trends in the Borana zone of southern Ethiopia. *Journal of Water and Climate Change*, *13*(8), 3132–3151. https://doi.org/10.2166/wcc.2022.173

Yan, Y., Le, X., Yang, T., & Yu, H. (2024). Interpretable PCA and SVM-Based Leak Detection Algorithm for Identifying Water Leakage Using SAR-Derived Moisture Content and InSAR Closure Phase. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 15136–15147. https://doi.org/10.1109/JSTARS.2024.3443127