K-Means Clustering Indonesia's Climate

Identifikasi Pola dan Integrasi Peta Peringatan Dini Curah Hujan Tinggi dan Kekeringan Meteorologis

Final Project for MSIB Data Analyst at Kampus Merdeka
Kemendikbud x GreatEdu Batch 5

GreenEnviro

SIB GREATEDU CYCLE 5

Profile Team

Radif Ramadan
Project Lead

Universitas Singaperbangsa Karawang

Mars Adefa
Analyst

Universitas Negeri Jakarta

Heres Ikhsanurijal
Analyst

Universitas Brawijaya

Eka Putri Lestari
Visualization

Universitas Jenderal Soedirman

Sarah Fadilla
Visualization

Institut Bisnis dan Informatika Kosgoro

SIB GREATEDU CYCLE 5

Table Of Content

Business Understanding

Indonesia, dengan iklim yang beragam, sering menghadapi tantangan bencana yang signifikan akibat perubahan iklim. Indonesia juga, memiliki sejumlah stasiun pengamatan di setiap provinsinya yang mencatat data iklim termasuk suhu, curah hujan, kecepatan angin, penyinaran matahari, dan kelembaban udara.

Tim kami akan menganalisis **Pola Cuaca** berdasarkan **data kuantitatif** tahun 2010-2020. Langkah ini diambil untuk mendukung pemerintah dalam menyusun **peringatan dini** dan strategi **mitigasi**, sehingga dapat mengurangi risiko bencana serta dampak negatif pada lingkungan.

Business Understanding

1 Purpose

- Mengetahui Pola Karakteristik Cuaca di Wilayah Indonesia Menggunakan Cluster.
- Membuat Peta Peringatan Dini Curah Hujan Tinggi dan Kekeringan Meteorologis di setiap wilayah Indonesia

Business Understanding

2 Benefit

- Menyediakan Peta Peringatan Dini dan Rekomendasi Mitigasi Bencana
- Mengurangi Resiko Kerugian Ekonomi dan Sosial

Data Understanding

1 Sumber Data

- Sumber dataset berasal dari kaggle
  https://www.kaggle.com/datasets/greegtitan/indonesia-climate/data
- Dataset berisi 3 tabel yang berisi climate_data, station_detail dan province_detail
- Data yang digunakan untuk analisis adalah tabel climate_data dengan **589.265 baris** dan **12 kolom**.

Data Understanding

2 Rename Kolom

Data yang digunakan akan direname untuk memudahkan pembacaan kolom data

- ‘date’ → ‘Date’
- ‘Tn’ → ‘MinTemp’
- ‘Tx’ → ‘MaxTemp’
- ‘Tavg’ → ‘AvgTemp’
- ‘RH_avg’ → ‘AvgHum’
- ‘RR’ → ‘RainFall’
- ‘ss’ → ‘’DurationSunshine’
- ‘ff_x’ → ‘MaxWindSpeed
- ‘ddd_x’ → ‘WindDirectionMaxSpeed’
- ‘ff_avg’ → ‘AvgWindSpeed’
- ‘ddd_car’ → ‘MostWindDirection’

Data Understanding

3 Deskripsi Data

Column	Dtype	Description
MinTemp	float64	min temperature (°C)
MaxTemp	float64	max temperature (°C)
AvgTemp	float64	average temperature (°C)
AvgHum	float64	average humidity (%)
RainFall	float64	rainfall (mm)
DurationSunshine	float64	duration of sunshine (hour)

Data Understanding

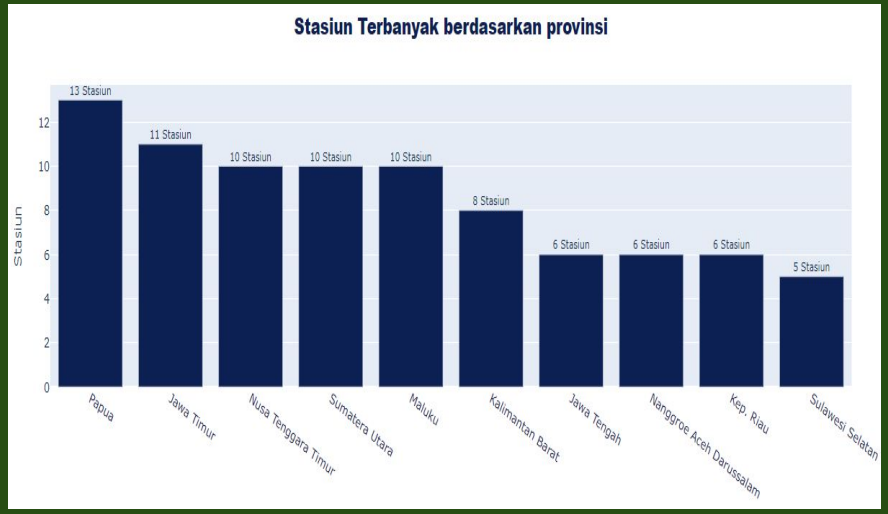
3 Deskripsi Data

Column	Dtype	Description
MaxWindSpeed	float64	max wind speed (m/s)
WindDirectionMaxSpeed	float64	wind direction at maximum speed (°)
AvgWindSpeed	float64	average wind speed (m/s)
MostWindDirection	object	most wind direction (°)
MaxWindSpeed	float64	max wind speed (m/s)
station_id	object	number of station

Data Understanding

4 Exploratory Data Analysis

Stasiun Terbanyak berdasarkan Provinsi



- Gambar di samping menunjukkan **10 provinsi** tertinggi yang memiliki stasiun cuaca paling banyak di Indonesia
- **Papua** menjadi provinsi dengan stasiun cuaca terbanyak yaitu **13 stasiun**

Data Understanding

4 Exploratory Data Analysis

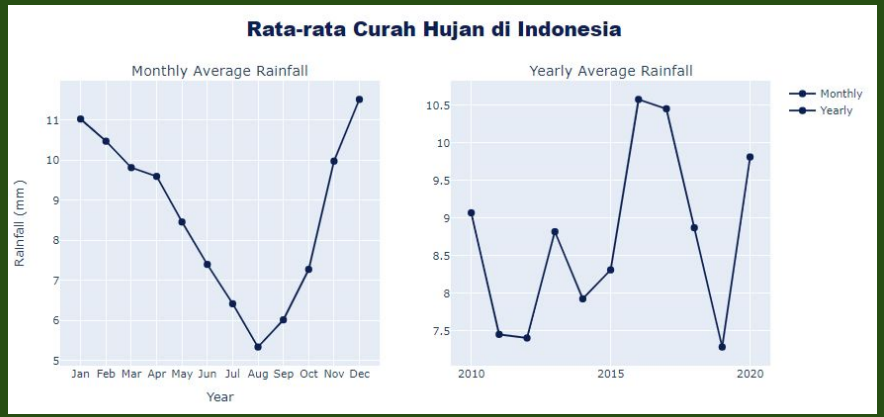
Stasiun Pengamatan Terbanyak berdasarkan Provinsi

- Gambar tersebut menunjukkan **5 stasiun cuaca** yang paling banyak melakukan pengamatan di Indonesia
- **Stasiun Meteorologi Aek Godang** menjadi stasiun yang paling banyak melakukan pengamatan dengan **4005** kali pengamatan

Data Understanding

4 Exploratory Data Analysis

Rata-rata Curah Hujan di Indonesia Bulanan dan Tahunan

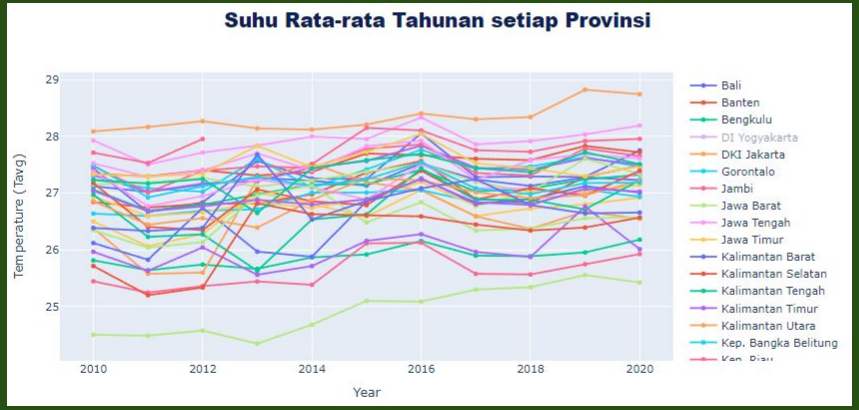


- Gambar di samping menunjukkan **rata-rata curah hujan di Indonesia** berdasarkan bulanan dan tahunan
- Berdasarkan bulanan, curah hujan tertinggi terjadi pada bulan **Desember**, sedangkan terendah pada bulan **Agustus**
- Berdasarkan tahunan dari 2010-2020, curah hujan tertinggi terjadi pada tahun **2016**, sedangkan terendah pada tahun **2019**

Data Understanding

4 Exploratory Data Analysis

Suhu Rata-rata Tahunan tiap Provinsi

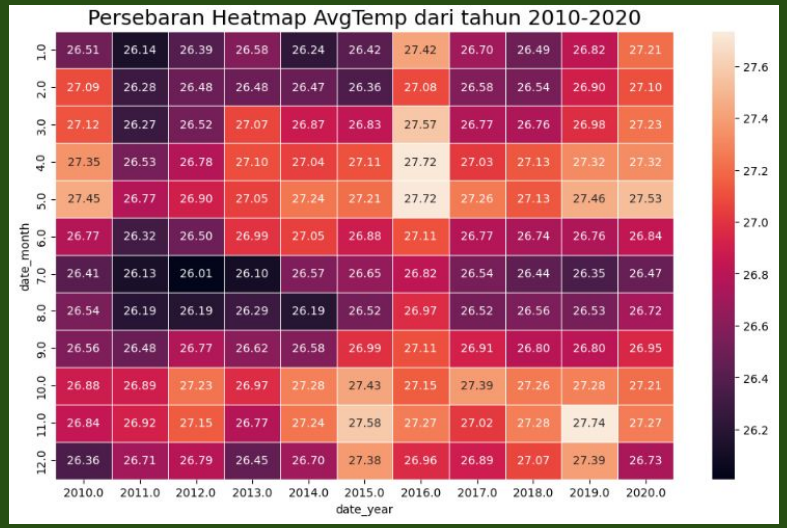


- Gambar di samping menunjukkan **suhu rata-rata tahunan tiap provinsi** di Indonesia dalam kurun waktu 2010-2020
- Suhu rata-rata tertinggi ada pada tahun **2019 di Provinsi DKI Jakarta**
- Suhu rata-rata terendah ada pada tahun **2013 di Provinsi Jawa Barat**

Data Understanding

4 Exploratory Data Analysis

Heatmap Persebaran AvgTemp



- Gambar di samping menunjukkan **Heatmap persebaran AvgTemp** di Indonesia dalam kurun waktu 2010 - 2020
- Adanya peningkatan suhu rata-rata pada tahun **2016** dari awal tahun hingga pertengahan tahun
- Adanya penurunan suhu rata-rata pada tahun **2011** dari awal tahun hingga pertengahan tahun

Data Understanding

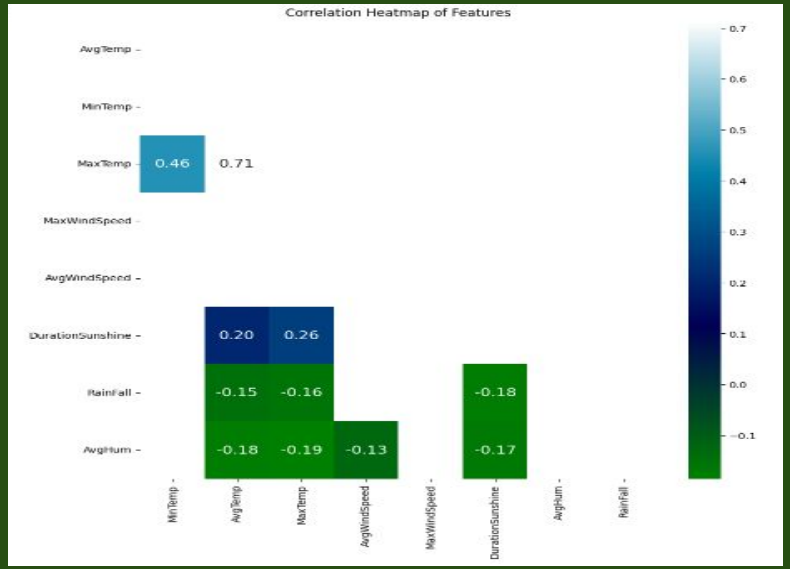
4 Exploratory Data Analysis

Rata-rata Lama Sinar Matahari Bulanan di Indonesia

- Gambar di samping menunjukkan **rata-rata durasi sinar matahari bulanan** di Indonesia
- Rata-rata durasi sinar matahari terlama ada pada bulan **Agustus**, sedangkan yang tersingkat ada pada bulan **Desember**

Data Understanding

5 Correlation

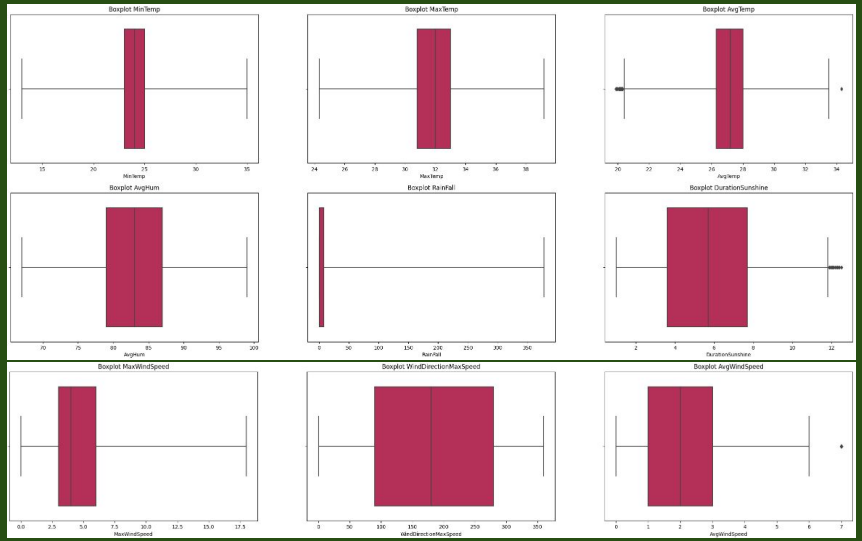


- Gambar di samping menunjukkan **korelasi antar variabel** yang ada dalam dataset
- Korelasi **positif** terkuat terjadi antara **MaxTemp** dengan **AvgTemp**, artinya semakin **tinggi** maximum temperaturnya maka semakin **tinggi** pula rata-rata temperaturnya begitupun sebaliknya
- Korelasi **negatif** terkuat terjadi antara **AvgHum** dengan **MaxTemp**, artinya semakin **tinggi** rata-rata kelembaban maka semakin **rendah** maximum temperaturnya begitupun sebaliknya

Data Preparation

1 Handling Data Quality

Handling Outlier



Setiap data yang **melebihi** batas bawah dan batas atas **dihapus** untuk menunjukan kualitas data yang lebih baik dan jauh dari **outlier**

Akibat dihapusnya beberapa data, maka jumlah baris berubah dari awalnya **589,265** menjadi **307,584 baris**

Data Preparation

1 Handling Data Quality

Handling Missing Values

Column	Sebelum diatasi	Setelah diatasi
MinTemp	0.039682	0
MaxTemp	0.064039	0
AvgTemp	0.076545	0
AvgHum	0.081766	0
RainFall	0.212780	0

Column	Sebelum diatasi	Setelah diatasi
DurationSunshine	0.074196	0
MaxWindSpeed	0.017333	0
WindDirectionMaxSpeed	0.022279	0
AvgWindSpeed	0.017186	0
MostWindDirection	0.023315	0

Setiap **data hilang** pada suatu kolom di **imputasi** dengan **median** untuk data **numerik** dan **mode** untuk data **kategori**. Sehingga tidak ada lagi data yang hilang dalam dataset

Data Preparation

2 Label Encoding

Dikarenakan adanya data **kategorik,** yaitu pada kolom**MostWindDirection** dan **WindDirectionMaxSpeed** maka diperlukan transformasi data menggunakan **replace()** berdasarkan tabel di bawah ;

MostWindDirection	WindDirectionMaxSpeed	Notasi
C	-	0
N	0	1
NE	$0 < X < 90$	2
E	90	3
SE	$90 < X < 180$	4

MostWindDirection	WindDirectionMaxSpeed	Notasi
S	180	5
SW	$180 < X < 270$	6
W	270	7
Nw	$270 < X < 360$	8

Data Preparation

3 Feature Selection

Data yang dipergunakan tidak melibatkan semua kolom yang ada dalam dataset. Maka beberapa variabel yang tidak terpakai akan dihapus, sementara beberapa kolom lainnya akan digunakan untuk proses modelling.

Kolom yang **dihapus** :

- ‘Date’
- ‘station_id’

Kolom yang **digunakan** :

1. ‘MinTemp’
2. ‘MaxTemp’
3. ‘AvgTemp’
4. ‘AvgHum’
5. ‘RainFall’
6. ‘DurationSunshine’

Kolom yang **digunakan** :

7. ‘MaxWindSpeed’
8. ‘WindDirectionMaxSpeed
9. ‘AvgWindSpeed’
10. ‘MostWindDirection’

Modelling

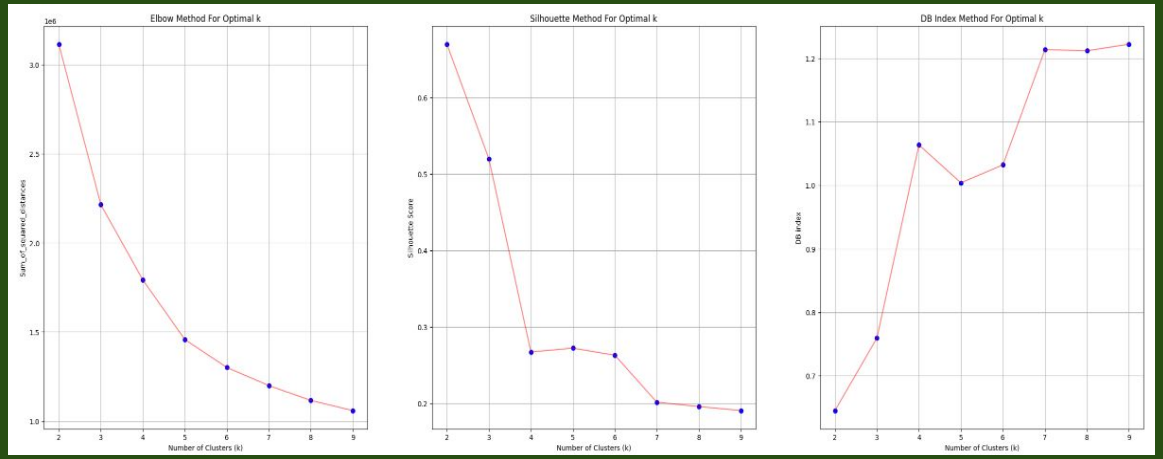
1 Model Selection

Algoritma yang digunakan adalah **K-Means** untuk mengelompokan data berdasarkan variabel yang telah dipilih dengan jumlah cluster yang kita tentukan, dimana model K-Means relatif cocok digunakan dengan jumlah data yang besar.

Yang menarik dari metode ini adalah kita dapat mengetahui jumlah cluster yang diinginkan menggunakan **Elbow method**, **Silhouette method** dan **Davies-Bouldin Index method**. Yang akan dibuat dengan grafik berdasarkan masing masing skornya pada slide berikutnya.

Modelling

2 Number of Cluster Optimal



	n_clusters	WSS	Davies-Bouldin Index	Silhouette Score
0	2	3.115949e+06	0.644922	0.669721
1	3	2.214888e+06	0.759274	0.519795
2	4	1.792331e+06	1.063683	0.267026
3	5	1.457719e+06	1.003866	0.272115
4	6	1.300602e+06	1.032192	0.262712
5	7	1.199046e+06	1.213849	0.201317
6	8	1.115512e+06	1.211940	0.195669
7	9	1.058453e+06	1.222471	0.190375

Berdasarkan analisis penentuan jumlah cluster menggunakan metode-metode yang sebelumnya dijelaskan didapatkan bahwa jumlah cluster optimal adalah dengan **3 cluster**.

Modelling

3 Centroid of Cluster

Dari Analisis **Centroid** pada setiap cluster didapatkan Pola Karakteristik pada setiap clusternya sesuai tabel disamping, sebagai berikut :

- **Cluster 1**, Jadi pada cluster ini merupakan cluster dengan Pola Cuaca pada **Musim Panas (Kemarau)**.
- **Cluster 2**, Jadi pada cluster ini merupakan cluster dengan Pola Cuaca pada **Musim Biasa (Hujan Ringan)** .
- **Cluster 3**, Jadi pada cluster ini merupakan cluster dengan Pola Cuaca pada **Musim Hujan Lebat (Ekstrem)** .

Fitur	Cluster 1	Cluster 2	Cluster 3
MinTemp	23.588155	23.223849	23.032201
MaxTemp	31.826497	31.155368	30.997985
AvgTemp	27.107995	26.428375	26.288341
AvgHum	81.865179	86.507491	87.250053
RainFall	2.012336	26.869062	75.803250
DurationSunshine	5.845377	4.609563	4.612513
MaxWindSpeed	4.816898	4.489766	4.483963
WindDirectionMaxSpeed	5.169071	5.498988	5.508652
AvgWindSpeed	1.984603	1.647312	1.653091
MostWindDirection	3.233223	3.052623	3.023106

Modelling

4 Feature Importance

Berikutnya adalah melihat **fitur penting** setiap variabel berdasarkan clusternya, hal ini untuk mendapatkan informasi variabel yang menyebabkan terbentuknya cluster pada analisis **Pola Karakteristik**.

- Berdasarkan hasil yang didapat bahwa **RainFall** dan **AvgHum** memiliki nilai yang lebih besar dibandingkan lainnya, hal ini menggambarkan bahwa kedua variabel tersebut yang menyebabkan terbentuknya cluster pada **Pola Karakteristik** Cuaca.

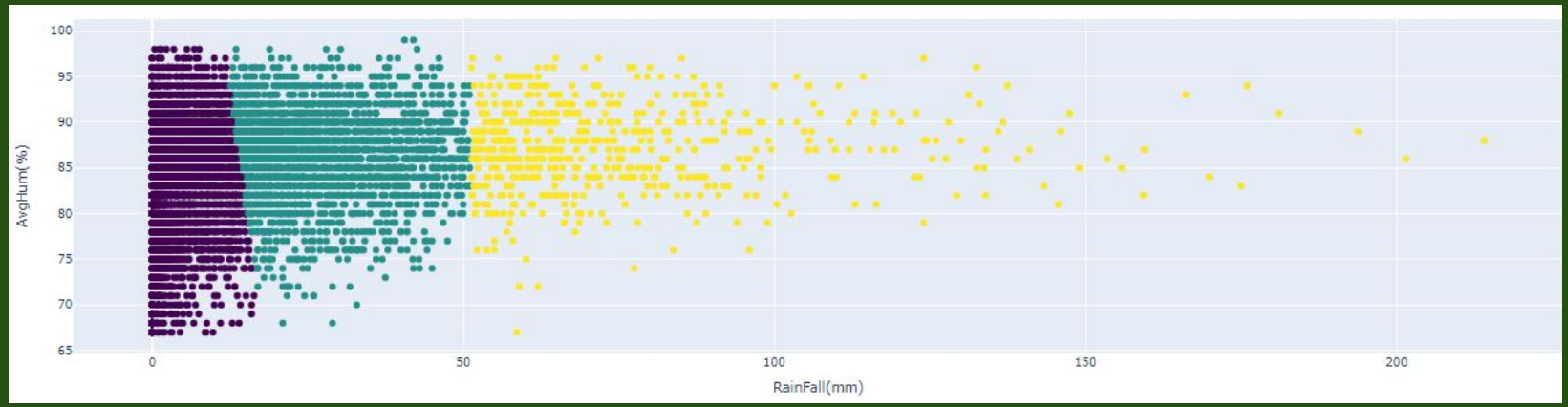
- Selanjutnya kedua variabel tersebut akan digunakan untuk visualisasi **persebaran** cluster yang didapat

Fitur	Nilai
RainFall	30.944399
AvgHum	3.068474
DurationSunshine	0.752103
MaxTemp	0.459773
AvgTemp	0.459397
MinTemp	0.284122
WindDirectionMaxSpeed	0.204194
AvgWindSpeed	0.203653
MaxWindSpeed	0.201564
MostWindDirection	0.119232

Modelling

5 Visualization Cluster

Scatter Plot



Modelling

5 Visualization Cluster

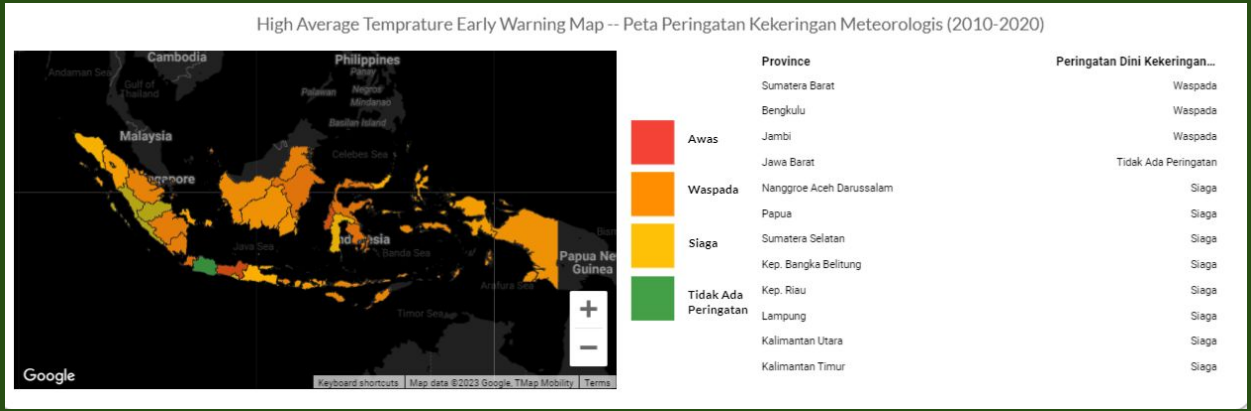
Peta Peringatan Dini Curah Hujan Tinggi

Setelah membuat visualisasi cluster, selanjutnya adalah membuat **Peta Peringatan Dini** menggunakan Looker dari hasil dari Rata-rata Curah Hujan selama 10 tahun.

Modelling

5 Visualization Cluster

Peta Peringatan Dini Kekeringan Metorologis



Setelah membuat visualisasi cluster, selanjutnya adalah membuat **Peta Peringatan Dini** menggunakan Looker dari hasil dari Rata-rata Temperatur selama 10 tahun.

Evaluation

1. **Keterkaitan Analisis terhadap SDGs**

- **SDG 13 - Tindakan terhadap Perubahan Iklim**

  Analisis Pola Karakteristik dan Peringatan dini berkaitan dengan cuaca ekstrem, baik curah hujan tinggi atau kekeringan. Hal ini dapat **membantu** pemerintah atau pemangku kepentingan lainnya dalam **mengambil tindakan** adaptasi terhadap perubahan iklim

- **SDG 11 - Kota dan Permukiman Berkelanjutan**

  Informasi yang didapat baik curah hujan tinggi atau kekeringan dapat **membantu perencanaan** kota dan permukiman terhadap cuaca ekstrem dan **mengurangi resiko** bencana di wilayah perkotaan dan pedesaan.

Evaluation

2. **Alasan Apakah Analisa Mendukung atau Bertentangan dengan Lingkungan**

   Akhir dari analisa ini tentu mendukung tujuan lingkungan. Hal ini dibuktikan dengan informasi ini akan membantu dalam **mengambil tindakan** adaptasi perubahan iklim, serta dapat **mereduksi** risiko bencana dan dampak lingkungan.

3. **Scoring Model**

   Model **K-means** yang digunakan dengan menggunakan 3 cluster didapatkan **0.519** untuk **Silhouette Score** dan **0.759** untuk **Davies-Bouldin Index**.

Recommendation

Tim kami, **GreenEnviro**, akan merekomendasikan target dan mitigasi bencana berdasarkan analisis clustering wilayah dan Peta Peringatan Dini yang telah dilakukan menggunakan algoritma K-Means.

Recommendation

Berdasarkan, analisis cluster yang telah tim kami buat.

Cluster-cluster tersebut dapat dibuat **rekomendasi mitigasi** yang dapat membantu dalam mengambil tindakan adaptasi terhadap perubahan iklim, serta dapat mereduksi risiko bencana dan dampak lingkungan.

Cluster	Recommendation
Cluster 1 (Musim Kemarau)	Dapat dilakukan mitigasi yang melibatkan : • pengelolaan air yang lebih baik, • penyuluhan konservasi air, • penghijauan • persiapan kebutuhan air yang lebih tinggi untuk menghindari kekeringan dan krisis air.
Cluster 2 (Musim Hujan Ringan)	Dapat dilakukan mitigasi dalam • mempersiapkan penanggulangan dengan memperbaiki sistem drainase • memonitor kondisi tanah serapan untuk menghindari potensi banjir dan tanah longsor.
Cluster 3 (Musim Hujan Lebat)	Dapat melakukan mitigasi yang lebih serius, hal ini dapat dilakukan dengan • mengevakuasi dan membuat rencana tanggap darurat yang disosialisasikan pada masyarakat. • mempersiapkan Infrastruktur dan sistem perairan yang diperkuat untuk menghindari banjir bandang.

Appendix

- https://www.kaggle.com/datasets/greegtitan/indonesia-climate/data
- https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/4810
- https://iklim.bmkg.go.id/id/
- https://jdpb.bnpb.go.id/index.php/jurnal/article/view/91/91
- https://shorturl.at/fpLT4
- https://journal.unnes.ac.id/nju/index.php/JG/article/view/17136
- https://medium.com/@MrBam44/how-to-evaluate-the-performance-of-clustering-algorithms-3ba29cad8c03
- https://chat.openai.com

Thank You!