KLASTERING KOPI ARABIKA MENGGUNAKAN ALGORITMA K-MEDOIDS

Sony Santana, Eddie Krishna Putra, Puspita Nurul Sabrina
Teknik Informatika, Universitas Jenderal Achmad Yani
Jalan Raya Karanglo km 2 Malang, Indonesia
Sonysann17@gmail.com

ABSTRAK

Industri kopi arabika merupakan salah satu usaha yang berkembang di indonesia. Usaha ini memiliki keuntungan yang menjanjikan seperti memberikan lapangan pekerjaan yang menguntungkan bagi masyarakat. Permasalahan yang muncul industri kopi yang semakin hari kreatif dan inovatif ini akan selalu mencari dan mengutamakan kualitas kopi yang menghasilkan produk yang mampu bersaing dengan lawan. Dataset yang diambil dalam penelitian ini bersumber dari Coffee Quality Institute dan dibagi ke dalam beberapa variabel yang dibutuhkan. Data tersebut akan melakukan pre-processing data seperti Cleaning Data, Transformasi Data, Normalisasi Data. Untuk mengatasi permasalahan tersebut dilakukan klasterisasi atau pengelompokan data dengan metode k-medoids. Output yang akan keluar akan menghasilkan nilai Cluster dari data kopi tersebut yang memiliki nilai kemurnian Cluster yang tinggi. Dari hasil penenelitian ini menunjukan klaster 1 menonjol dengan nilai terbaik untuk aroma, Flavor, acidity, dan Uniformity, meliputi negara-negara tertentu seperti Guatemala, Honduras, Ethiopia, dan lainnya. Klaster 2, diwakili oleh negara-negara seperti Colombia, Brazil, dan Costa Rica, menunjukkan nilai terbaik pada atribut seperti aftertaste, body, Balance, clean cup, Sweetness, cupper point, dan total cup point. Klaster 3, yang meliputi sejumlah negara termasuk Mexico, Taiwan, dan United States (Hawaii), menunjukkan nilai terendah pada semua atribut. Jadi hasil akhir dari Cluster 2 memiliki nilai atribut rata-rata yang tinggi, sementara Cluster 1 memiliki nilai rata-rata yang sedang, dan Cluster 3 menampilkan nilai rata-rata yang rendah.

Kata Kunci : *Kopi Arabika, Clustering, K-Medoids, Silhouette Index, Pre-processing*

1. PENDAHULUAN

Kopi arabika merupakan kopi yang berasal dari negara Afrika yang terletak di kawasan Etiopia. Tanaman kopi ini dapat ditemukan pada wilayah dataran Yaman bagian selatan jazizar Arab [1]. Pasar internasional dalam industri kopi diakui sebagai produk yang paling banyak dikonsumsi di dunia dan sudah menjadi gaya hidup. Industri bisnis dibidang produksi kopi akan selalu mengutamakan kualitas kopi untuk menciptakan produk yang mampu bersaing. Dalam menentukan kualitas kopi, bisnis usaha kopi masih menggunakan cara pengenalan analisis pribadi dan manual. Hal ini membuat proses pemilihan kopi yang berkualitas memakan banyak waktu dan tenaga. Sementara itu, diperlukan proses yang lebih cepat dan akurat untuk meningkatkan produksi. Bagaimana kemampuan dari metode yang dipakai *clustering* ini berpengaruh baik atau tidak dalam penelitian kopi. Peran teknologi disini dapat membantu industri bisnis kopi yang lebih efisien dan meningkatkan kualitas produksi. Penentuan kualitas kopi membutuhkan suatu sistem yang tepat untuk menganalisa masalah ini [2].

Sebuah sistem yang dapat dibangun untuk mengatasi masalah tersebut dengan *clustering* dengan metode K-Medoids [3].

Pemilihan topik dalam penelitian ini dilandasi oleh penelitian sebelumnya dengan mengidentifikasi metode dan cara implementasi yang berbeda, serta bisa menyempurnakan metode yang belum dicoba sebelumnya. Pembahasan yang akan muncul dalam penelitian ini yaitu mengetahui pembentukan *cluster* yang paling optimal dalam klasterisasi data kopi arabika dari yang dihasilkan menggunakan metode yang dipakai serta mencari kualitas kopi terbaik yang bisa dijadikan referensi menghasilkan produk yang dapat bersaing dengan pasar.

Berbeda dengan Hasil penelitian ini bertujuan untuk membantu pengelompokan data kopi arabika dengan metode yang dipakai yaitu *clustering*, serta menguji nilai *cluster* dari K-Medoids dengan menguji kemurnian *cluster* dari metode, dan diakhir bisa berkontribusi terhadap ilmu pengetahuan khususnya dibidang IT pada data mining. Algoritma K-Medoids merupakan salah satu algoritma *Clustering* yang digunakan untuk menyelesaikan masalah *clustering* dengan menggunakan representasi pusat dari setiap *cluster*. Penggunaan algoritma pengelompokan kopi arabika K-Medoids bertujuan untuk memudahkan pemilihan kualitas kopi arabika sesuai dengan preferensi konsumen dan menjamin kualitas produk kopi arabika yang konsisten. Keunikan dalam penelitian ini yaitu kasus kopi Arabika yang diambil masih sedikit. Penelitian ini berfokus pada *pre processing* data dari dataset numerik. Dengan menguji kemurnian *cluster* yang dipakai seperti K-Medoids dengan dengan *silhouette score.*

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Beberapa penelitian sebelumnya telah melakukan pengujian dengan metode algoritma SGD, *random forest* dan *naive bayes*. Dalam penelitian ini membandingkan ketiga metode dengan *dataset* yang dipakai terdiri dari beberapa atribut. Penelitian ini

dapat menjadi celah untuk dikembangkan kembali dengan menguji metode yang belum dipakai dan dengan menambahkan fitur seleksi ke dalam penelitian selanjutnya [1].

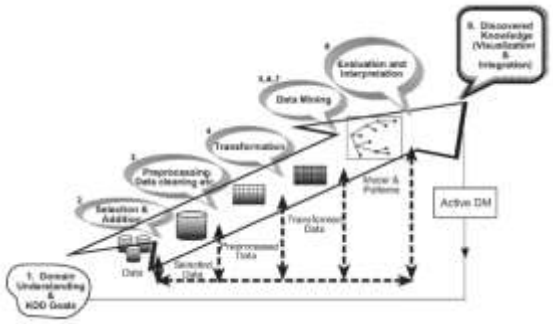
Penelitian berikutnya menguji dengan metode random forest dengan pendekatan data science dengan hasil akhir akurasi random forest dalam memprediksi kualitas kopi pada penelitian ini sebesar 79% [2]. Kemudian, pada penelitian sebelumya berkaitan dengan KNN dengan *case* studi yang diambil harga komiditi kopi arabika. Dan untuk pengujian disini akan dilakukan peramalan harga komoditi kopi arabika. Hasil akhir dalam penelitian ini metode yang dipakai bisa memperkecil kesalahan nilai dengan fitur optimasi dan tanpa seleksi fitur [3].

Pada penelitian terdahulu sudah menerapkan beberapa metode algoritma *data mining* seperti dalam pengelompokan bibit tanaman kopi arabika yang dilakukan dengan metode *K-Means clustering* [4]. Kemudian terdapat optimalisasi penentuan sumber pasokan kopi arabika gayo melalui pendekatan *hirarchical clusteting* [5]. Ada juga prediksi tingkat produksi kopi menggunakan regresi linear [6] menggunakan algoritma C 4.5 [7] menggunakan metode *backpropagation* dan algoritma *support vector machine* [8] Metode Asosiasi *data mining* menggunakan algoritma *apriori* [9] memakai algoritma *random forest* [2].

2.2. Knowledge Discovery In Databases (KDD)

Knowledge discovery in databases (KDD) adalah proses mengekstraksi model dan pola dari database besar. Abstraksi model dari database besar dan mencari pola dan gejala yang valid, baru, dan nontrivial dalam model yang diabstraksi. Istilah data mining (DM) sering digunakan sebagai sinonim untuk proses KDD meskipun sebenarnya ini hanyalah sebuah langkah dalam proses KDD. Data mining mengacu pada proses penerapan algoritma penemuan pada data. Abstraksi model dari database besar dan mencari pola dan gejala yang valid, baru, dan nontrivial dalam model yang diabstraksi [10].

Dalam analisis eksplorasi dan pemodelan otomatis dari repositori data besar dan proses terorganisir untuk mengidentifikasi pola yang valid, baru, berguna, dan dapat dipahami dari kumpulan data yang besar dan kompleks.



Gambar 1. Alur Knowledge Discovery in Database

Knowledge discovery in databases (Gambar 2.1) bersifat berulang dan interaktif, terdiri dari sembilan langkah. Perhatikan bahwa prosesnya berulang pada setiap langkah, artinya mungkin diperlukan langkah mundur untuk menyesuaikan langkah sebelumnya. Prosesnya memiliki banyak aspek "artistik" dalam artian seseorang tidak dapat menyajikan satu formula atau membuat taksonomi lengkap untuk pilihan yang tepat untuk setiap langkah dan jenis aplikasi. Oleh karena itu diperlukan pemahaman mendalam tentang proses dan berbagai kebutuhan serta kemungkinan dalam setiap langkah. Taksonomi untuk metode Data Mining membantu dalam proses ini [11].

2.3. Clustering

Clustering merupakan pengelompokan data, pengamatan atau kasus ke dalam kelas objek yang serupa. Cluster merupakan kumpulan record data yang memiliki kemiripan satu sama lain dan berbeda dari record data di cluster lain. Clustering memiliki perbedaan dengan klasifikasi karena tidak memiliki variabel target untuk melakukan clustering. Karena tugas cluster bukan hanya mengklasifikasikan atau memprediksi suatu variabel target. Algoritme pengelompokan mencoba mengelompokkan seluruh kumpulan data ke dalam subkelompok atau cluster yang relatif homogen, dengan memaksimalkan kesamaan catatan di dalam cluster dan meminimalkan kesamaan dengan catatan di luar cluster tersebut [12].

2.4. K-Medoids

K-Medoids adalah sebuah metode algortima yang berguna untuk menemukan *Medoids* dalam sebuah kelompok *cluster* yang menjadi inti dari pengelompokan *Cluster*. Algoritma ini lebih baik dari pada *K-Means*. Hal ini karena *K-Medoids* mencari k sebagai objek dari representatif yang berfungsi meminimalkan jumlah tidak sama pada objek data, *K-Medoids* memakai penjumlahan jarak *euclidean* dari objek data [13]. Algoritma *K-Medoids* membentuk *cluster* dengan perhitungan kesamaan jarak antara objek *Medoid* dan *non-Medoid*. Analisis ini meminimalkan ketidakmiripan setiap objek dalam *cluster* dengan menggunakan nilai *absolute error* (E) [14].

Langkah-langkah algoritma *K-Medoids* sebagai berikut:

a. Inisialisasi k pusat *cluster* (jumlah kluster).
b. Gunakan pengukuran jarak *euclidean* pada persamaan berikut untuk menetapkan semua data (objek) ke *cluster* terdekat : x

$$E = \sum_{c=1}^{k} \sum_{i=1}^{n_c} |p_{ic} - 0_c| \qquad (1)$$

Penjelasan:
nc = jumlah objek yang terdapat pada *cluster* ke-c.
pic = objek-objek *non-Medoids* pada *cluster* ke-c.
Oc = Untuk nilai *Medoids* dari *cluster* ke-c.

Algoritma *K-Medoids* adalah sebagai berikut:

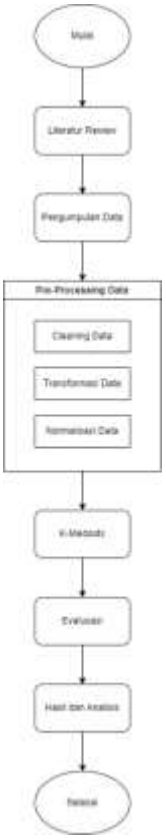
a. Proses pemilihan k objek menjadi O_c dilakukan dengan memilih k objek sebagai *medoid* dari masing-masing *cluster* ke-c, dengan c = 1, 2, 3, . . ., k.
b. Dilakukan perhitungan kemiripan antara objek *Medoid* dengan objek *non-medoid* dengan memakai jarak *euclidean* menggunakan persamaan yang ditentukan.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2} \quad (2)$$

Penjelasan:
d = jarak antara objek dan *centroid.*
p = data atau objek yang akan dihitung jaraknya.
q = titik acuan yang digunakan untuk menghitung jarak dengan objek p.

c. Tetapkan objek *non-medoid* ke grup yang paling dekat dengan *medoid* pilih O_{random}, dengan O_{random}, adalah sebuah objek *non-medoids* untuk O_c awal.
d. Hitung kesamaan antara objek non-O_{random} dengan objek O_{random} menggunakan jarak *euclidean.*
e. Tempatkan objek non-O_{random} ke dalam grup yang paling mirip dengan O_{random}.
f. Hitung nilai kesalahan mutlak sebelum dan sesudah mengganti O_c dengan O_{random} Jika $E_{random} < E_c$ maka tukar O_j dengan O_{random} tetapi jika $E_{random} > E_c$ maka O_c tetap ada.

3. METODE PENELITIAN



Gambar 2. Metode Penelitian

Metode yang dipakai pada penelitian ini adalah langkah-langkah dalam mengklasifikasikan data kopi arabika menggunakan algoritma k-medoids dan output yang akan disimpulkan bahwa terdapat 3 cluster yang memiliki nilai kualitas tinggi, sedang, dan rendah. Pada tahapan ini dimulai dengan literatur review, pengumpulan data, pre-processing yang terdiri dari tiga tahapan yaitu cleaning, transformasi data, normalisasi data yang akan dilanjutkan dengan implementasi algoritma K-Medoids serta pengukuran evaluasi klaster.

3.1. Pengumpulan Data

Dalam proses pencarian dataset dilakukan dengan melakukan studi literatur pada jurnal dan website open source pada internet. Dataset kopi arabika yang digunakan dalam penelitian ini adalah data yang dapat diakses secara publik dan diambil dari situs kaggle. Pemilihan kopi arabika sebagai dataset dipertimbangkan dengan penelinitan yang ada sebelumnya dari dataset yang sama. Terakhir kali data ini diperbarui pada tahun 2020. Tujuan dibuatnya data ini untuk kopi arabika dilakukan clustering menggunakan K-Medoids sehingga menghasilkan cluster yang memiliki kemurnian yang tinggi, mengoptimalkan kualitas kopi arabika menggunakan analisis data mining untuk meningkatkan nilai bisnis dan pengalaman pelanggan, mengekstraksi informasi dan menemukan pola yang bermanfaat dari dataset kopi arabika menggunakan metode K-Medoids. Berikut dataset kopi arabika yang terdiri atas 43 atribut termasuk kelas data dan memiliki 1312 baris data.

3.2. Pre-processing Data

Gambar 3. Alur Pre-Processing

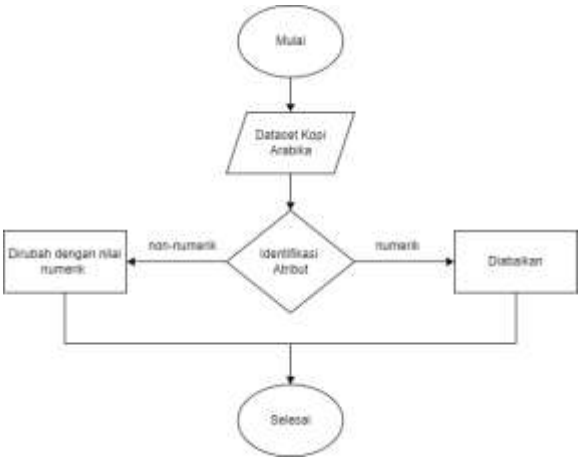
Setelah tahap pengumpulan dataset dan penjelasan tentang definisi atribut dan identifikasi atribut, langkah berikutnya adalah melibatkan proses pre-processing data yang terdiri dari tiga langkah, yakni Data Cleaning, Data Transformation, dan Normalization, yang dilakukan secara berurutan. Pada Gambar 3.2 menjelaskan tahapan yang akan dilalui dalam proses pre-processing Data tersebut. Tahap awal dari proses pre-processing data adalah data cleaning yang melibatkan seleksi atribut, penghapusan nilai null, dan penghapusan nilai duplikat. Setelah itu, dilanjutkan dengan proses Transformasi data yang mengubah nilai atribut dari Country.of.Origin yang sebelumnya merupakan string menjadi data numerik.

3.3. Data Cleaning

Dalam proses pembersihan data memiliki beberapa proses didalam nya seperti *feature selection*, *duplicat value*, dan pengecekan *missing value* dan data kosong.

3.4. Transformasi Data

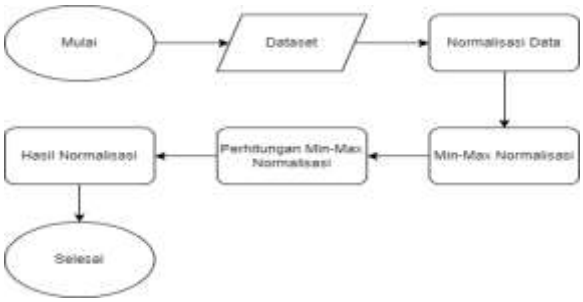
Pada tranformasi data langkah pertama yang akan dilakukan masuk ke dalam fitur encoding data, didalam encoding data ini dataset yang sudah di proses pada fitur cleaning Data akan dilanjutnya dengan proses transformasi data dan akan berlanjut kedalam proses normalisasi data. Pada tahap ini dilakukan transformasi data untuk mengubah format data agar dapat digunakan pada proses data mining, pada dataset yang dipakai sudah kategori numerik jadi untuk transformasi data hanya mentrasformasikan data yang masih nominal. Tujuan dari transformasi data pada clustering K-Medoids bertujuan untuk mempersiapkan dataset agar sesuai dengan asumsi dan persyaratan algoritma k-medoids. Dengan melakukan transformasi data yang tepat, dapat meningkatkan performa dan kualitas klustering yang dihasilkan.



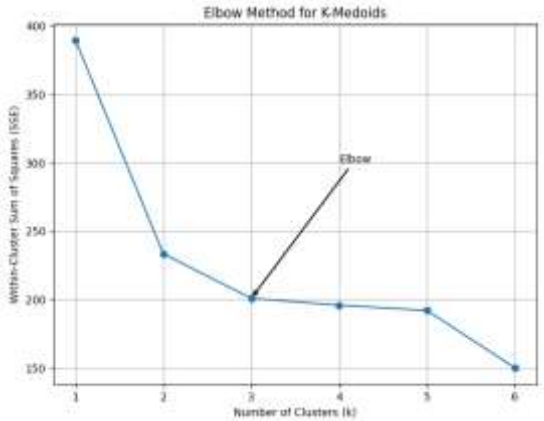
Gambar 4. Alur Transformasi Data

Langkah pertama dalam proses pengkodean label adalah mengidentifikasi atribut dalam kumpulan data yang memiliki nilai kategorikal atau non-numerik, seperti “Negara Asal” atau atribut lain yang tidak dapat langsung diartikan sebagai angka. Setelah mengidentifikasi atribut kategorikal, langkah selanjutnya mengimplementasikan pengkodean label menggunakan library *scikit-learn* Python. Dalam hal ini, gunakan modul LabelEncoder *scikit-learn* untuk mengonversi nilai kategorikal menjadi angka. Terapkan pengkodean label ke atribut yang diidentifikasi sebagai kategorikal. Fungsi fit_transform() LabelEncoder digunakan untuk mengubah nilai kategorikal menjadi angka baru. Hasil pengkodean label biasanya disimpan di kolom baru DataFrame. Setelah proses pengkodean label selesai, menyimpan hasil konversinya pada bingkai data yang dimodifikasi dalam format file yang sesuai seperti CSV dan menggunakannya dalam proses analisis atau pengelompokan selanjutnya menggunakan algoritma K-Medoids. Proses dan alur *encoding* data dapat dilihat pada Gambar 3.3

3.5. Normalisasi Data



Gambar 5. Alur Normalisasi Data



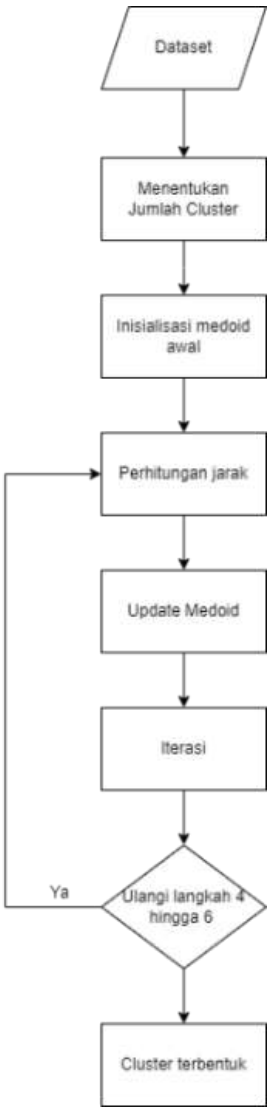
Gambar 6. Nilai Metode Elbow

Pada tahap ini data sudah siap untuk digunakan dalam proses data mining, data sudah melalui proses pembersihan dan transformasi sehingga akan langsung bisa digunakan. Normalisasi data dapat dilakukan ketika terdapat perbedaan skala atau rentang nilai antara atribut-atribut dalam dataset. Jika terdapat setidaknya satu atribut dengan rentang nilai yang jauh lebih besar daripada atribut lainnya, normalisasi data digunakan untuk menyamakan skala atribut-atribut tersebut. Tujuan normalisasi adalah agar semua atribut memiliki kontribusi yang seimbang dalam perhitungan jarak atau kesamaan antar data. Selain itu, normalisasi

data juga berguna jika terdapat asumsi bahwa atribut-atribut dalam dataset memiliki distribusi yang mirip atau mendekati distribusi normal. Dengan melakukan normalisasi, data akan lebih sesuai dengan asumsi tersebut, sehingga analisis statistik atau metode clustering yang bergantung pada asumsi tersebut dapat memberikan hasil yang lebih akurat dan bermakna.

3.6. K-Medoids

Pada tahap clustering menggunakan metode K-Medoids, clustering K-Medoids adalah sebuah metode dalam analisis data yang digunakan untuk mengelompokkan objek-objek ke dalam beberapa kelompok berdasarkan kemiripan atribut atau karakteristik yang dimiliki. Dalam penelitian ini kita akan menggunakan metode K-Medoids untuk melakukan clustering pada dataset kopi arabika. Tahapan dari proses k-medoids dapat dilihat pada Gambar 3.5



Gambar 7. Alur K-Medoids

3.6. Evaluasi

Saat mengevaluasi pengelompokan dalam konteks K-Medoids pada data kopi Arabika, proses utamanya adalah memahami seberapa baik kluster yang dihasilkan mewakili pola dan struktur yang ada dalam data. Metode evaluasi ini melibatkan beberapa langkah, antara lain penggunaan indikator internal dan eksternal. Metrik internal seperti WCSS (containment sum of squares) dan skor siluet digunakan untuk mengukur seberapa kompak cluster internal dan seberapa jauh jarak cluster. Peringkat yang baik menunjukkan bahwa klaster-klaster tersebut homogen dan terpisah secara jelas satu sama lain, sehingga memungkinkan pemahaman yang lebih baik mengenai karakteristik kopi Arabika yang diwakili oleh masing-masing klaster.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Perhitungan Nilai Metode Elbow

Berdasarkan hasil perhitungan dibawah elbow method optimal untuk menentukan titik cluster yaitu beradapa pada 3 cluster.

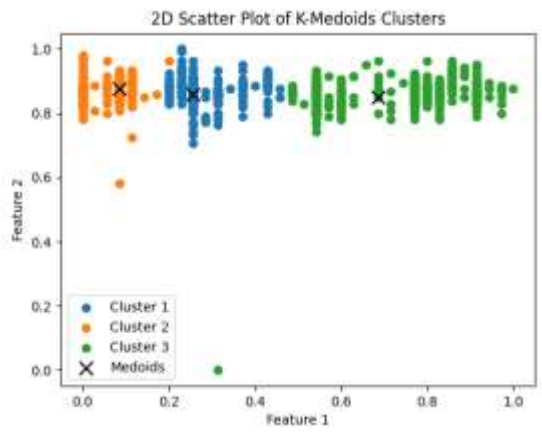
Dari hasil Metode Elbow, diperoleh nilai WCSS (Within-Cluster Sum of Squares) untuk beberapa jumlah kluster yang telah ditentukan. Berikut adalah nilai-nilai WCSS:

- K1 = 389.5473995864868
- K2 = 233.36886577194753
- K3 = 200.88852579860497
- K4 = 195.88348567392427
- K5 = 192.05682027552004
- K6 = 150.08912019411298

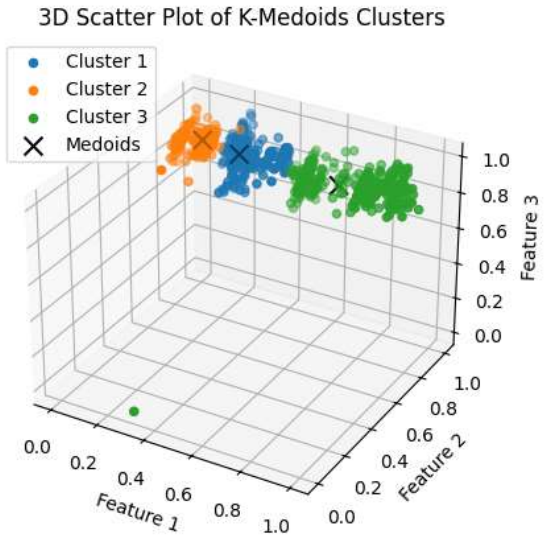
WCSS mencerminkan seberapa dekat titik data dalam satu kluster dengan pusat klusternya. Semakin kecil nilai WCSS, semakin homogen dan padat klusternya. Metode Elbow bertujuan menemukan titik di mana penurunan nilai WCSS tidak lagi signifikan seiring dengan penambahan jumlah kluster. Pada kasus ini, perhatikan penurunan yang besar dari k=1 ke k=2, kemudian penurunan yang lebih lambat dari k=2 hingga k=6. Titik di mana penurunan ini melambat secara signifikan pada grafik disebut "elbow," dan itulah jumlah kluster yang optimal untuk data tersebut. Dalam hal ini, k=2 atau k=3 bisa menjadi pilihan yang tepat karena setelah titik tersebut, penurunan WCSS tidak lagi signifikan dengan penambahan jumlah kluster.

4.2. Hasil Pengujian Algoritma K-Medoids

Pada pengujian algoritma k-meoids ini menunjukkan tiga klaster yang terbentuk dari 12 atribut yang diamati. Setiap klaster ditampilkan dengan warna berbeda: Klaster 1 sebagai biru, Klaster 2 sebagai orange, dan Klaster 3 sebagai hijau. Plot ini berguna untuk memvisualisasikan pola dan distribusi data secara tiga dimensi serta membedakan klaster berdasarkan atribut yang digunakan.



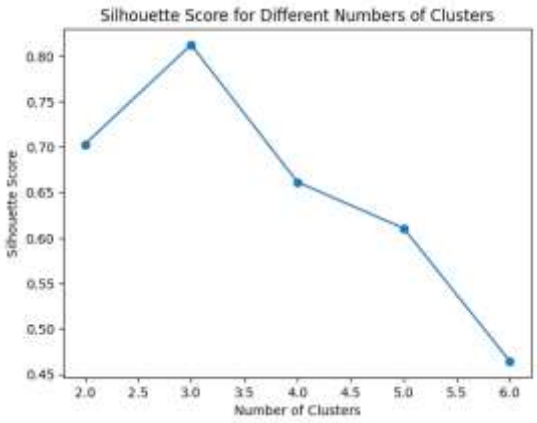
Gambar 8. Hasil Visualisasi 2D Scatter Plot K-Medoids Cluster



Gambar 9. Hasil Visualisasi 3D Scatter Plot K-Medoids Cluster

4.3. Hasil Evaluasi

Evaluasi menggunakan silhouette coefficient dalam clustering K-Medoids kopi Arabika membantu menilai seberapa baik data berada dalam klusternya dibandingkan dengan yang lain. Rentang nilai koefisien silhouette dari -1 hingga 1; semakin tinggi nilainya, semakin baik kesesuaian data dengan kluster tempatnya. Ketika nilai mendekati 1, itu menandakan bahwa kluster kelompok data secara jelas terpisah. Evaluasi ini mempertimbangkan jarak antara data dalam kluster dan seberapa baik data tersebut sesuai dengan kluster lainnya. Dalam konteks K-Medoids untuk kopi Arabika, nilai yang positif menunjukkan bahwa data cenderung homogen di dalam kluster, menegaskan pembentukan kluster yang baik untuk sampel data kopi Arabika.

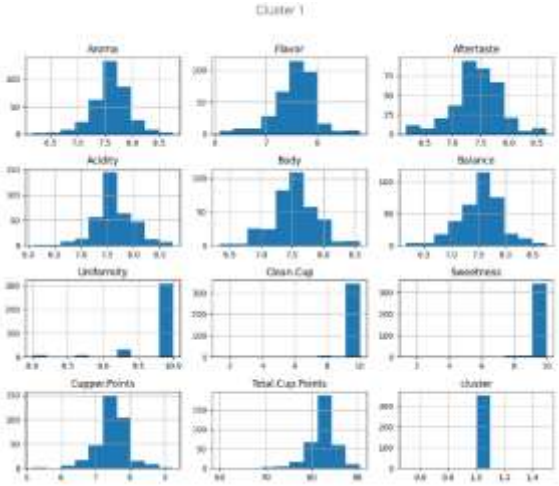


Gambar 10. Hasil Silhouette Score

Dari hasil *silhouette score* diatas ukuran seberapa baik setiap data cocok dengan klaster yang telah dibentuk. Rentang nilainya dari -1 hingga 1, dimana skor lebih tinggi menunjukkan klaster yang lebih terpisah dan padat. Dalam penelitian ini saat jumlah klaster adalah 3, nilai silhouette_score tertinggi sekitar 0.8123, menandakan bahwa pembagian menjadi 3 klaster memberikan hasil yang lebih baik dibandingkan dengan jumlah klaster lainnya. Nilai tersebut menunjukkan bahwa klaster tersebut memiliki pemisahan yang baik dan titik-titik dalam klaster yang sama lebih dekat satu sama lain dibandingkan dengan klaster lain. Jumlah klaster lain seperti 2, 4, 5, dan 6, memiliki nilai silhouette_score yang lebih rendah, menunjukkan adanya tumpang tindih atau pemisahan yang kurang jelas antar klaster atau klaster yang kurang padat. Dengan demikian, berdasarkan evaluasi *silhouette coefficient*, jumlah klaster yang paling optimal adalah 3.

4.4. Hasil Interpretasi

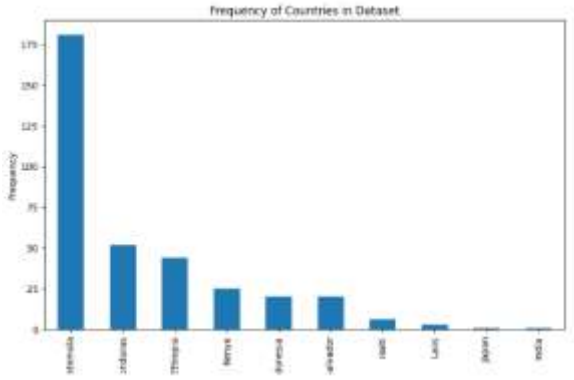
Interpretasi hasil dari clustering melibatkan menganalisis pola-pola dan ciri-ciri yang muncul dari kelompok-kelompok yang dibentuk oleh proses pengelompokan data. Ini mencakup pemahaman terhadap kesamaan atau perbedaan di antara kelompok, serta signifikansinya dalam konteks yang relevan. Dalam kopi Arabika, interpretasi hasil klastering dapat mengidentifikasi kelompok-kelompok kopi berdasarkan beragam karakteristik seperti aroma, Flavor atau acidity. Hal ini membantu dalam memahami perbedaan kualitas atau karakteristik antar-kelompok dan dapat digunakan sebagai dasar untuk pengambilan keputusan, seperti dalam pemasaran atau evaluasi kualitas.



Gambar 8. Hasil Interpretasi Cluster 1

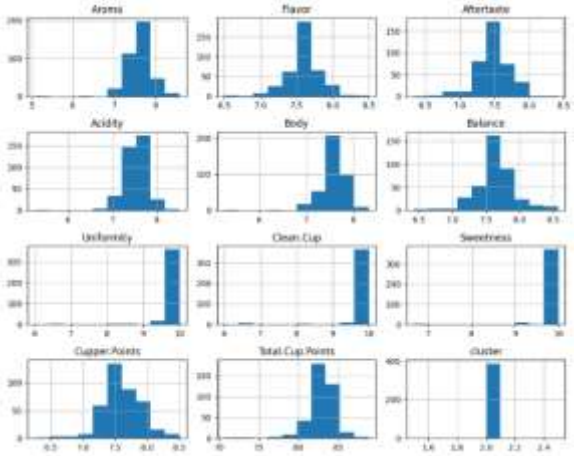
Dalam Nilai-nilai dari atribut dalam *cluster* 1 menunjukkan variasi frekuensi yang cukup besar. Pada, atribut Aroma dengan nilai 7.50 memiliki frekuensi terbanyak sebanyak 55 kali, sementara nilai-nilai seperti 6.17, 6.50, 6.42, 7.81, 8.33, dan 8.75 hanya muncul dalam frekuensi yang sangat rendah. Demikian pula, untuk atribut *Flavor*, nilai 7.50 juga memiliki frekuensi tertinggi sebanyak 46 kali, sedangkan beberapa nilai lainnya seperti 6.08, 6.58, 6.75, 6.92, 7.88, dan 8.83 hanya muncul dalam frekuensi yang terbilang minim. Hal serupa juga terlihat pada atribut *Aftertaste* , di mana nilai 7.33 mendominasi dengan frekuensi tertinggi sebanyak 43 kali, sementara nilai-nilai seperti 6.42, 6.58, 7.56, 8.17, 8.58, dan 8.67 muncul dalam frekuensi yang lebih rendah. Pada atribut Acidity pada *Cluster* 1 memiliki nilai terbanyak pada angka 7.67 dengan frekuensi 42, sementara nilai-nilai seperti 6.08, 6.50, 8.58, dan 8.75 memiliki frekuensi paling rendah. Pada atribut Body, nilai 7.50 menjadi yang paling dominan dengan frekuensi 44, sedangkan nilai-nilai seperti 6.33, 6.50, 7.63, dan 8.58 memiliki frekuensi terendah.

Pada atribut *Balance*, nilai 7.50 memiliki frekuensi tertinggi sebanyak 48 kali, sementara nilai 6.33, 6.58, 8.25, 8.58, dan 8.75 memiliki frekuensi yang paling minim. Atribut *Uniformity* dan *Sweetness* menunjukkan nilai tertinggi pada angka 10.00 dengan frekuensi 308 dan 313, namun beberapa nilai seperti 8.00, 8.67, 1.33, 6.00, dan 6.67 memiliki frekuensi paling rendah. Terakhir, pada atribut *Cupper Points* dan *Total Cup Points*, nilai-nilai tertinggi pada angka 7.50 dan 83.17 masing-masing dengan frekuensi 43 dan 14, sementara nilai-nilai seperti 5.17, 5.42, 6.00, 6.58, 8.83, 8.75, 9.00, 9.25, 59.83, 81.08, 81.75, 82.08, dan 89.92 memiliki frekuensi yang paling minim. Total data dalam *cluster* 1 adalah sebanyak 353, dengan variasi yang signifikan pada setiap atributnya. Guatemala menjadi negara dengan jumlah data terbanyak di *cluster* 1, sementara India memiliki jumlah data terendah.

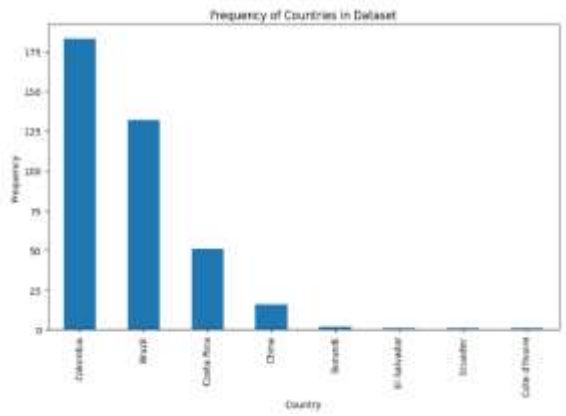


Gambar 9. Hasil Interpretasi Country 1

Sama seperti atribut dalam cluster 1 sebelumnya dalam cluster 2 ini terdapat bebrapa atribut dalam Nilai-nilai dari atribut dalam cluster 2 menunjukkan variasi frekuensi yang cukup besar dimulai dari yang paling tinggi sampai paling rendah. Atribut aroma dari cluster 2 memiliki nilai puncak pada 7.67, muncul sebanyak 65 kali, sementara nilai-nilai lain seperti 5.08, 6.33, 6.83, 6.92, 7.81, 8.25, 8.33, dan 8.58 muncul jarang. Dalam hal Flavor, nilai tertingginya adalah 7.67 dengan frekuensi 67 kali, namun beberapa nilai seperti 6.50, 6.75, 6.58, 6.92, 7.81, 8.17, 8.42, dan 8.50 muncul dengan frekuensi yang paling minim. Seterusnya, atribut aftertaste menunjukkan nilai tertinggi 7.50 dengan frekuensi 64, sementara beberapa nilai seperti 6.33, 6.58, 6.67, 6.75, 7.88, 8.08, dan 8.42 muncul dengan frekuensi yang sangat sedikit. Demikian pula, atribut acidity dan body memiliki nilai tertinggi yang sama dengan frekuensi 64 dan 65, namun muncul nilai-nilai yang jarang seperti 5.25, 6.75, 6.92, 8.33, 8.50, 7.38, dan 8.25.

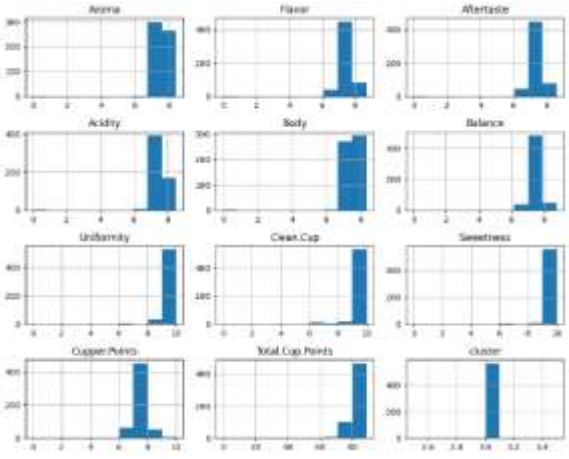


Gambar 10. Hasil Interpretasi Cluster 2

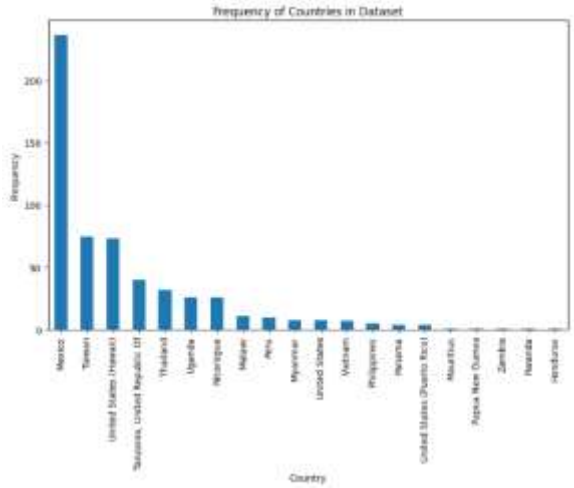


Gambar 11. Hasil Interpretasi Country 2

Dalam atribut *Uniformity* menampilkan nilai tertinggi sebesar 10.00, dengan frekuensi terbanyak mencapai 360 kali, sementara nilai 6.00 dan 9.50 muncul paling jarang. Hal serupa terjadi pada atribut *Clean.Cup* yang menampilkan nilai puncak 10.00 dengan frekuensi 366, namun nilai 6.00 dan 8.67 muncul sangat jarang. Selain itu, atribut *Sweetness* juga memiliki nilai tertinggi 10.00 dengan frekuensi tertinggi sebesar 372, sedangkan nilai 8.67 muncul paling jarang. Pada atribut *Cupper.point*, nilai tertinggi terdapat pada 7.67 dengan frekuensi 53, dan nilai 6.25, 6.92, dan 8.13 muncul dengan frekuensi yang minim. Terakhir, atribut *Total.Cup.Points* menunjukkan nilai tertinggi 83.00 dengan frekuensi 18, sementara beberapa nilai seperti 70.67, 86.33, 86.43, 86.92, dan 87.17 muncul sangat jarang dalam *Cluster* ini. Keseluruhan atribut dalam *Cluster* 2 memiliki total 387 data, dengan jumlah tertinggi dari negara Colombia dan jumlah terendah dari negara Cote d'Ivoire.



Gambar 12. Hasil Interpretasi Cluster 3



Gambar 13. Hasil Interpretasi Country 3

Yang terakhir pada cluster 3 atribut aroma menunjukkan nilai tertinggi sebesar 7.50 dengan frekuensi terbanyak sebesar 66, sementara nilai 0.00, 6.50, dan 6.67 memiliki frekuensi paling rendah. Atribut Flavor dari Cluster ini menampilkan nilai 7.58 dengan frekuensi tertinggi 71, namun nilai 0.00, 6.33, 6.42, 6.50, 7.88, 8.08, dan 8.67 muncul dengan frekuensi paling minim. Kemudian, atribut aftertaste menunjukkan nilai 7.33 dengan frekuensi tertinggi 67, sementara nilai 0.00, 6.17, 6.25, 6.33, 6.50, 7.38, 8.25, dan 8.58 memiliki frekuensi paling rendah. Selanjutnya, atribut acidity menampilkan nilai 7.58 dengan frekuensi tertinggi 68, namun nilai 0.00, 6.25, 6.67, dan 7.63 muncul dengan frekuensi yang sangat minim. Atribut body memiliki nilai tertinggi 7.50 dengan frekuensi 93, dan nilai 0.00, 6.42, 6.83, dan 8.33 muncul dengan frekuensi yang sangat jarang. Sementara pada atribut Balance, nilai tertinggi adalah 7.50 dengan frekuensi 78, dan nilai 0.00, 6.08, 6.50, 6.67, dan 8.75 muncul dengan frekuensi paling rendah. Dan untuk atribut Uniformity menampilkan nilai tertinggi 10.00 dengan frekuensi 459, sedangkan nilai 0.00 dan 9.00 memiliki frekuensi paling minim dalam cluster ini.

Untuk atribut Clean.Cup, Sweetness, Cupper.point, dan Total.Cup.Points pada Cluster 3 menampilkan nilai dan frekuensi masing-masing atribut. Clean.Cup menunjukkan nilai terbesar 10.00 dengan frekuensi tertinggi 502, sementara nilai 0.00, 2.67, 5.33, 7.33 memiliki frekuensi paling minim. Sweetness memiliki nilai tertinggi 10.00 dengan frekuensi 532, sedangkan nilai 0.00, 6.67, 6.00 memiliki frekuensi terendah dalam cluster ini. Cupper.point menunjukkan nilai tertinggi 7.50 dengan frekuensi 57, sementara nilai 0.00, 5.25, 6.17, 6.50 memiliki frekuensi terendah dalam cluster ini. Terakhir, Total.Cup.Points menampilkan nilai tertinggi 83.00 dengan frekuensi 14, dan nilai 0.00, 77.33, 77.50, 77.58, 77.67 memiliki frekuensi yang paling rendah. Secara keseluruhan, Cluster 3 memiliki 570 data dengan Mexico memiliki jumlah data terbanyak, sementara Mauritius, Papua New Guinea,

Zambia, Rwanda, dan Honduras memiliki jumlah data yang paling minim dalam cluster ini.

5. KESIMPULAN DAN SARAN

Dalam penelitian klastering kopi Arabika menggunakan metode k-medoids, analisis terhadap dua belas atribut mengungkapkan tiga klaster yang memiliki perbedaan yang signifikan dalam karakteristik kopi. Klaster pertama, terdiri dari negara-negara seperti Guatemala, Honduras, Ethiopia, Kenya, Indonesia, El Salvador, Haiti, Laos, Japan dan India menunjukkan kualitas unggul pada atribut aroma, Flavor, acidity, dan Uniformity. Klaster kedua, yang mewakili Colombia, Brazil, Costa Rica, China, Burundi, El Savador, Ecuador dan Cote d'lvoire menunjukkan nilai tertinggi pada sejumlah atribut seperti aftertaste, body, Balance, clean cup, Sweetness, cupper point, dan total cup point. Sebaliknya, Klaster ketiga yang melibatkan negara-negara seperti Mexico, Taiwan, United States (Hawaii), United Republic Of Tanzania, Thailand, Uganda, Nicaragua, Malawi, Peru, Myanmar, United States, Vietnam, Philippines, Panama, United States (Puerto Rico), Mauritius, Papua New Guinea, Zambia, Rwanda, dan Honduras menampilkan nilai terendah pada semua atribut yang diamati. Dari hasil tersebut, dapat disimpulkan bahwa laster 2 menonjol dengan nilai atribut rata-rata yang tinggi, klaster 1 memiliki nilai rata-rata menengah, sementara klaster 3 menunjukkan nilai rata-rata yang rendah dalam ciri khas kopi Arabika yang diamati. Untuk langkah selanjutnya dalam studi klastering kopi arabika, disarankan untuk memilih atribut alternatif yang mempengaruhi kualitas kopi, seperti nutrisi tanah atau kondisi mikroklimat. Bisa dengan pembandingan metode klastering lain seperti mengadakan perbandingan antara K-Medoids dengan metode clustering lain seperti Fuzzy C-Means atau Density-Based clustering untuk mengevaluasi keunggulan relatif dari masing-masing pendekatan. Dengan mengembangkan penelitian selanjutnya, akan memberikan wawasan yang lebih dalam tentang karakteristik kopi arabika dan berkontribusi pada pemahaman yang lebih luas yang memengaruhi kualitas kopi.

DAFTAR PUSTAKA

[1] V. Sari, F. Firdausi, and Y. Azhar, "Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes," *Edumatic J. Pendidik. Inform.*, vol. 4, no. 2, pp. 1–9, 2020, doi: 10.29408/edumatic.v4i2.2202.

[2] K. Ciptady, M. Harahap, J. Jonvin, Y. Ndruru, and I. Ibadurrahman, "Prediksi Kualitas Kopi Dengan Algoritma Random Forest Melalui Pendekatan Data Science," *Data Sci. Indones.*, vol. 2, no. 1, 2022, doi: 10.47709/dsi.v2i1.1708.

[3] A. Bode, "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017, doi: 10.33096/ilkom.v9i2.139.188-195.

[4] B. Ginting and F. Riandari, "Implementasi Metode K-Means Clustering Dalam Pengelompokan Bibit Tanaman Kopi Arabika," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 3, no. 2, pp. 151–157, 2020, doi: 10.32672/jnkti.v3i2.2381.

[5] A. Saputra, "Optimalisasipenentuan Sumber Pasokan Kopi Arabika Gayo Melalui Pendekatan Hirarchicalclustering Data Mining," *J. Optim.*, vol. 1, no. 1, pp. 24–31, 2018, doi: 10.35308/jopt.v1i1.166.

[6] P. Katemba and R. K. Djoh, "Prediksi Tingkat Produksi Kopi Menggunakan Regresi Linear," *J. Ilm. FLASH*, vol. 3, no. 1, pp. 42–51, 2017, [Online]. Available: http://jurnal.pnk.ac.id/index.php/flash/article/view/136.

[7] I. Ifongki, "Penerapan Data Mining Menggunakan Algoritma C4.5 Tehadap Pengaruh Penjualan Kopi Pada Pt. Jpw Indonesia," *J. Sist. Inf. dan Inform.*, vol. 3, no. 1, pp. 40–54, 2020, doi: 10.47080/simika.v3i1.836.

[8] H. A. Sihombing and I. C. Buulolo, "Pengenalan Buah Kopi Berdasarkan Parameter Warna Menggunakan Algoritma Backpropagation Dan Algoritma Support Vector Machine (Svm)," *Seminastika*, vol. 3, no. 1, pp. 26–32, 2021, doi: 10.47002/seminastika.v3i1.234.

[9] F. Nurchalifatun, "Penerapan Metode Asosiasi Data Mining Menggunakan Algoritma Apriori Untuk Mengetahui Kombinasi Antar Itemset Pada Pondok Kopi," *Data Min.*, vol. 99, no. 99, pp. 1–10, 2015.

[10] R. Sarker, H. Abbass, and C. Newton, *Introducing Data Mining and Knowledge Discovery*, no. July. 2011.

[11] C. Zhang and J. Han, *Data Mining and Knowledge Discovery*. 2021.

[12] G. A. Marcoulides, *Discovering Knowledge in Data: an Introduction to Data Mining*, vol. 100, no. 472. 2014.

[13] D. Marlina, N. F. Putri, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," vol. 4, no. 2, pp. 64–71, 2018.

[14] D. Ayu, I. Cahya, D. Ayu, and K. Pramita, "Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," vol. 9, no. 3, 2019.