OPEN ACCESS
J. OF ICT, VOL. 5, NO. 1, PP.001-010, JUN 2023

ISSN : 2686-1089 (ONLINE)
DOI: 10.52661

Journal of Informatics and Communications Technology (JICT)

# *Clustering* Data Cuaca Ekstrim Indonesia dengan *K-Means* dan Entropi

Ahmad Chusyairi [#1]

*# Ilmu Komputer, IPB University, Universitas Bina Insani*
*Kampus IPB Dramaga Bogor, Jl. Raya Dramaga, Babakan, Bogor, Jawa Barat*

[1] ahmadchusyairi@ipb.ac.id

**Abstract**

Weather is an important factor in all aspects of life. Significant weather changes affect human activities. Some activities related to the weather such as agricultural activities, plantations, and operations. As information technology develops in agriculture, more and more patterns of database systems are manual or computerized. However, the amount of data available does not always match the knowledge that can be generated. The community really needs weather information regardless of the format if it is reliable and valid. One is to determine local climate patterns. The data collection method uses a document study method, then uses data analysis techniques with a scoring system on rainfall data, maximum temperature, minimum temperature, average temperature, humidity, and wind speed. The k-means clustering method is a technique for grouping data. From the analysis carried out, there are 3 classes to cluster the weather levels produced by the entropy test to avoid bias towards non-optimal precision and accuracy. The division of the number of clusters can be captured as a type of potential vulnerability, namely high, medium, and mild where the total of all weather data from all BMKG stations throughout Indonesia is 123 data, 52 of which are classified as areas that have the potential to experience extreme high weather, 31 are classified as areas that experience extreme weather, moderate extreme weather potential and 40 are classified as areas with mild extreme weather potential.

**Keywords:** bmkg, clustering, entropy, extreme climate, k-means

## I. INTRODUCTION

Cuaca menjadi faktor penting dalam segala aspek kehidupan. Perubahan cuaca yang signifikan berpengaruh terhadap kegiatan manusia. Beberapa aktivitas yang berhubungan dengan cuaca seperti aktivitas pertanian, perkebunan dan penerbangan. Seiring berkembangnya teknologi informasi di bidang pertanian, semakin banyak pola sistem database yang bersifat manual atau terkomputerisasi namun jumlah data yang tersedia tidak selalu sesuai dengan pengetahuan yang dapat dihasilkan, misalnya, data yang ada hanya merupakan bagian dari arsip atau laporan, tetapi setelah diproses, data tersebut dapat memberikan wawasan yang sebelumnya tidak terpikirkan. masyarakat sangat membutuhkan informasi cuaca apapun formatnya jika dapat dipercaya dan valid, salah satunya adalah menentukan pola iklim setempat. Data cuaca ekstrim di Indonesia bersumber dari BMKG. *Clustering* untuk data cuaca yang bersifat ekstrim di Indonesia dengan *k-means* dan *entropi* bertujuan untuk memudahkan petani dalam menentukan waktu yang tepat dalam melakukan penanaman di bidang pertanian, sehingga meminimalisir kesalahan pada saat penanaman dimana cuacanya dalam keadaan ekstrim dan dengan adanya pengetahuan mengenai data cuaca yang ekstrim diharapkan dapat meningkatkan hasil produktivitas pertanian.

## II. LITERATURE REVIEW

*Data mining* adalah pendekatan yang digunakan dalam mengekstraksi data yang bermakna dari kumpulan data yang diberikan menggunakan satu atau lebih pendekatan seperti teknik statistik, pembelajaran mesin, matematika atau kecerdasan buatan, *data mining* juga didefinisikan sebagai proses pencarian pola yang menarik dan tersembunyi dalam kumpulan besar data yang disimpan dalam database seperti: gudang data dan area penyimpanan data lainnya. *Data mining* juga didefinisikan sebagai bagian dari proses penggalian pengetahuan dari database yang dikenal sebagai *Knowledge Discovery in Database* (KDD) [1].

Cuaca menjadi faktor penting dalam segala aspek kehidupan. Perubahan cuaca yang signifikan berpengaruh terhadap kegiatan manusia. Beberapa aktivitas yang berhubungan dengan cuaca seperti aktivitas pertanian, perkebunan dan penerbangan. Cuaca menjadi penting untuk dimonitor sehingga potensi seperti hujan deras, petir dapat dipersiapkan solusinya [2]. Dampak kondisi penyimpangan iklim yang nyatanya adalah meningkatnya intensitas curah hujan, banjir bandang, dan banjir pasang, badai lokal, suhu perkotaan naik, kekeringan dan tanah longsor [3].

Masyarakat sangat membutuhkan informasi cuaca apapun formatnya, asalkan terpercaya dan valid. Salah satunya adalah untuk menentukan pola iklim lokal, tetapi karena sulit untuk menilai secara manual dan siklus cuaca, diperlukan perangkat lunak yang dapat bertugas membantu memberi informasi dalam hal pengelompokan data meteorologi berdasarkan kesamaan data dengan metode *K-Means* yang memilih k jumlah awal *centroid* secara acak yang setiap titik ditempatkan pada klaster yang terdekat.

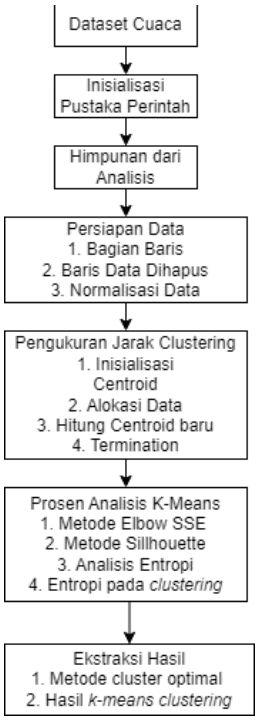
Teknik *data mining* yang dapat digunakan untuk mengelompokkan data cuaca berdasarkan kemiripan data menggunakan teknik *clustering*. Teknik *clustering* digunakan dalam data mining untuk mengelompokkan objek yang memiliki kemiripan pada kelas atau segmen yang sama, objek pada kelas yang berbeda memiliki sifat yang berbeda [4]. *Clustering* juga merupakan alat yang ampuh dalam analisis data, digunakan untuk menemukan struktur *cluster* dalam kumpulan data dengan kesamaan terbesar dalam *cluster* yang sama, tetapi perbedaan terbesar antara cluster yang berbeda. Umumnya, analisis klaster menjadi cabang dari analisis multivariat statistik, dan merupakan pendekatan *unsupervised learning* untuk *machine learning* [5].

Pengujian dengan entropi diperlukan, karena data yang diperoleh mengandung faktor ketidakteraturan atau keacakan dan paling sering diterapkan pada seluruh penelitian pada subjek terkait [6]. Konsep entropi memberikan wawasan mendalam ke arah perubahan spontan pada fenomena sehari-hari [7] salah satunya adalah cuaca. Sebagian besar penelitian terdahulu pada subjek terkait dilakukan dengan menggunakan cara otomatis yaitu memanfaatkan aplikasi Rapidminer salah satunya yang dilakukan oleh Khairunnisa [8], dan aplikasi Matlab oleh Sari [9] pada artikel ini, akan dibangun perangkat lunak menggunakan bahasa *python* dalam hal implementasi entropi permutasi dan K-means *clustering* bertujuan untuk menentukan pola dari kelas terkait pada data cuaca ekstrim dari seluruh stasiun BMKG di Indonesia dari awal hingga dapat mengeluarkan hasil.

## III. RESEARCH METHOD

Penelitian ini dilakukan dengan menggunakan data yang diperoleh pada tanggal 23 Oktober 2021 dari pusat basis data Badan Meteorologi Klimatologi dan Geofisika (https://dataonline.bmkg.go.id/), data di analisis menggunakan metode *K-Means Clustering* sebagai teknik untuk melakukan pengelompokan data. K-Means merupakan teknik cluster mengacu pada partisi dataset objek ke dalam kelompok objek yang paling mirip. Objek-objek penelitian ini adalah numerik [10]. Selanjutnya, prosedur klaster data didasarkan pada derajat keanggotaan masing-masing komponen [11]. Analisis ini dilakukan dengan menggunakan bahasa *Python* versi 3.10.8 dan perangkat lunak *Visual Studio Code* terintegrasi *Jupyter Notebook*.

Tahapan penelitian dilakukan sebagaimana yang tercantum dibawah ini:




Gambar 1. Tahapan Penelitian

1. Pada tahap dataset cuaca, data yang diperoleh (disebut sebagai *dataset*) cuaca ekstrim yang berisi data curah hujan (rr), tempratur maksimum (tx), tempratur minimum (tn), tempratur rata-rata (tAvg), kelembapan (rh), kecepatan angin (wx), dan tanggal pengambilan tiap informasi dari stasiun BMKG seluruh Indonesia yang diperoleh dari situs web BMKG (https://dataonline.bmkg.go.id/)
2. Tahap inisialisasi pustaka perintah yang diperlukan yaitu *Pandas, Folium, urllib, json,* dan *sklearn* akan diinstal dan dieksekusi.
3. Tahap himpunan analisis, ketika ada data yang dominan dibandingkan dengan yang lain, maka himpunan tertentu itu dibuat menjadi 1 kelompok dan dikeluarkan dari proses analisis.
   Data yang diperoleh pada tahap 1 dimuat lebih lanjut pada *Python* menggunakan *pandas*.

```
Import pandas as pd
nama_kolom : 'nama_stasiun', 'rr', 'tgl_pengamatan_ch', 'tx',
    'tgl_pengamatan_tx' , 'tn', 'tgl_pengamatan_tn' ,'tAvg'
    ,'tgl_pengamatan_tAvg' ,'rh' ,'tgl_pengamatan_rh' ,'wx'
    ,'tgl_pengamatan_wx'
dataset -> pd.read_csv('extreme-climate-dataset.csv', sep=';', header=None,
    skiprows=1, names=nama_kolom)
```

4. Persiapan Data
   a. Bagian baris adalah data observasi, kolom adalah variabel atau *features*.
   b. Baris data dihapus dengan ketentuan sebagai berikut:
      Untuk menghapus baris data yang tidak ada pada database lokasi stasiun yang bersumber dari BMKG (pengambilan data langsung bersumber menggunakan layanan GIS dari BMKG berbasis REST API (https://gis.bmkg.go.id), konsep dari alur ini adalah mendapatkan daftar stasiun lokasi BMKG termasuk koordinatnya yang tersimpan pada *database* BMKG, lalu melakukan pencarian pada daftar tersebut berdasarkan nama stasiun, jika tersedia, maka ambil data koordinat, jika tidak ada, maka baris data tersebut dihapus. Data mengandung nilai 8888 (tidak terukur), 9999 (data tidak terukur), dan data kosong (*null*).

```
dataset.dropna() untuk null
dataset.drop(dataset[dataset['tx'] == 8888 dan 9999].index, inplace=True)
```

c. Normalisasi data (penskalaan *MinMax*) dilakukan untuk membuat *feature* menjadi sebanding pada *feature* yang diperlukan saja.
Menggunakan fungsi dari *sklearn* yaitu:

```
X = dataset[['tn', 'tx','rr', 'tAvg', 'rh', 'wx']].values
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
```

Pada Rumus 1 menampilkan persamaan untuk menghilangkan dominasi atribut pada $X$ yang menempatkan data pada rentang 0 sebagai nilai minimal (*minValue*) hingga nilai 1 sebagai nilai maksimum (*maxValue*) [12].

$$X *= \frac{x - minValue}{maxValue - minValue} \quad (1)$$

5. Pengukuran jarak *clustering* dilakukan dengan menggunakan jarak *Euclidean* yang mana proses ini dijalankan dengan menggunakan fungsi *KMeans* turunan sklearn, adapun proses yang terjadi pada fungsi ini antara lain [12]:
   a. Inisialisasi *centroid*
   Wajib menentukan jumlah *cluster* (pada proses mendatang, nilai yang digunakan adalah nilai pada perulangan), proses ini akan secara acak menentukan *centroid* yang mana menggunakan rumus *mean* sebagai deskripsi nilai rata-rata.

$$\mu_1 = \frac{x_1 + x_2 + x_3 + \ldots + x_j}{j} \quad (2)$$

dengan $x_j$ adalah nilai objek dalam *cluster* dan jumlah $j$ dari objek dalam *cluster*, dan rumus untuk *single median* data sebagai berikut, jika datanya ganjil dan n adalah banyaknya data maka rumus median menjadi:

$$ME = \frac{X_{n+1}}{2} \quad (3)$$

Jika jumlah data genap maka rumus median menjadi:

$$ME = \frac{1}{2}\left(x\left(\frac{n}{2}\right) + x\left(\frac{n}{2} + 1\right)\right) \quad (4)$$

Median data dengan membagi yang terkecil dan terbesar nilai tambah dengan rumus:

$$X *= \frac{smallestvalue - largestvalue}{2} \quad (5)$$

   b. Alokasi data
   Setiap data atau objek akan ditempatkan pada *cluster* terdekat. Jarak antara dua benda menentukan kedekatan benda tersebut. Pengukuran jarak dengan jarak *Euclidean* adalah ukuran kedekatan yang paling umum digunakan, meskipun pengukuran lain seperti *Manhattan* ukuran dan koefisien *Jaccard* dapat digunakan. Jarak Euclidean ($d$) antara dua titik data $x(x_1, x_2, \ldots x_n)$ dan c($c_1, c_2, \ldots c_n$) dengan $n$ atribut dijelaskan dalam rumus Euclidean Distance sebagai berikut:

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ \ldots + (x_n - c_n)^n} \quad (6)$$

dengan $d$ adalah jarak titik $x$ dan $c, x_n$ adalah data kriteria dan $c_n$ adalah centroid dari cluster ke-$n$.
   c. Hitung *centroid* baru
   Untuk setiap *cluster*, setelah tahap alokasi data iterasi pertama selesai maka baru *centroid* dihitung untuk alokasi data iterasi berikutnya. *Centroid* baru ini adalah data yang paling representatif titik semua titik data dalam *cluster* diperoleh dari rata-rata *cluster*. Secara matematis, langkah ini dapat dinyatakan sebagai meminimalkan *Sum Squared Error* (SSE) dari semua titik data dalam sebuah *cluster* ke *cluster* pusat. Tujuan keseluruhan dari langkah ini adalah untuk meminimalkan SSE masing-masing kelompok. Ini rumus penentuan *centroid* baru:

$$\mu_1 = \frac{1}{2} \sum_{x \epsilon C_i} X \quad (7)$$

dengan $\mu_1$ adalah titik pusat (*centroid*), $X$ adalah objek di*cluster*, $C_i$ jumlah objek dalam *cluster*.

*d. Termination*
Langkah menghitung *centroid* baru, dan langkah untuk menetapkan titik data ke *centroid* baru adalah diulang sampai tidak ada lagi perubahan dalam tugas titik data terjadi. Dengan kata lain, tidak signifikan perubahan *centroid* dicatat. *Centroid* akhir dideklarasikan sebagai *cluster* prototipe dan digunakan untuk menggambarkan keseluruhan model pengelompokkan.

6. Proses analisis *K-means* guna mendapatkan nilai K yang optimal dapat digambarkan sebagai berikut:
   a. Metode *Elbow* SSE dengan rumus [12]:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_i - \mu_i||^2 \quad (8)$$

rumus SSE dengan $xj$ adalah objek pada *cluster ci* dan *centroid* dari *cluster*, yang mana dapat dilakukan dengan menjalankan kode sebagai berikut:

```
from scipy.stats import entropy
from kneed import KneeLocator

def calculate_entropy(labels):
    value,counts = np.unique(labels, return_counts=True)
    return entropy(counts)

for k in range(2, 10):
    kmeanModel = KMeans(n_clusters=k).fit(X_scaled)
    kmeanModel.fit(X_scaled)
    print('k =', k, 'entropy =', calculate_entropy(kmeanModel.labels_))

kmeans_kwargs = {
    "init": "random",
    "n_init": 10,
    "max_iter": 300,
    "random_state": 42,
}

# A list holds the SSE values for each k
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, **kmeans_kwargs)
    kmeans.fit(X_scaled)
    sse.append(kmeans.inertia_)

# plot sse against k
plt.plot(range(1, 11), sse)
plt.xticks(range(1, 11))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
# tick recommended k
kl = KneeLocator(range(1, 11), sse, curve="convex", direction="decreasing")
plt.vlines(kl.elbow, plt.ylim()[0], plt.ylim()[1], linestyles='dashed')
plt.show()
```

b. Metode *Sillhouette* dengan persamaan berikut [12]:

$$SI(k) = \frac{1}{k}\sum_{h=1}^{k} SI_h \quad (9)$$

$where,$

$$SI_h = \frac{1}{|C_h|}\sum_{i=1}^{|C_h|}\left[\frac{b_i^h - a_i^h}{\max\{a_i^h, b_i^h\}}\right] \quad (10)$$

$$a_i^h = \frac{1}{|C_h| - 1}\sum_{l=1, l\neq i}^{|C_h|} d\left(x_i^h, x_l^h\right) \qquad b_i^h = j \min \in \{1, \ldots, k \neq h\}, \frac{1}{|C_j|}\sum_{l=i}^{|C_h|} d\left(x_i^h, x_l^j\right) \quad (11)$$

dengan $x_i^h$dan $x_l^h$ adalah titik data ke-$i$ dan ke-$l$ pada cluster ke-$h$, $d(.)$ singkatan dari jarak *Euclidean*, $a_i^h$ menunjukkan jarak rata-ratadari titik data ke-$h$ ke titik data lain di *cluster* ke-$h$,dan $b_i^h$ menunjukkan jarak rata-rata dari titik ke- ke titik-titik di *cluster* lain. Nilai *Silhouette* berkisar antara -1 dan 1. Kemudian, jumlah *cluster*, yang sesuai dengan SI($k$) terbesar, adalah ditentukan sebagai $k$ nilai yang optimal. Untuk setiap titik, SI menghitung lebar mengenai keanggotaannya dalam *cluster* manapun. Lalu, lebar SI diperhitungkan sebagai rata-rata atas semua pengamatan, sehingga dapat dieksekusi dengan menjalankan kode berikut sekaligus menampilkan visual dari *Silhouette*:

```
silhouette = []
K = range(2, 10)
for k in K:
   kmeanModel = KMeans(n_clusters=k).fit(X_scaled)
   kmeanModel.fit(X_scaled)
   silhouette.append(silhouette_score(X_scaled, kmeanModel.labels_))

plt.plot(K, silhouette, 'bx-')
plt.xlabel('k')
plt.ylabel('Silhouette')
plt.title('The Silhouette Method showing the optimal k')
plt.show()
```

c. Analisis entropi
*Cluster* yang optimal dilihat dari visual garis (merupakan nilai k) yang turun drastis pada grafik visualisasi *Elbow* dan *Sillhouette plot.* Analisis entropi dengan persamaan berikut [13]:

$$\mathrm{H(X)} = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \quad (12)$$

Jika $X = \{x_i | i = 1, \ldots, n\}$ adalah fitur non-*numerik* atau label kelas, yaitu: entropi X.

d. Entropi pada *clustering*
Entropi pada *clustering* adalah jumlah bobot dari entropi klaster, di mana bobot untuk setiap klaster adalah probabilitas klaster, yang merupakan jumlah tupel dalam klaster dibagi dengan jumlah total tupel, hal ini dilakukan untuk mengetahui tingkat variabilitas ketidakteraturan suatu data dengan rumus, untuk melakukan kalkulasi entropi bisa dijalankan menggunakan kode berikut:

```
from scipy.stats import entropy

def calculate_entropy(labels):
   value,counts = np.unique(labels, return_counts=True)
   return entropy(counts)
```

7. Ekstraksi hasil
   a. Metode *cluster* optimal
      Berdasarkan pendekatan metode *cluster* optimal pada langkah sebelumnya, maka akan diperoleh *cluster* yang optimal. Jumlah cluster digunakan untuk menghitung nilai *clustering k-means*. Sebagai contoh, pada tahap sebelumnya, nilai dari k = 3 didapat, maka:

```
kmeans = KMeans(n_clusters=3, random_state=42)
y_kmeans = kmeans.fit_predict(X)
```

   b. Hasil *k-means clustering*
      Berdasarkan hasil tersebut akan diperoleh hasil *k-means clustering*. Hasil ini dapat divisualisasikan beserta peta persebarannya menggunakan kode:

```
peta_persebaran = folium.Map(location=[-2.5, 118], zoom_start=5)

color = ['red', ‘yellow’, 'green']

# add cluster
for i in range(0, len(y_kmeans)):
    clusternya = y_kmeans[i]
    folium.CircleMarker(
        location=[lat[i], lon[i]],
        radius=13,
        color=color[clusternya],
        fill=True,
        fill_color=color[clusternya],
        fill_opacity=0.7,
        popup='Cluster ' + str(y_kmeans[i]) + ' ' +
str(dataset['nama_stasiun'].iloc[i])
    ).add_to(peta_persebaran)
#tampilkan peta persebaran dengan memanggil variabel
peta_persebaran
```

IV. RESULTS AND DISCUSSIONGambar 2 merupakan titik koordinat untuk lokasi stasiun yang menyediakan data iklim lengkap yang mana divisualisasikan pada bentuk peta, dapat dilihat bahwa hampir setiap provinsi memiliki stasiun pemantauan sendiri yang terdiri dari lebih dari satu. Adapun stasiun ini terbagi atas 4 jenis stasiun yaitu stasiun meteorologi, stasiun klimatologi, stasiun geofisika, stasiun *Global Atmosphere Watch* (GWA), dan Balai Besar MKG.

Gambar 2. Peta titik persebaran stasiun BMKG

Berdasarkan dari hasil analisis statistik deskriptif yang ditampilkan pada Tabel 1 dari 123 stasiun BMKG seluruh Indonesia, dapat diketahui bahwa curah hujan yang dialami pada sebagian daerah adalah 47 hingga 327

mm perhari, tempratur maksimal antara 27 hingga 38.4°C perhari, tempratur minimum antara 9.4 hingga 23°C perhari, tempratur rata-rata antara 21.5 hingga 33.6°C perhari, kelembapan antara 46 hingga 83% perhari, dan kecepatan angin antara 1 hingga 3 knot perhari. Curah hujan (rr) memiliki *mean* tertinggi, standar *deviation* tertinggi, nilai minimum tertinggi dan nilai maksimum tertinggi sedangkan kecepatan angin (w) memiliki *mean* terendah, standar *deviation* terendah, nilai minimum terendah dan nilai maksimum terendah.

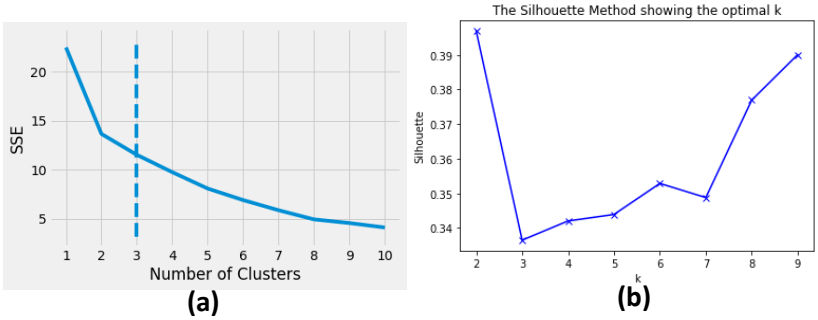
Tabel 1
Tabel statistik deskriptif data Iklim masing-masing daerah

| ***Feature*** | **Data terhitung** | ***Mean*** | ***Std. Deviation*** | ***Min*** | ***Max*** |
|---|---|---|---|---|---|
| rr (curah hujan) | 123 | 157 | 55.6 | 47 | 327 |
| tx (tempratur maksimal) | 123 | 38 | 26.9 | 27 | 38.4 |
| tn (tempratur minimum) | 123 | 19 | 2.4 | 9.4 | 23 |
| tAvg (tempratur rata-rata) | 123 | 29 | 1.8 | 21.5 | 33.6 |
| rh (kelembapan) | 123 | 65 | 7.4 | 46 | 83 |
| w (kecepatan angin) | 123 | 1.6 | 0.5 | 1 | 3 |

Penentuan jumlah optimal k grup ditentukan dengan menggunakan 2 (dua) pendekatan yang paling umum digunakan, yaitu *Elbow* dan *Silhouette*. Hasilnya dapat dilihat sebagaimana yang ditampilkan pada

Gambar *3* bagian a dan b. Berdasarkan hasil visualisasi yang ditampilkan oleh

**Gambar** *3*, metode *elbow* menunjukan nilai *k* yang optimal adalah *k* = 3 dikarenakan memiliki nilai k jatuh paling tajam gambar 2(a), sedangkan *silhouette* memberikan hasil yang sama yaitu menunjukan nilai *k* = 3 sebagai nilai k yang optimal dikarenakan nilai rata-ratanya semakin tinggi gambar 2(b), sehingga berdasarkan hasil dari kedua metode tersebut dapat disimpulkan bahwa nilai k yang optimal untuk membentuk *cluster* adalah tiga. Selanjutnya hasil analisis *clustering* menggunakan *K-means* dengan k = 3.



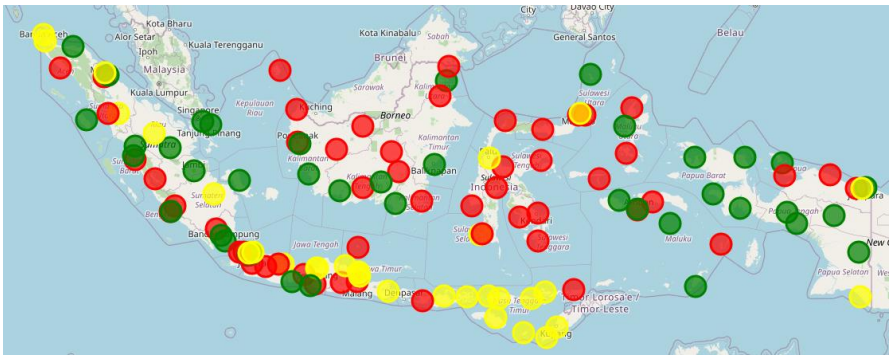

Gambar 3. Hasil dari (a) metode *elbow* dan (b) metode *sillhouette* untuk menemukan nilai K optimal

Dapat diketahui bahwa pada Gambar 4 entropi dengan nilai k sebesar 3 menghasilkan entropi yang tidak terlalu besar dan tidak terlalu kecil, sehingga tingkat ketidakteraturan datanya cenderung baik.

```
k = 2 entropy = 0.6811685453222742
k = 3 entropy = 1.076627811715253
k = 4 entropy = 1.2352701287991745
k = 5 entropy = 1.5507349837291144
k = 6 entropy = 1.676858284421512
k = 7 entropy = 1.797251639822941
k = 8 entropy = 1.9559190356238674
k = 9 entropy = 2.0217089369637016
```

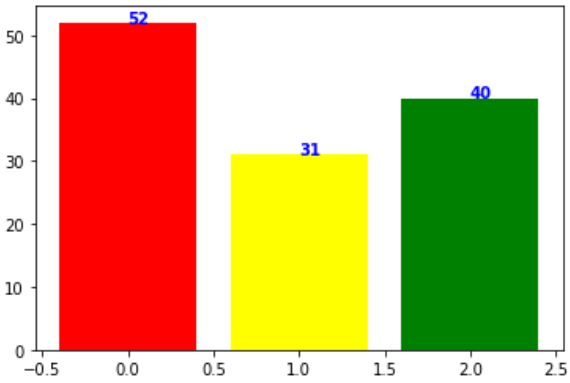
Gambar 4. Nilai entropi menguji nilai K

Untuk memastikan bahwa nilai K yang dihasilkan oleh kedua metode tersebut optimal. Pengujian K dengan uji entropi dilakukan. Hasil pemilihan K sangat disarankan untuk mengambil nilai entropi yang tidak mendekati nilai terendah maupun tinggi pada hasil ujinya, sehingga nilai k = 3 merupakan jumlah *cluster* yang tepat untuk menghindari bias pada presisi dan akurasi dengan nilai entropi yang optimal. Sebagaimana yang ditampilkan pada Gambar 4 mengenai visual peta sebaran stasiun dari seluruh daerah yang terbagi atas 3 (tiga) *cluster*, yaitu *cluster* 0 (merah), *cluster* 1 (kuning) dan cluster 2 (hijau).




Gambar 5. Hasil *cluster* cuaca ekstrim

Dan Gambar 5 mengenai jumlah data pada masing-masing *cluster* bahwa *cluster* 0 (merah) berjumlah 52 dengan potensi kerentanan tinggi yang dominan berada di Wilayah Indonesia Barat, cluster 1 (kuning) berjumlah 31 dengan potensi kerentanan sedang, dan cluster 2 (hijau) berjumlah 40 potensi kerentanan rendah dari seluruh data yang berjumlah 123, dapat dilihat bahwa cluster 0 berjumlah lebih banyak sedangkan cluster 1 paling sedikit.




Gambar 6. Jumlah data pada masing-masing *cluster*

Beberapa bagian pada studi yang dilakukan sejalan dengan Malino, et.al. [14], bahwa wilayah Sulawesi lebih cenderung berpotensi rentan terhadap cuaca ekstrim, yang mana faktor rentannya suatu wilayah terhadap perubahan iklim adalah kawasan perkotaan yaitu kawasan ini dihuni oleh sebagian besar penduduk. Begitu juga dengan wilayah Selatan Indonesia yaitu Jawa, Bali, Nusa Tenggara Timur dan Barat yang sering berafilisasi dengan eksistensi siklon (badai) tropis dan efek anomali iklim, baik El Nino atau La Nina sebagaimana yang tercantum pada jurnal Rahmawati, et.al. [15].

## V. Conclusion

Berdasarkan pembahasan pada bab-bab sebelumnya mengenai pengelompokan data iklim dari stasiun BMKG yang tersebar di seluruh Indonesia, *clustering* digunakan untuk mengelompokkan data cuaca berdasarkan kemiripan dan menggunakan algoritma *K-Means* untuk menentukan pola dari kelas terkait pada data cuaca ekstrim dari seluruh stasiun BMKG di Indonesia, sehingga dapat ditarik kesimpulan adalah data iklim dapat

dikelompokan menggunakan algoritma *K-Means* menghasilkan 3 kelompok yaitu cuaca ekstrim dengan potensi kerentanan tinggi, sedang, dan ringan. Pengujian dengan entropi diperlukan, karena data yang diperoleh mengandung faktor ketidakteraturan atau keacakan dan paling sering diterapkan pada seluruh penelitian pada subjek terkait. Menurut hasil visualisasi diatas pada data iklim tersebut, maka dapat diketahui bahwa beberapa daerah memiliki cenderung memiliki cuaca ekstrim tinggi terhadap perubahan iklim berdasarkan riwayat dan juga penelitian-penelitian yang sudah dilakukan.

## REFERENCES

[1] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Comput. Sci. Rev.*, vol. 40, pp. 1–23, May 2021, doi: 10.1016/j.cosrev.2021.100402.

[2] F. Hamami and A. I. Dahlan, "Klasifikasi Cuaca Provinsi DKI Jakarta menggunakan Algoritma Random Forest dengan Teknik Oversampling," *J. Teknoinfo*, vol. 16, no. 1, pp. 87–92, 2022.

[3] M. Mananohas, M. D. Bobanto, and F. Ferdy, "Hubungan Cuaca dan Tanaman Pangan Menggunakan Regresi Linear di Kota Tondano," *J. Mat. dan Apl.*, vol. 8, no. 2, pp. 169–175, 2019, [Online]. Available: https://ejournal.unsrat.ac.id/index.php/decartesian

[4] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar, "From clustering to clustering ensemble selection: A review," *Eng. Appl. Artif. Intell.*, vol. 104, pp. 1–21, Sep. 2021, doi: 10.1016/j.engappai.2021.104388.

[5] K. P. Sinaga, I. Hussain, and M. S. Yang, "Entropy K-Means Clustering with Feature Reduction under Unknown Number of Clusters," in *IEEE Access*, 2021, vol. 9, pp. 67736–67751. doi: 10.1109/ACCESS.2021.3077622.

[6] B. Poernomo and R. D. I. Sari, "Penerapan Data Mining untuk Prakiraan Cuaca di Kota Malang Menggunakan Algoritma Iterative Dichotomiser Tree (ID3)," *Jouticla*, vol. 3, no. 2, pp. 101–108, 2017.

[7] B. Purvis, Y. Mao, and D. Robinson, "Entropy and its Application to Urban Systems," *Entropy*, vol. 21, no. 1, pp. 1–15, Jan. 2019, doi: 10.3390/e21010056.

[8] S. Khairunnisa and M. I. Jambak, "Pengelompokan Cuaca Kota Palembang Menggunakan Algoritma K-Means Clustering Untuk Mengetahui Pola Karakteristik Cuaca," *J. Media Inform. Budidarma*, vol. 6, no. 4, pp. 2352–2360, Oct. 2022, doi: 10.30865/mib.v6i4.4810.

[9] H. L. Sari, "Fuzzy Clustering Dalam Pengclusteran Data Curah Hujan Kota Bengkulu Dengan Algoritma C-Means," *J. Ilm. Matrix*, vol. 16, no. 2, pp. 115–124, 2014.

[10] A. Chusyairi and P. Ramadar Noor Saputra, "Pengelompokan Data Puskesmas Banyuwangi Dalam Pemberian Imunisasi Menggunakan Metode K-Means Clustering," *Telematika*, vol. 12, no. 2, pp. 139–148, Aug. 2019, doi: 10.35671/telematika.v12i2.848.

[11] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, Jun. 2022, doi: 10.1007/s11135-021-01176-w.

[12] E. Umargono, J. E. Suseno, and V. Gunawan, "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula," in *International Seminar on Science and Technology (ISSTEC)*, 2020, pp. 121–129.

[13] M. Wei, T. W. S. Chow, and R. H. M. Chan, "Heterogeneous feature subset selection using mutual information-based feature transformation," *Neurocomputing*, vol. 168, pp. 706–718, Nov. 2015, doi: 10.1016/j.neucom.2015.05.053.

[14] C. R. Malino, M. Arsyad, and P. Palloan, "Analisis Parameter Curah Hujan dan Suhu Udara di Kota Makassar Terkait Fenomena Perubahan Iklim," *J. Sains dan Pendidik. Fis.*, vol. 17, no. 2, pp. 139–145, 2021.

[15] A. Rahmawati, B. Theo, T. Pamungkas, and D. Partini, "Pemetaan Tingkatan Cuaca Ekstrim Masing-Masing Kecamatan di Kota Kupang," *J. geoedusains*, vol. 2, no. 1, pp. 1–10, 2021.