Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.1, July 2024, pp. 62~70
e-ISSN: 2548-6861

Implementation of K-Means, Hierarchical, and BIRCH Clustering Algorithms to Determine Marketing Targets for Vape Sales in Indonesia

Justin Laurenso [1]*, Danny Jiustian [2]*, Felix Fernando [3]*, Vartin Suhandi [4]*, Theresia Herlina Rochadiani [5]*
* Informatika, Universitas Pradita
justin.laurenso@student.pradita.ac.id [1], danny.jiustian@student.pradita.ac.id [2], felixfernando@student.pradita.ac.id [3],
vartin.suhandi@student.pradita.ac.id [4], theresia.herlina@pradita.ac.id [5]

Article Info

Article history:

Received 2022-11-24
Revised 2024-04-17
Accepted 2024-05-07

Keyword:

Algoritma K-Means,
BIRCH,
Clustering,
Elbow Method,
Silhouette Score.

ABSTRACT

In today's era, smoking is a common thing in everyday life. Along with the development of the times, an innovation emerged, namely the electric cigarette or vape. Electric cigarettes or vapes use electricity to produce vapor. The e-cigarette business is very promising in today's business world due to the consistent increase in market demand. However, determining the target buyer is one of the things that is quite important in determining the success of a business. In this analysis, the background of each region in Indonesia has different diversity; therefore, observation of data is needed to find out which regions in Indonesia have the potential to increase marketing based on profits (margins) to support the target market analysis process so that companies do not suffer losses and increase business success. In this study, the analysis will be carried out using vape quantity, margin, and purchasing power data in each region, which is processed using 3 algorithms: K-Means, Hierarchical, and BIRCH. The results of the clustering of the three algorithms produce two clusters. The K-means, Hierarchical, and BIRCH algorithms produce the same clusters: a potential cluster consisting of 18 cities and a non-potential cluster consisting of 45 cities. To see the performance of the model results, an evaluation was carried out using the Silhouette score, Davies Bouldin, Calinski Harabasz, and Dunn index, which obtained results of 0.765201, 0.376322, 315.949434, and 0.013554. From these results, it can be concluded that the clustering results are not too good and not too bad because the greater the Silhouette Score, Calinski Harabasz, and Dunn Index value, the better the clustering results while for Davies Bouldin the smaller the value means the better the clustering results.

I. PENDAHULUAN

Transaksi merupakan sebuah proses terjadinya sebuah perpindahan keuangan yang biasanya terjadi pada perusahaan seperti melakukan penjualan, pembelian, hingga melakukan pemberian upah kepada pekerja di perusahaan dan juga mencakup hal lainnya seperti melakukan pembayaran untuk berbagai macam tagihan, menurut Ramadhani,Bunyamin dan fitriani[1].Sebuah Perusahaan melakukan banyak sekali kegiatan transaksi. Kegiatan melakukan penjualan digunakan oleh sebuah perusahaan untuk mendapatkan sebuah keuntungan (margin) dalam setiap Transaksi yang dilakukan. Kegiatan Penjualan bukan hanya mengenai bagaimana sebuah perusahaan mendapatkan sebuah laba atau keuntungan. Penjualan juga merupakan sebuah strategi untuk melakukan hubungan antara pembeli dengan perusahaan[2]. Menurut Widharta dan Sugiharto melakukan sebuah penjualan adalah sebuah kegiatan dimana seorang pembeli akan dipengaruhi dengan cara mengarahkan agar apa yang dibutuhkan oleh pembeli dapat sesuai dengan produk yang ditawarkan. Dengan adanya kesepakatan antara kebelah dua pihak yang mendapatkan keuntungan masing-masing maka itu juga disebut sebagai sebuah penjualan[2].

Dalam Kegiatan langsung di lapangan sebuah penjualan dapat dipengaruhi oleh beberapa hal seperti kondisi dan kemampuan menjual, kondisi pasar, modal, kondisi

organisasi perusahaan serta faktor lain seperti periklanan, peragaan dan kampanye[2]. Kondisi pasar meliputi dari serangkaian faktor seperti jenis pasar, kelompok pembeli, segmen pasar,daya beli, frekuensi pembelian dan kebutuhannya.

Menentukan target pasar merupakan salah satu hal yang cukup penting untuk dipertimbangkan dalam menentukan keberhasilan sebuah penjualan. Target penjualan dapat dilakukan dengan melakukan analisis mengenai faktor - faktor yang ada di dalam sebuah kegiatan penjualan seperti daya beli sebuah daerah agar produk yang dilempar dapat sesuai dengan pasar yang ada. selain itu dengan adanya daya beli dibutuhkan sebuah Kelompok pembeli yang memiliki kebutuhan yang sesuai dengan produk yang ditawarkan dan beberapa hal lainnya yang ada di dalam target pasar. Untuk melakukan analisis observasi pada lapangan diperlukan data-data untuk mengetahui sebesar apa potensi yang ada di daerah tersebut.

Vape adalah salah satu dari sebagian banyak teknologi dari rokok elektrik yang sedang diminati oleh banyak orang saat ini, *vape* terbagi menjadi 2 bagian *Mod* dan *atomizer* dimana *Mod* merupakan badan yang digunakan sebagai sumber daya utama yang dialirkan kepada a*tomizer*, dimana atomizer adalah sebuah penampung *liquid* yang berisi kawat dan kapas[3].

Vape Store adalah usaha dengan memberikan tempat dimana produk-produk *vape* disediakan, seperti *Mod, atomizer, liquid* dan lain-lainnya. biasanya *vape store* merupakan *offline store* yang dimana orang akan datang untuk membeli secara langsung untuk merasakan melakukan pengaturan pada *vape* yang dimiliki oleh pembeli.

PT X merupakan distributor yang bergerak di dalam dunia bisnis *vape*. perusahaan ini sudah bergerak cukup lama sekitar lima tahun hingga saat ini. Perusahaan ini sudah memiliki puluhan hingga ratusan *vape store* sebagai pembeli tetap selama perusahaan bergerak.Perusahaan ini sudah melakukan ribuan Penjualan setiap tahunnya. target pasar yang dibentuk sudah ada tetapi jauh dari sempurna. Perusahaan ini memiliki kesulitan untuk melakukan pemilihan daerah untuk melakukan mengembangkan sayap yang lebih besar lagi.

Data science secara sederhana merupakan sebuah seni dan kecerdasan agar mendapatkan sebuah pengetahuan melalui sebuah data, sementara secara luas ilmu data atau *data science* merupakan rangkaian sebuah data diambil dan digunakan untuk mendapatkan sebuah pengetahuan agar dapat melakukan sebuah keputusan, menerka masa yang akan datang, memahami masa yang sudah lewat atau yang akan datang dan menciptakan sebuah inovasi dalam industri maupun produk[4]. dalam data science terdapat berbagai macam pemodelan atau algoritma antara lain seperti *Association, Clustering* dan lainnya.

Clustering adalah adalah metode pada data mining yang digunakan untuk melakukan pengelompokkan data yang memiliki karakteristik mirip ataupun sama menjadi sebuah cluster sedangkan untuk data yang memiliki karakteristik berbeda akan dikelompokkan menjadi *cluster* baru[5]. Metode *clustering* memiliki banyak sekali *modeling* yang digunakan seperti Algoritma K-means, Algoritma Hierarchical, Algoritma BIRCH dan masih banyak yang lainnya. Setiap model dapat menampilkan hasil yang berbeda-beda.

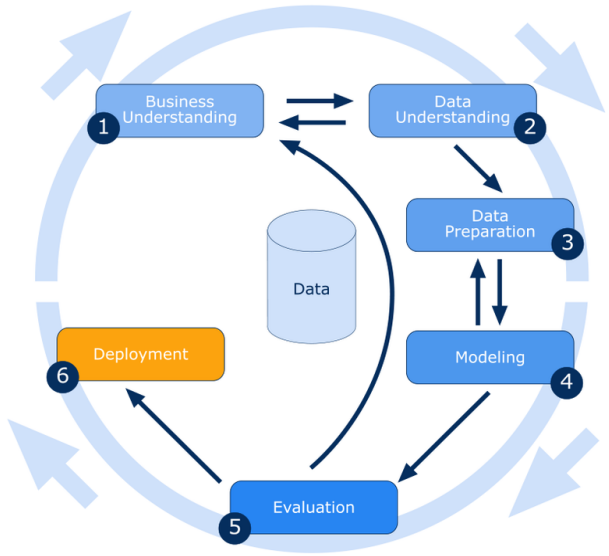
Dalam penelitian yang dilakukan oleh [6], dalam penelitian yang berjudul *Customer Segmentation Using K-Means Algorithm As A Basis For A Marketing Strategy In The Store* Rumah Tua VAPE, dimana pada penelitian tersebut menggunakan data yang berasal dari toko Bernama Rumah Tua. Dalam penelitian ini, dilakukan proses *clustering* untuk mendapatkan segmentasi pelanggan yang menghasilkan 2 *cluster*. *Cluster* 1 sebesar 60% dan *cluster* 2 sebesar 40%.

Kemudian penelitian oleh [7] tentang Segmentasi Pelanggan Restoran Menggunakan Metode *Clustering Simple K-Means* (Studi Kasus XYZ). Pada penelitian ini menggunakan data yang berasal dari survei yang telah dilakukan kepada 50 responden pengunjung restoran XYZ. Penelitian ini dilakukan untuk menentukan faktor perilaku pelanggan pada sebuah bisnis sehingga pemilik dapat menyesuaikan dari faktor tersebut dalam bisnisnya. Hasil yang didapatkan dari penelitian ini adalah 5 *cluster*. *Cluster* 1 mendapatkan persentase sebesar 15%, *cluster* 2 sebesar 11%, *cluster* 3 sebesar 25%, *cluster* 4 sebesar 28%, dan *cluster* 5 sebesar 21%. Sehingga hasil dari penelitian ini adalah *cluster* 4 memiliki hasil yang lebih dominan yaitu sebesar 28%.

Berdasarkan latar belakang masalah yang telah dikemukakan, penelitian ini dilakukan untuk membantu PT X untuk melakukan prediksi mengenai wilayah-wilayah di Indonesia yang memiliki potensi cukup tinggi untuk dilakukannya pemasaran di wilayah tersebut. Data yang digunakan untuk penelitian ini adalah data penjualan dua tahun terakhir dari PT X. Dengan menggunakan *data science* diharapkan dapat melakukan prediksi dan membantu dalam mengambil keputusan untuk peningkatan penjualan pada perusahaan tersebut.

II. METODE PENELITIAN

Penelitian ini menggunakan metodologi CRISP-DM *(Cross Industry Standard for Data Mining)* yang merupakan sebuah standarisasi dari teknik pengolahan data dan cukup umum untuk digunakan dalam pemecahan kasus yang diinginkan, CRISP-DM ini terbentuk pada awal tahun 1996 yang dibangun oleh beberapa perusahaan seperti *Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA.* Pada penelitian ini menggunakan tiga algoritma diantaranya ada K-Means, Hierarchical, dan Birch, di dalam penelitian ini dilakukan dengan 6 tahapan seperti yang ditunjukkan oleh Gambar 1[8].



Gambar 1. Tahapan Penelitian [9]

A. *Business Understanding*

Pada tahap ini merupakan tahap awal dari model CRISP-DM yang akan dikerjakan, dalam tahap ini akan dilakukan sebuah studi pustaka mengenai objek penelitian yang akan diteliti, agar hasil yang nantinya telah dilakukan dapat maksimal dan sesuai dengan tujuan dari proses ini, dengan adanya tahap ini diharapkan bisa menjadi efektif, beberapa solusi juga diharapkan sesuai dengan tujuan dari permasalahan yang telah dikemukakan agar dapat mencapai atau mempersiapkan strategi awal untuk kedepannya.

B. *Data Understanding*

Pada tahap kedua ini yaitu *data understanding*. Pada tahap *data understanding* ini akan dilakukan pengumpulan data awal yang dimana itu digunakan sebagai tolak ukur dari sebuah kualitas data yang di proses [8]. Secara garis besarnya itu digunakan untuk memeriksa data, sehingga dapat mengidentifikasi masalah yang ada dalam data tersebut. Beberapa kasus *clustering* dan pengelompokkan yang berkaitan dengan penjualan *vape* akan dikaji sehingga solusi yang ditawarkan diharapkan dapat menjadi efektif. Proses data mining ditawarkan sebagai solusi dengan tujuan agar dapat menyelesaikan sebuah permasalahn yang telah dikemukan.

C. *Data Preparation*

Pada tahap ini yaitu dengan dilakukannya preprocessing data yang telah ada agar bisa dilanjutkan ke tahap berikutnya yaitu tahap *modelling*, tahap ini juga membagi data secara umum menjadi dua data, yaitu data training dan data testing. Pada tahap ini juga terdapat beberapa tahapan, seperti :

1) *Data Selection*, pada proses ini berfungsi untuk menidentifikasi atribut yang akan dipakai dan yang tidak akan di pakai atau yang tidak mempunyai korelasi dengan tujuan dari penelitian ini.
2) *Data preprocessing*, pada proses ini berfungsi untuk mengidentifikasi dan membersihkan data dari missing values serta juga memeriksa data yang tidak memiliki sebuah konsistensi
3) *Transformation*, pada proses ini berfungsi untuk melakukan pengelompokan atribut ke dalam data yang baru yang kemudian dilakukan integrasi dengan hasil data processing, serta juga mentransformasi data yang akan diproses lebih lanjut.

D. *Modeling*

Pada tahap *modelling* ini memiliki tujuan untuk memilih model serta teknik yang terbaik untuk nantinya digunakan. pada tahap ini juga menerapkan metode clustering dengan algoritma yang akan digunakan di antaranya *K-Means*, *Hierarchical*, dan *BIRCH*..

Algoritma *K-Means* merupakan salah satu bentuk dari pengelompokan data yang sederhana, K-means juga sebuah algoritma yang membutuhkan sebuah parameter input sebanyak k dan membagi dengan sekumpulan n objek yang ada di dalam *cluster* k. Kemiripan dari setiap anggota terhadap *cluster* dapat diukur dengan pendekatan objek dengan nilai *mean* pada setiap cluster yang disebut sebagai *centroid cluster*. Konsep dasar dari k-means ini juga mencari pusat dari cluster secara iteratif, pusat dari *cluster* juga ditetapkan berdasarkan jarak dari setiap data ke pusat *cluster*. *K-Means* juga merupakan sebuah metode clustering dengan proses *modeling* tanpa adanya pengawasan dari beberapa golongan dan masing masing golongan sendiri juga mempunyai keunikannya sendiri[10].

Algoritma *Hierarchical* merupakan salah satu teknik *clustering* yang membentuk sebuah hirarki atau berdasarkan tingkatan dalam data yang nantinya menyerupai struktur pohon. Dengan demikian sebuah proses pengelompokan dapat dilakukan secara bertingkat atau secara bertahap sesuai dengan data yang dimiliki. Pada umumnya, metode ini digunakan dengan jumlah data yang tidak terlalu banyak serta juga *cluster* yang nantinya akan dibentuk juga belum diketahui. Algoritma Hierarchical ini juga memiliki dua jenis pengelompokkan data yaitu ada *agglomerative* dan *divisive*, *agglomerative* merupakan sebuah strategi yang digunakan dalam pengelompokan yang dimulai dengan setiap objek dalam sebuah *cluster* yang terpisah kemudian akan semakin membesar, sedangkan *divisive* merupakan sebuah strategi yang digunakan untuk pengelompokan yang dimulai dari semua objek yang dijadikan *cluster* tunggal yang kemudian dipisahkan sampai setiap objek berapa dalam sebuah cluster yang terpisah[11].

Algoritma *BIRCH* (*Balanced Iterative Reducing and Clustering Using Hierarchies*) menggunakan konsep *Cluster Feature* (CF) yang meringkas sebuah informasi mengenai subcluster - subcluster yang ada sehingga dapat mengurangi skala dari sebuah clustering. Algoritma BIRCH memiliki keunggulan dalam menangani data yang besar dan akurat karena dapat mengatasi sebuah *outliers* atau pencilan. Area kepadatan dari suatu data akan menentukan apakah data

tersebut dianggap sebagai outlier atau dianggap subcluster. Jika sebuah area padat dengan data, maka akan dianggap sebagai subcluster, sedangkan area yang sangat jarang akan dianggap sebagai outlier[12]. Algoritma BIRCH ini memiliki dua cara kerja yaitu dengan birch scan atau membaca, dimana data yang di bangun akan masuk kesalam sebuah inisial memory CF tree, yang akan dipandang sebagai kompresi multilevel dari data. sedangkan pada cara kerja kedua adalah dengan menggunakan sebuah penyeleksian terhadap algoritma yang akan dipakai dalam bentuk cluster leaf node dari CF tree [13].

E. Evaluation

Pada tahap *evaluation* ini akan menganalisis hasil dari pemodelan yang telah digunakan sebelumnya. Pada metode *evaluation* ini menggunakan metode *silhouette score, davies-bouldin, calinski harabasz,* dan *dunn index. Silhouette score* merupakan sebuah metode yang digunakan untuk melihat sebuah kualitas dan kekuatan sebuah *cluster*, metode ini menggabungkan dari metode separasi dan kohesi [14], Apabila menggunakan metode ini maka nilai yang nantinya akan optimal harus semakin tinggi nilainya. *Davies-bouldin* merupakan sebuah teknik untuk menghitung rata rata dari nilai pada setiap titik data, perhitungan nilai dari setiap titik ini akan dibagi dengan jarak antara kedua titik pusat pada sebuah *cluster* yang berguna sebagai *separation*[15]. Apabila menggunakan ini maka untuk menentukan yang terbaik yaitu dengan melihat nilai yang semakin kecil. *calinski harabasz* merupakan sebuah metode yang digunakan untuk menghitung perbandingan antara nilai dari *Sum of Square between cluster (SSB)* sebagai separation dan nilai *Sum of Square within cluster (SSW)* sebagai *compactness* yang nantinya akan di kalikan dengan faktor dari normalisasi. Apabila menggunakan metode ini maka untuk menentukan nilai terbaik ditentukan dengan nilai yang semakin besar., dan *dunn index* adalah metode untuk mengukur validasi pengelompokan internal yang mengukur tingkat kekompakan klaster dan tingkat pemisahan antar klaster[16]. *Dunn index* merupakan sebuah metode yang menghitung sebuah nilai minimum dari sebuah perbandingan antara nilai sebuah fungsi di kedua cluster yang berguna sebagai *separation*. Apabila menggunakan metode ini maka dapat dikatakan yang terbaik yaitu dengan nilai yang dihasilkan semakin besar [17].

Dengan menggunakan metode tersebut maka dapat melakukan uji kualitas dari setiap model yang didapatkan, dengan adanya *evaluation* ini diharapkan model yang dibangun sudah sesuai dengan tujuan yang telah ditentukan dari awal.

F. Deployment

Pada tahap ini merupakan tahap terakhir dari penelitian, pada tahap ini maka akan mencari tingkat penyebaran terhadap hasil dari seluruh tahap yang telah dikerjakan dalam penelitian ini. Pada tahap *Deployment* akan dilakukan dengan cara membuat laporan presentasi dari yang sudah diteliti sebelumnya dengan berdasarkan pemodelan dan *evaluation* yang didapatkan. Hasil yang nantinya didapatkan dapat digunakan untuk menentukan tindakan selanjutnya dalam Penjualan *Vape* di Indonesia [12].

III. HASIL DAN PEMBAHASAN

Hasil dari penelitian yang telah dilakukan adalah melakukan pengelompokan *customer* dalam Upaya Menentukan Fokus *Target* Penyebaran Penjualan Vape di Indonesia dengan menggunakan algoritma k-means *clustering*, *hierarchical clustering*, dan BIRCH *clustering*. Hasil pengelompokkan mendapatkan 2 cluster yaitu berpotensi dan tidak berpotensi. Selanjutnya hasil dari pengelompokkan tersebut divisualisasikan menggunakan library matplotlib.pyplot dan pandas, serta dievaluasi menggunakan library sklearn yaitu *silhouette coefficient, davies bouldin, calinski harabasz*, dan *dunn index*. Seluruh tahapan menggunakan bahasa pemrograman python.

A. Business Understanding

Pada tahapan *business understanding* dilakukan pemahaman tujuan kebutuhan berdasarkan penilaian bisnis. Kemudian diubah menjadi sebuah rencana awal yang dirancang untuk mencapai tujuan. Berikut langkah dalam pemahaman bisnis, yaitu :

1) Determine Business Objectives

Dalam penelitian ini, yang menjadi target pengguna yaitu distributor *vape*. Adapun dari perspektif bisnis, yang dibutuhkan oleh pengguna untuk membantu dalam menentukan fokus target penyebaran penjualan *vape* adalah grafis plot berdasarkan jumlah penjualan, keuntungan penjualan, dan daya beli *vape* di Indonesia.

2) Assess Situation

Adapun kondisi saat ini dilihat dari fakta di lapangan menunjukkan bahwa Distributor *Vape* belum mengetahui kota mana saja yang memiliki penjualan, keuntungan, dan daya beli tertinggi sehingga tidak mengetahui fokus target kota yang berpotensi dilakukan promosi.

3) Determine Data Mining Goals

Tujuan dari data mining yang dilakukan dalam penelitian ini yaitu melakukan pengklasteran kota dengan menggunakan metode *clustering* dengan menerapkan beberapa algoritma, yaitu k-means, hierarchical, dan BIRCH yang akan divisualisasikan kemudian untuk mempermudah dalam melihat penjualan, keuntungan, dan daya beli yang paling tinggi.

4) Plan Activities

Guna mencapai tujuan data mining dalam pengklasteran kota di Indonesia, tools yang akan digunakan yaitu *Google Collaboratory* dengan bahasa pemrograman python.

B. Data Understanding

Tahapan berikutnya adalah tahap pemahaman data awal, mengenal, dan memahami data yang dimiliki serta melakukan sebuah analisis pada data tersebut. Data yang digunakan pada penelitian ini adalah Data Penjualan *Vape* yang bersumber

dari distributor *vape*. Data yang digunakan untuk analisis yaitu data pada tahun 2021 hingga 2022. Contoh data dapat dilihat pada Tabel I.

TABEL I
DATA PENJUALAN VAPE

Qty	Modal	Jual	Diskon	Deleted At	Name	Sku	Nama Produk	Alamat
1	45000	55000	0	-	Dika tangerang vapers (new)	L1117	saltnic ELO soda lemon lime 15ml	Perumahan Sudirman, Blok A9 No 14 Gg. Demak, Banten
1	45000	55000	0	-	Dika tangerang vapers (new)	L987	saltnic ELO strawberry bubblegum 15ml	Perumahan Sudirman, Blok A9 No 14 Gg. Demak, Banten
1	80000	95000	0	-	Dika tangerang vapers (new)	L1187	saltnic KOPICINO 30ml 95000	Perumahan Sudirman, Blok A9 No 14 Gg. Demak, Banten
3	23000	28000	0	2022-06-04 12:00:54	Dika tangerang vapers (new)	OCC039	Coil Vinci Voopoo 0.2 (Drag)	Perumahan Sudirman, Blok A9 No 14 Gg. Demak, Banten
4	40000	55000	0	-	Dika tangerang vapers (new)	L1136	saltnic SOJU v1 original 30ml	Perumahan Sudirman, Blok A9 No 14 Gg. Demak, Banten

Untuk memahami data yang akan diolah dalam penelitian ini, dilakukan pendeskripsian data mengenai atribut, tipe data, dan keterangan data. Hasil deskripsi data dapat dilihat seperti pada Tabel II.

TABEL II
DESKRIPSI DATA PENJUALAN VAPE

Atribut	Tipe Data	Keterangan
QTY	*Float*	Jumlah produk yang terjual
MODAL	*Float*	Harga beli produk
JUAL	*Float*	Harga jual produk
Diskon	*Float*	Diskon produk
DELETED AT	*Object*	Berisi tanggal produk tersebut dibatalkan penjualannya
NAME	*Object*	Nama dalam *invoice*
SKU	*Object*	Kode unik setiap produk
NAMA PRODUK	*Object*	Nama produk yang sedang dijual
ALAMAT	*Object*	Alamat pembeli

C. *Data Preparation*

Pada tahap data preparation dilakukan kegiatan persiapan data dengan membangun suatu data agar dapat menyesuaikan dengan kebutuhan data pemodelan dari data mentah yang belum diolah.

1) *Data Selection*

Data yang diambil yaitu data Penjualan *Vape* di Indonesia. Data ini disimpan dalam bentuk excel, lalu dibersihkan dengan menghapus atribut-atribut yang tidak diperlukan pada penelitian ini. Dalam data Penjualan *Vape* akan dilakukan penghapusan atribut NAME.

```
new_data = data.drop(['NAME'],axis=1)
```

2) *Data Preprocessing*

Berdasarkan tahap *data understanding* dan verifikasi data menunjukkan bahwa adanya data dengan kualitas tidak baik dan harus diperbaiki. Kemudian langkah selanjutnya yaitu memilih data yang DELETED AT nya tidak terisi, karena data yang digunakan merupakan data penjualan maka data DELETED AT tidak boleh terisi, jika terisi maka data tersebut tidak valid.

```
filter_notnull_del_at = new_data[(new_data['DELETED AT']!= "-")].index
new_data.drop(filter_notnull_del_at,inplace=True)
```

Selanjutnya akan dilakukan penghapusan data yang memiliki nilai kosong atau *missing value* dan melakukan pengecekan nilai *missing value* berikut.

```
filter_null = new_data.dropna(axis=0)
filter_null
```

\

```
filter_null.isnull().sum()
```

```
QTY           0
MODAL         0
JUAL          0
Diskon        0
DELETED AT    0
SKU           0
NAMA PRODUK   0
ALAMAT        0
dtype: int64
```

Kemudian akan dilakukan pengelompokkan data. Data akan dikelompokkan berdasarkan kota/kabupaten dan kategori produk dengan upaya mendapatkan atribut yang akan

digunakan untuk tahap pemodelan yaitu QTY, MARGIN, dan DAYA BELI. Hasil pengelompokkan data dapat dilihat pada Tabel III.

TABEL III
DATA INTEGRATION

Kota	Qty	Margin	Daya Beli
Bali	158	14942500	20685000
Banda Aceh	294	39175500	46912000
Bandung	1517	111772200	149263500
Banjar	2198	163244400	195765100
Banjarmasin	19	1875000	3285000

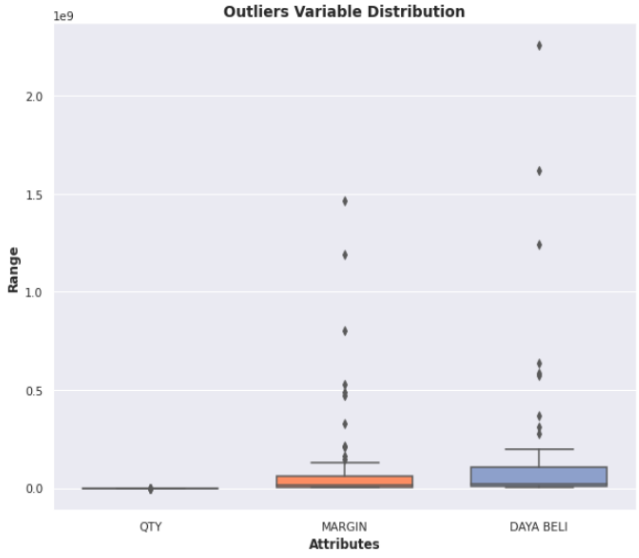
3) Data Transformation

Data yang digunakan penelitian ini adalah QTY, MARGIN, dan DAYA BELI sehingga data yang digunakan tidak memiliki data kategorik. Kemudian akan dilakukan penanganan data numerik agar mempunyai rentang nilai yang sama dan agar tidak terdapat data yang terlalu besar sampai terlalu kecil menggunakan *StandardScaler*.

D. Modeling

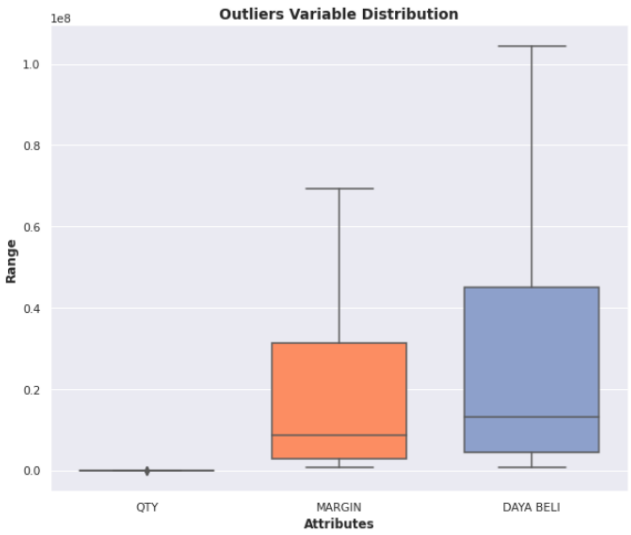
Tahap selanjutnya adalah *Modeling* yang merupakan sebuah fase penerapan *data mining* yang digunakan sesuai dengan tujuan. Pada penelitian ini teknik *data mining* yang digunakan yaitu clustering dengan menggunakan beberapa algoritma, yaitu K-Means, Hierarchical, dan BIRCH.

Tahap pertama dalam fase ini yaitu melakukan pengecekan nilai *outliers*. *Outliers* atau pencilan adalah suatu observasi pada sekumpulan data yang memiliki pola atau nilai yang berbeda dari kumpulan data lainnya. Proses ini dilakukan karena algoritma K-Means, Hierarchical, dan BIRCH sensitif terhadap sebuah *outliers*. Setelah melakukan pengecekan *outliers*, maka dapat disimpulkan bahwa atribut MARGIN dan DAYA BELI memiliki *outliers*. Seperti pada Gambar 2.



Gambar 2. Hasil pengecekan *Outliers* pada setiap atribut

Tahap kedua dalam fase ini adalah menghapus *outliers* pada setiap atribut agar data yang akan digunakan bersih dari *outliers.* Setelah dilakukan penghapusan *outliers*, data telah bersih dan siap digunakan untuk modeling. Seperti pada Gambar 3.

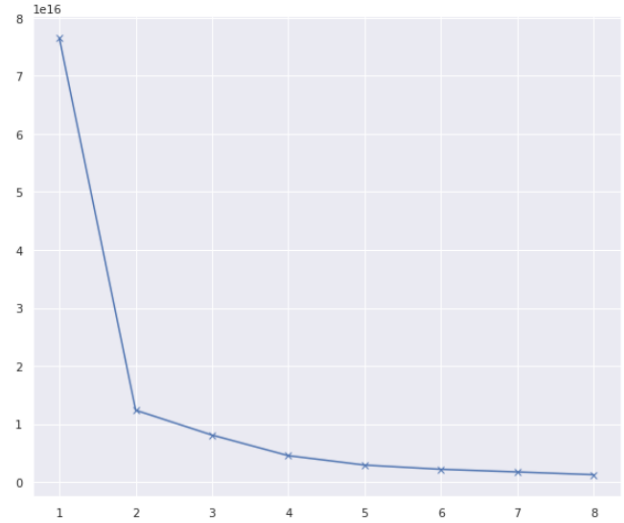


Gambar 3. Hasil setelah melakukan penghapusan *Outliers*

Tahap ketiga dalam fase ini adalah menentukan jumlah *cluster*. Untuk menentukan jumlah cluster pada K-Means akan menggunakan metode *elbow*, Hierarchical akan menggunakan metode *dendogram* dan BIRCH akan menggunakan *silhouette score*. Setelah mendapatkan jumlah *cluster* yang optimal, akan dilanjutkan ke tahap modeling.

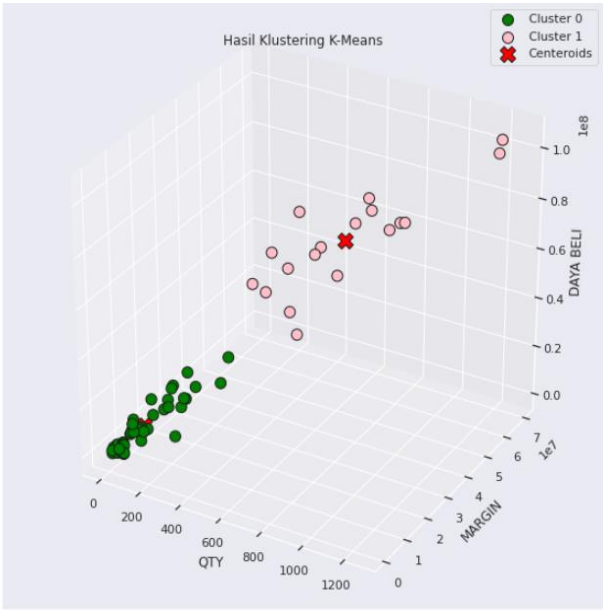
1) K-Means

Dalam mencari jumlah *Cluster* yang optimal, akan digunakan metode *elbow*. Pada Gambar 4 terlihat garis mengalami patahan yang membentuk elbow atau siku pada saat K = 2. Maka dapat disimpulkan 2 adalah jumlah cluster yang optimal untuk algoritma K-Means.



Gambar 4. Metode *Elbow*

Setelah didapatkan jumlah *cluster* yang optimal, maka akan dilakukan modeling menggunakan algoritma K-Means. Hasil dari proses *clustering* akan ditampilkan dalam bentuk *3D Scatter Plot* seperti pada Gambar 5.



Gambar 5. Hasil *Clustering* menggunakan *K-Means*

Pada Tabel IV menunjukkan 5 hasil dari proses *clustering* data menggunakan algoritma *K-Means*.

TABEL IV
HASIL CLUSTERING

Kota	Qty	Margin	Daya Beli	Cluster
Bali	158	14942500	20685000	0
Banda Aceh	294	39175500	46912000	1
Banjarmasin	19	1875000	3285000	0
Banyumas	129	3852900	9892100	0
Barru	9	1230000	2455000	0

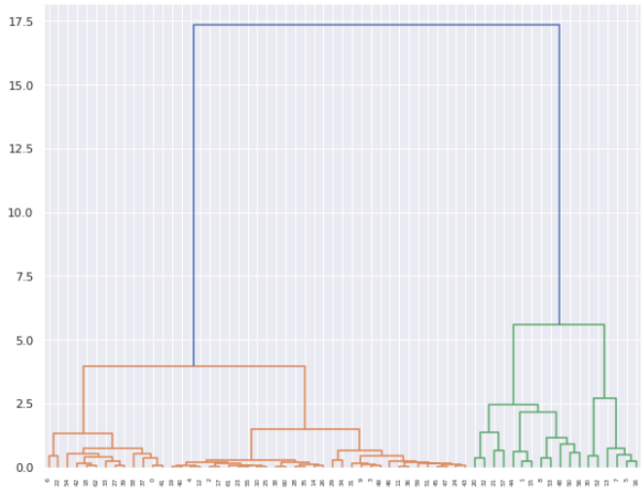
Kemudian untuk membedakan hasil pengelompokkan yang terbentuk maka dilakukan profilisasi dengan mencari nilai rata- rata dari setiap atribut. Penamaan *cluster* akan ditentukan dari atribut yang digunakan dalam *modeling* yaitu QTY, MARGIN, dan DAYA BELI. Sehingga dapat diketahui karakteristik dari setiap *cluster* yang terbentuk. Pada *cluster* 0 terdapat rata - rata dengan QTY terjual sebesar 111.5, MARGIN 6.221.737.9, dan DAYA BELI 9.932.363.3. *Cluster* 0 memiliki karakteristik dengan QTY rendah, MARGIN rendah, dan DAYA BELI rendah. Berdasarkan hal tersebut, maka *cluster* 0 sebagai *cluster* tidak berpotensi.

Pada *cluster* 1 terdapat rata - rata dengan QTY terjual sebesar 683.8, MARGIN 46.133.988.9, dan DAYA BELI 68.191.147.2. *Cluster* 1 memiliki karakteristik dengan QTY tinggi, MARGIN tinggi, dan DAYA BELI tinggi. Berdasarkan hal tersebut, maka *cluster* 1 sebagai *cluster* berpotensi.

CLUSTER	QTY	MARGIN	DAYA BELI
0	111.5	6221737.8	9932363.3
1	683.8	46133988.9	68191147.2

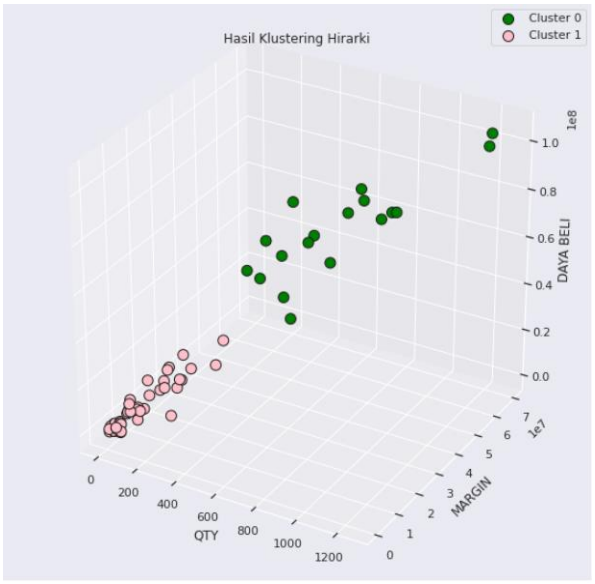
2) Hierarchical

Dalam mencari jumlah *Cluster* yang optimal, akan digunakan metode *dendogram*. Gambar 6 terlihat terbentuk beberapa hirarki atau tingkatan antar objek dan mendapatkan hasil yaitu 2 *cluster*.



Gambar 6. Dendogram

Setelah didapatkan jumlah *cluster* yang optimal, maka akan dilakukan modeling menggunakan algoritma Hierarchical. Hasil dari proses *clustering* akan ditampilkan dalam bentuk *3D Scatter Plot* gambar 7.



Gambar 7. Hasil *Clustering* menggunakan *Hierarchical*

Tabel V menunjukkan 5 hasil dari proses clustering data menggunakan algoritma *Hierarchical*.

TABEL V
HASIL CLUSTERING

Kota	Qty	Margin	Daya Beli	Cluster
Bali	158	14942500	20685000	1
Banda Aceh	294	39175500	46912000	0
Banjarmasin	19	1875000	3285000	1
Banyumas	129	3852900	9892100	1
Barru	9	1230000	2455000	1

Kemudian untuk membedakan hasil pengelompokkan yang terbentuk maka dilakukan profilisasi dengan mencari nilai rata- rata dari setiap atribut. Penamaan *cluster* akan ditentukan dari atribut yang digunakan dalam *modeling* yaitu QTY, MARGIN, dan DAYA BELI. Pada Gambar 15 dapat diketahui karakteristik dari setiap *cluster* yang terbentuk.

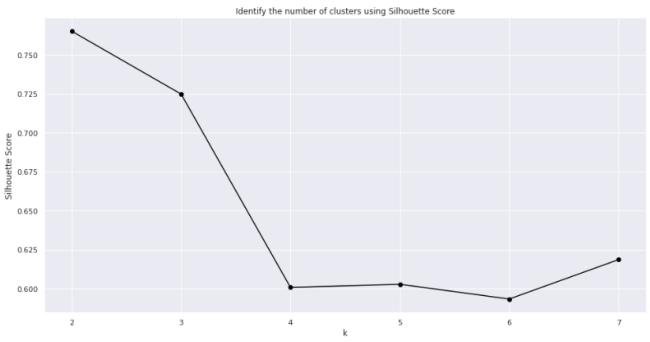
Pada *cluster* 0 terdapat rata - rata dengan QTY terjual sebesar 683.8, MARGIN 46.133.988.9, dan DAYA BELI 68.191.147.2. *Cluster* 1 memiliki karakteristik dengan QTY tinggi, MARGIN tinggi, dan DAYA BELI tinggi. Berdasarkan hal tersebut, maka *cluster* 10 sebagai *cluster* berpotensi.

Pada *cluster* 1 terdapat rata - rata dengan QTY terjual sebesar 111.5, MARGIN 6.221.737.9, dan DAYA BELI 9.932.363.3. *Cluster* 0 memiliki karakteristik dengan QTY rendah, MARGIN rendah, dan DAYA BELI rendah. Berdasarkan hal tersebut, maka *cluster* 1 sebagai *cluster* tidak berpotensi.

CLUSTER	QTY	MARGIN	DAYA BELI
0	683.8	46133988.9	68191147.2
1	111.5	6221737.8	9932363.3

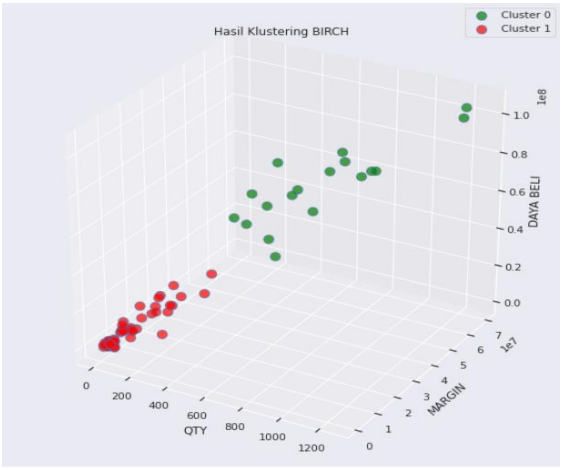
3) BIRCH

Dalam mencari jumlah *Cluster* yang optimal, akan digunakan *silhouette score*. Gambar 8 terlihat bahwa K = 2 memiliki nilai *silhouette score* yang paling tinggi, maka 2 merupakan *cluster* yang optimal.



Gambar 8. Karakteristik *Cluster*

Setelah didapatkan jumlah *cluster* yang optimal, maka akan dilakukan modeling menggunakan algoritma BIRCH. Hasil dari proses *clustering* akan ditampilkan dalam bentuk *3D Scatter Plot* seperti pada gambar 9.



Gambar 9. Hasil *Clustering* menggunakan *BIRCH*

Tabel VI menunjukkan 5 hasil dari proses clustering data menggunakan algoritma *BIRCH*.

TABEL VI
HASIL CLUSTERING

Kota	Qty	Margin	Daya Beli	Cluster
Bali	158	14942500	20685000	1
Banda Aceh	294	39175500	46912000	0
Banjarmasin	19	1875000	3285000	1
Banyumas	129	3852900	9892100	1
Barru	9	1230000	2455000	1

Kemudian untuk membedakan hasil pengelompokkan yang terbentuk maka dilakukan profilisasi dengan mencari nilai rata- rata dari setiap atribut. Penamaan *cluster* akan ditentukan dari atribut yang digunakan dalam *modeling* yaitu QTY, MARGIN, dan DAYA BELI. Pada Gambar 19 dapat diketahui karakteristik dari setiap *cluster* yang terbentuk.

Pada *cluster* 0 terdapat rata - rata dengan QTY terjual sebesar 683.8, MARGIN 46.133.988.9, dan DAYA BELI 68.191.147.2. *Cluster* 1 memiliki karakteristik dengan QTY tinggi, MARGIN tinggi, dan DAYA BELI tinggi. Berdasarkan hal tersebut, maka *cluster* 10 sebagai *cluster* berpotensi.

Pada *cluster* 1 terdapat rata - rata dengan QTY terjual sebesar 111.5, MARGIN 6.221.737.9, dan DAYA BELI 9.932.363.3. *Cluster* 0 memiliki karakteristik dengan QTY rendah, MARGIN rendah, dan DAYA BELI rendah. Berdasarkan hal tersebut, maka *cluster* 1 sebagai *cluster* tidak berpotensi.

CLUSTER	QTY	MARGIN	DAYA BELI
0	683.8	46133988.9	68191147.2
1	111.5	6221737.8	9932363.3

E. Evaluation

Setelah melakukan tahapan *modeling* maka perlu dilakukan evaluasi untuk mengetahui tingkat keberhasilan pemodelan yang diterapkan apakah sudah sesuai dengan tujuan yang sudah dirancang pada tahap *business understanding.*

1) Evaluate Result

Hasil *clustering* dengan algoritma K-Means, Hierarchical, dan BIRCH akan dievaluasi dengan menggunakan *silhouette score, davies-bouldin, calinski harabasz*, dan *dunn index* dapat terlihat pada Tabel VII. Hasil evaluasi performa yang didapatkan dengan *cluster* 2 untuk algoritma K-Means, Hierarchical, dan BIRCH memiliki nilai yang sama, yaitu *silhouette score* 0.765201, *davies-bouldin* 0.376322, *calinski harabasz* 315.949434, dan *dunn index* 0.013554.

TABEL VII
HASIL EVALUASI

Algorithm	Silhouette Score	Davies Bouldin	Calinski Harabasz	Dunn Index
K-Means	0.765201	0.376322	315.949434	0.013554
Hierarchical	0.765201	0.376322	315.949434	0.013554
BIRCH	0.765201	0.376322	315.949434	0.013554

V. KESIMPULAN

Penerapan 3 algoritma *clustering*, yaitu *K-Means*, *Hierarchical*, dan *BIRCH* untuk menentukan target pemasaran penjualan vape di Indonesia berdasarkan keuntungan di setiap daerah dengan hasil yang sama mendapatkan 2 *cluster* yaitu *cluster* berpotensi dan *cluster* tidak berpotensi. *Cluster* berpotensi mendapatkan sebanyak 18 kota sedangkan *cluster* tidak berpotensi mendapatkan sebanyak 45 kota. Hasil evaluasi performa *clustering* ketiga algoritma yaitu, dengan *silhouette score* berada pada nilai 0.765201, *davies-bouldin* berada pada nilai 0.376322, *calinski harabasz score* berada pada nilai 315.949434 dan *dunn index* dengan nilai 0.013554. Visualisasi menggunakan library matplotlib dan seaborn dengan bahasa pemrograman python untuk memudahkan dalam memahami hasil dari *clustering* menggunakan *K-Means*, *Hierarchical*, dan *BIRCH*.

DAFTAR PUSTAKA

[1] A. R. Ramadhani, H. Bunyamin, and L. Fitriani, "Perancangan Aplikasi Persediaan Barang dan Transaksi Penjualan Barang di Alya Store," *Jurnal Algoritma*, vol. 13, no. 2, pp. 284–390, Feb. 2017, doi: 10.33364/algoritma/v.13-2.384.

[2] W. P. WIDHARTA, "Penyusunan Strategi Dan Sistem Penjualan Dalam Rangka Meningkatkan Penjualan Toko Damai," *Jurnal Strategi Pemasaran*, vol. 1, no. 2, pp. 1–15, 2013, [Online]. Available: https://publication.petra.ac.id/index.php/manajemen-pemasaran/article/view/720

[3] M. A. Budhi, "Pemilihan Lokasi Usaha Vaporstore Menggunakan Metode Weighted Product," *Jurnal Sistem dan Informatika (JSI)*, vol. 14, no. 1, pp. 9–15, Nov. 2019, doi: 10.30864/jsi.v14i1.230.

[4] I. P. Sari, "Implementasi Data Science dalam Ritel Online: Analisis Customer Retention dan Clustering Customer dengan Metode K-Means," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 5, 2021, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jsakti/article/view/333

[5] M. C. Untoro, L. Anggraini, M. Andini, H. Retnosari, and M. A. Nasrulloh, "Penerapan metode k-means clustering data COVID-19 di Provinsi Jakarta," *Teknologi*, vol. 11, no. 2, pp. 59–68, Apr. 2021, doi: 10.26594/teknologi.v11i2.2323.

[6] R. Fadly and I. Wahyudin, "Customer Segmentation Using K-Means Algorithm As A Basis For A Marketing Strategy In The Store Rumah Tua VAPE," 2020. [Online]. Available: https://iocscience.org/ejournal/index.php/mantik

[7] Z. G. Prastyawan, M. Ridho Bagaskara, and D. Fitriati, "SEGMENTASI PELANGGAN RESTORAN MENGGUNAKAN METODE CLUSTERING SIMPLE K-MEANS (STUDI KASUS XYZ)," 2018. [Online]. Available: https://conference.upnvj.ac.id/index.php/seinasikesi/article/view/75/pdf

[8] H. and H. R. and J. M. and J. J. Kurniadewi, "Pemetaan UMKM dalam Upaya Pengentasan Kemiskinan dan Penyerapan Tenaga Kerja Menggunakan Algoritma K-Means," *Journal of Applied Informatics and Computing*, vol. 6, pp. 113–119, Aug. 2022, [Online]. Available: https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/4227

[9] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *Journal of Applied Informatics and Computing*, vol. 5, no. 2, pp. 103–108, Oct. 2021, doi: 10.30871/jaic.v5i2.3200.

[10] M. N. Sutoyo, "Algoritma K-Means." [Online]. Available: https://fti.usn.ac.id/sinau/assets/files/KMeans.pdf

[11] Tri Binty N, "Algorithm Agglomerative Hierarchical Clustering — and Practice with R," Jun. 30, 2019. https://medium.com/@story.of.stats/algorithm-agglomerative-hierarchical-clustering-31d2cea14d9

[12] M. Aldo Shauma, Y. Purwanto, A. iovianty, and S. Komputer, "Deteksi Anomali Trafik Menggunakan Algoritma BIRCH dan DBSCAN menggunakan Streaming Traffic Anomaly Traffic Detection with BIRCH dan DBSCAN Algorithm for Streaming Traffic," *e-Proceeding of Engineering*, vol. 3, p. 5004, 2016, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/3132#

[13] Munar Hudaya, "14.pertemuan Keempatbelas." UNIKOM, Bandung, May 10, 2013. [Online]. Available: https://repository.unikom.ac.id/42096/1/14.Pertemuan%20Keempatbelas.doc

[14] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, pp. 102–109, Nov. 2019, doi: 10.31940/matrix.v9i3.1662.

[15] N. H. Harani, C. Prianto, and F. A. Nugraha, "Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 10, no. 2, pp. 133–146, Oct. 2020, doi: 10.34010/jamika.v10i2.2683.

[16] C.-E. ben Ncir, A. Hamza, and W. Bouaguel, "Parallel and scalable Dunn Index for the validation of big data clusters," *Parallel Comput*, vol. 102, p. 102751, May 2021, doi: 10.1016/j.parco.2021.102751.

[17] SRI ARISTA PANGGOLA, "Penentuan K Opitmum dengan nilai Dunn Index dan Davies Bouldin Index, Serta Evaluasi Model Cluster Menggunakan Average Within dan Average Between di R," Jul. 18, 2020. https://medium.com/@aristap/penentuan-k-opitmum-dengan-nilai-dunn-index-dan-davies-bouldin-index-serta-evaluasi-model-cluster-d1cde2f9e828