

## Penerapan Kecerdasan Buatan dalam Komunikasi Pemasaran: Segmentasi Pelanggan dengan Agglomerative Clustering

Received Feb. 8, 2025; Revised on Month. xx, 20xx; Accepted Month. xx, 20xx

### Abstrak

*Komunikasi pemasaran merupakan salah satu topik yang biasa diteliti dalam bidang ilmu komunikasi. Salah satu tujuan dari komunikasi pemasaran ini adalah untuk membangun hubungan yang baik kepada pelanggan. Cara membangun hubungan ini bisa dengan melakukan analisis perilaku pelanggan (customer behaviour). Hasil dari analisis ini bisa memiliki luaran berupa segmentasi pelanggan. Menghasilkan segmentasi pelanggan dengan analisis perilaku pelanggan secara konvensional atau tradisional memiliki batasan seperti banyaknya data pelanggan yang perlu untuk dianalisis, cukup lamanya waktu untuk dilakukan analisis tersebut, dan kompleksitas yang tinggi. Batasan tersebut terjadi karena data pelanggan pada era saat ini termasuk ke dalam kategori big data. Untuk melakukan analisis terhadap big data tersebut, maka diperlukan peran dan kajian bersifat multidisciplinary. Kedua bidang tersebut adalah menggabungkan ilmu komunikasi dengan data science. Tugas dari data science adalah melakukan analisis data dan melakukan prediksi dari big data ini menggunakan kecerdasan buatan (artificial intelligence). Salah satu subbagian pada kecerdasan buatan adalah pembelajaran mesin (machine learning). Pada contoh kasus segmentasi pelanggan, metode kecerdasan buatan yang dapat diimplementasikan adalah Clustering karena kita tidak hanya memprediksi hasil dari satu variabel, namun juga mengetahui karakteristik dan dapat melakukan pengelompokan pada masing-masing pelanggan. Penelitian ini bertujuan untuk membangun kecerdasan buatan dengan menggunakan algoritma Clustering, yaitu Agglomerative Clustering pada segmentasi pelanggan dan juga melakukan evaluasi hasilnya dengan metrik evaluasi Clustering. Pendekatan yang dilakukan adalah kuantitatif menggunakan workflow data science dengan data sekunder yang didapatkan dari sumber terbuka Kaggle. Metode penerapan kecerdasan buatan ini diharapkan agar dapat mempermudah kompleksitas dan mempercepat analisis segmentasi pelanggan yang dapat diterapkan pada komunikasi pemasaran agar dapat melakukan strategi pemasaran yang tepat sesuai dengan kelompok pelanggan.*

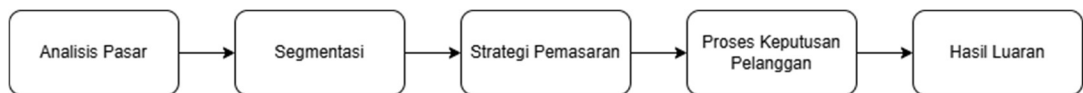
Kata kunci: kecerdasan buatan, komunikasi pemasaran, segmentasi pelanggan, clustering, agglomerative clustering

### PENDAHULUAN

Bidang Ilmu Komunikasi memiliki banyak topik yang bervariasi dalam penelitiannya. Salah satu topik tersebut adalah komunikasi pemasaran. Komunikasi pemasaran memainkan peranan yang penting tentang bagaimana perusahaan atau organisasi dapat menyampaikan pesan ke khalayak (audiens) termasuk dengan membangun hubungan yang baik, promosi produk, dan juga dalam pemerekan (*branding*). Untuk menyampaikan produk kepada pelanggan atau audiens, sangat penting untuk berfokus pada hasil produk yang bisa berkompetisi yang bisa ditinjau dari harga, penggunaan, kenyamanan, desain, dll. Pada era sekarang, komunikasi pemasaran tidak hanya dilakukan secara langsung tanpa media, tetapi juga secara tidak langsung atau menggunakan media. Topik-topik yang sering

dibahas dalam komunikasi pemasaran antara lain komunikasi pemasaran terintegrasi (*integrated marketing communication*) dan komunikasi pemasaran digital. Persamaan dari kedua topik tersebut adalah menyoroti pentingnya teknologi digital ataupun media digital sebagai strategi komunikasi pemasaran. Kemajuan teknologi memungkinkan bisnis untuk membuat berbagai macam tipe konten, komunikasi secara cepat dan *real-time* dari organisasi ke audiens atau sebaliknya, dan mencapai berbagai macam orang-orang dari berbagai dunia, dan dapat memfasilitasi komunikasi dua arah (Shankar dkk., 2022).

Seperti yang telah dibahas sebelumnya, membangun hubungan yang baik dengan pelanggan merupakan salah satu tujuan dari komunikasi pemasaran. Untuk mencapai tujuan tersebut, perusahaan atau perusahaan wajib menciptakan nilai atau *value* komunikasi secara efektif. Perusahaan harus dapat menguasai alat (*tools*) dan strategi pemasaran yang tepat di era pasar yang kompetitif. Alat atau strategi tersebut tidak hanya menjawab kebutuhan pelanggan, namun juga dapat meningkatkan loyalitas, kepercayaan, dan kepuasan pelanggan. Pengintegrasian dan pengoordinasian teknologi yang berbeda cukup penting untuk menghasilkan sinergi, memastikan bahwa pelanggan menyadari nilai unik yang diberikan oleh organisasi atau perusahaan. Komunikasi efektif yang berdasarkan kepada nilai pelanggan berpengaruh terhadap tingkat loyalitas pelanggan yang lebih tinggi, sehingga membantu perusahaan dengan menjadi lebih menonjol (*standout*) dalam pasar yang kompetitif. Merancang strategi komunikasi pemasaran yang efektif membutuhkan pemahaman tentang preferensi pelanggan, termasuk yang berkaitan dengan harga, kualitas produk, kenyamanan, dan hubungan emosional. Selain itu, umpan balik dan masukan dari pelanggan menjamin penerapan strategi komunikasi yang efisien. Pada akhirnya, komunikasi pemasaran bertindak sebagai jembatan antara pelanggan dan bisnis untuk memfasilitasi pertukaran nilai dalam membangun hubungan yang tidak hanya untuk keuntungan saja tetapi juga agar memiliki hubungan yang dapat bertahan lama (Kovanoviene dkk., 2021).



**Gambar 1.** Strategi komunikasi dan perilaku pelanggan

Salah satu cara untuk membangun hubungan yang baik dengan pelanggan adalah dengan analisis perilaku pelanggan atau *customer behavior analysis*. Perilaku pelanggan adalah studi tentang pelanggan yang mencakup individu atau kelompok, proses pelanggan tersebut dalam menggunakan produk, layanan, pengalaman, atau ide untuk memuaskan kebutuhan, serta juga meninjau dampak dari proses-proses tersebut pada pelanggan. Perilaku pelanggan bersifat kompleks dan multidimensi karena perilaku tersebut tidak hanya dipengaruhi oleh faktor internal tetapi juga eksternal. Perilaku pelanggan biasanya diaplikasikan bersama dengan strategi pemasaran yang melibatkan pengembangan, regulasi, dan efek. Proses strategi pemasaran dan perilaku pelanggan dapat dilihat pada Gambar 1. Segmentasi atau disebut juga segmentasi pelanggan juga termasuk dalam proses tersebut. Oleh karena itu, segmentasi dilakukan setelah analisis pasar. Setelah itu, hasil segmentasi digunakan untuk strategi pemasaran agar tujuan komunikasi pemasaran menjadi efektif. Segmentasi memiliki empat langkah, yaitu mengidentifikasi kebutuhan

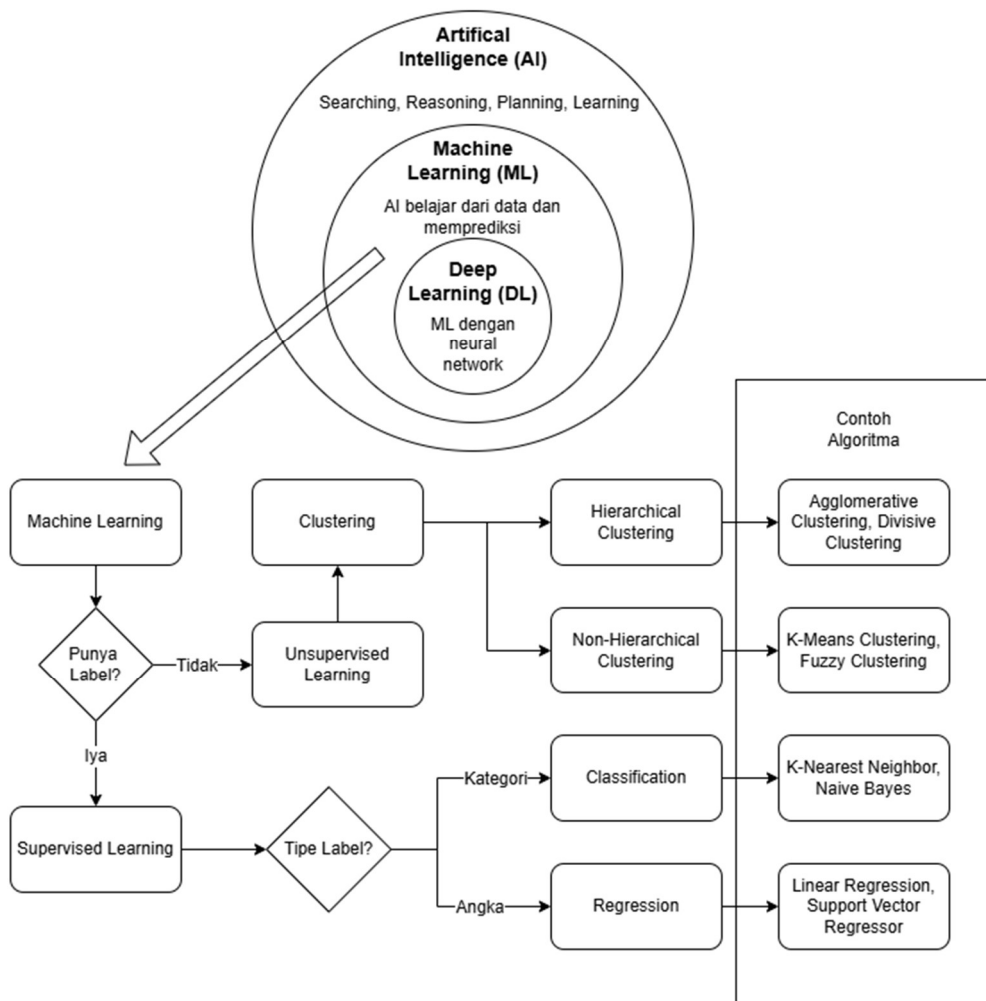
dari perspektif yang berhubungan dengan produk, mengelompokkan pelanggan dengan karakteristik yang sama, mendeskripsikan setiap kelompok, dan memilih kelompok yang akan dilayani berdasarkan karakteristik serta membuat strategi berdasarkan karakteristik tersebut. (Mothersbaugh & Hawkins, 2015).

Menghasilkan segmentasi pelanggan dengan analisis perilaku pelanggan secara konvensional memiliki banyak keterbatasan seperti banyaknya data pelanggan yang perlu dianalisis, lamanya waktu (*time consuming*) untuk melakukan analisis, dan kompleksitas tipe data yang tinggi. Keterbatasan ini terjadi karena data pelanggan di era saat ini dapat dikategorikan sebagai *big data*. Hal ini berarti data tersebut tidak hanya berasal dari sumber heterogen saja, namun juga mencakup format heterogen seperti terstruktur, tidak terstruktur, dan semi-terstruktur. *Big data* memiliki tiga karakteristik yang biasanya sering disebut dengan 3V yaitu Volume, Velocity, and Variety. Volume berarti besar atau banyaknya data digital dari jutaan perangkat atau aplikasi, Velocity berarti data yang diciptakan bersifat cepat dalam waktu yang singkat dan diperlukan pemrosesan secara instan, dan Variety berarti data berasal dari berbagai macam sumber dan format. Lebih lanjut, pendekatan *big data* ada karena pendekatan tradisional atau konvensional kurang efisien, kurang memiliki skalabilitas, akurasi yang rendah, dsb. Sifat kompleksitas *big data* ini membutuhkan teknologi dan algoritma yang canggih. Tantangan dalam pengolahan *big data* dapat ditemukan di berbagai tingkatan mulai dari pengumpulan, penyimpanan, analisis, manajemen, dan visualisasi (Oussous dkk., 2018).

Untuk menganalisis *big data* dibutuhkan peran dan kajian dari berbagai bidang ilmu (*multidisciplinary*). Bidang-bidang tersebut menggabungkan ilmu komunikasi dengan data science. Ilmu komunikasi diperlukan untuk melihat masalah dari perspektif komunikasi pemasaran dan perilaku pelanggan, sedangkan data science diperlukan untuk melihat masalah dari perspektif *big data* dan kecerdasan buatan. Tugas dari data science adalah menganalisis data dan membuat prediksi dari *big data* tersebut dengan menggunakan kecerdasan buatan. Perbedaan antara data science dan data analyst adalah data analyst biasanya hanya menganalisis data pada tahap data exploration dan data processing, lalu luaran yang dihasilkan adalah insight dan visualisasi data. Data science menganalisis data pada langkah-langkah lebih jauh lagi, termasuk menggunakan kecerdasan buatan untuk membuat model prediksi dan mengevaluasi kecerdasan buatan tersebut dengan metrik evaluasi. *Big data* digunakan untuk melatih kecerdasan buatan tersebut.

Terdapat kesalahpahaman (miskonsepsi) umum tentang kecerdasan buatan (*artificial intelligence* atau AI) yang mengatakan bahwa kecerdasan buatan hanya ada atau berkembang baru-baru ini. Kecerdasan buatan telah ada dalam kehidupan sehari-hari dan telah digunakan sebelumnya seperti mesin penerjemah (*machine translation*), visi komputer (*computer vision*), deteksi spam, sistem rekomendasi, pengolahan bahasa alami (*natural language processing*), dll. Namun, penelitian tentang GenAI (Generative Artificial Intelligence) dan algoritma Transformer benar-benar meningkatkan popularitas kecerdasan buatan. Secara umum, kecerdasan buatan dapat dibagi menjadi empat kategori yaitu Searching, Reasoning, Planning, dan Learning. Yang paling terkenal adalah Learning atau bisa juga disebut sebagai Machine Learning (pembelajaran mesin). Machine Learning

merupakan bagian dari kecerdasan buatan yang menggunakan data historis sebagai data latih (*data training*) untuk memprediksi data lain yang akan datang. Dengan demikian, Machine Learning dapat belajar berdasarkan pola dan menjadi cerdas berdasarkan data (Russel & Norvig, 2021; Suyanto, 2021).



**Gambar 2.** Kecerdasan buatan atau *Artificial Intelligence* (AI), istilah, dan hubungannya

Machine Learning dapat dibagi menjadi dua berdasarkan ada atau tidaknya label. Label atau beberapa orang juga menyebutnya *ground truth*, target, kelas, variabel terikat, atau variabel dependen, merupakan luaran atau *output* dari model Machine Learning selama proses pelatihan (*training*) untuk memprediksi hasil yang mempresentasikan nilai kebenaran atau jawaban yang benar berdasarkan data input atau fitur data atau variabel independen, atau variabel terikat. Machine Learning memiliki subkategori yaitu Supervised Learning dan Unsupervised Learning. Supervised Learning adalah yang memiliki label sedangkan Unsupervised Learning yang tidak memiliki label. Sehingga, Unsupervised Learning memprediksi data sesuai dengan label, namun dapat mengelompokkan data berdasarkan karakteristik dan kemiripan antara masing-masing nilai data. Metode atau

algoritma yang paling terkenal dinamakan Clustering (Russel & Norvig, 2021). Untuk lebih mengetahui atau mempermudah pemahaman dari yang sudah dibahas sebelumnya tentang kecerdasan buatan atau *artificial intelligence* (AI) dan juga istilah-istilah asing serta hubungan-hubungannya dapat melihat Gambar 2.

Clustering secara umum dapat dibagi ke dalam dua kategori berdasarkan penggunaan hierarki, yaitu Hierarchical Clustering dan Non-Hierarchical Clustering. Non-Hierarchical Clustering tidak menggunakan hierarki tetapi menggunakan metode lain seperti *centroid-based* dengan algoritma yang paling populernya adalah K-Means Clustering ataupun juga dengan menggunakan probabilitas dengan algoritma yang paling populernya adalah Fuzzy Clustering. Hierarchical Clustering dapat dibagi menjadi dua pendekatan. Pendekatan pertama menggunakan top-down yang disebut Divisive Clustering dan pendekatan kedua menggunakan bottom-up yang disebut Agglomerative Clustering. Hierarchical Clustering biasanya digunakan dalam bidang Ilmu Komputer dan Informatika, Arkeologi, Studi Kemanusiaan dan Sosiologi, atau bahkan Kedokteran, Farmasi, dan Biologi. Clustering dalam penerapannya dapat digunakan untuk berbagai macam contoh kasus seperti deteksi objek, deteksi zona seismik, fluktuasi suhu, matematika-politik, klasifikasi bunga iris, dan reproduksi *Escherichia coli* (Scitovski dkk., 2021).

Agglomerative Clustering digunakan dalam penelitian ini karena algoritma atau metode ini belum sering digunakan untuk kasus segmentasi pelanggan dibandingkan dengan algoritma Clustering yang lain. Lebih lanjut, Agglomerative Clustering memiliki banyak parameter yang bisa diatur, sehingga algoritma ini bisa dilakukan *Hyperparameter Tuning* untuk mendapatkan hasil Clustering yang lebih baik. Selain itu juga, metode Clustering ini sangat fleksibel untuk menentukan jumlah cluster atau grup dan tidak hanya terpaku pada angka tetap (*fixed number*), sehingga dapat disesuaikan dengan data pelanggan, keadaan di lapangan, dan banyaknya strategi komunikasi pemasaran yang dapat dilakukan.

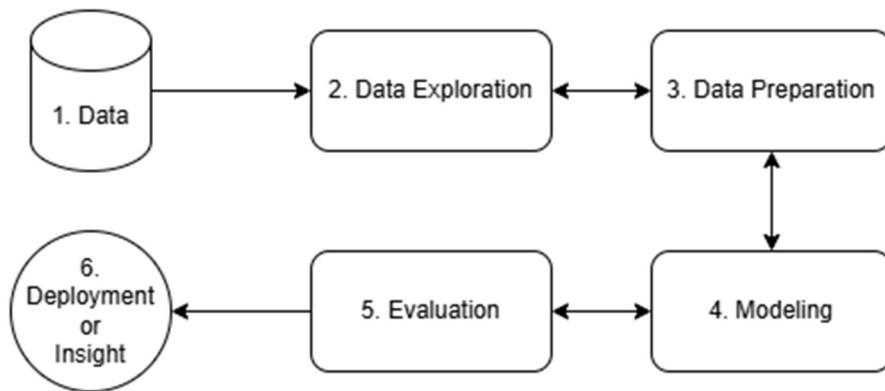
Beberapa penelitian sebelumnya telah meneliti tentang Clustering, Agglomerative Clustering, atau segmentasi pelanggan. Untuk segmentasi pelanggan, ada beberapa metode Clustering yang digunakan seperti K-Means Clustering dalam kasus E-commerce berdasarkan data perilaku pembelian pelanggan (Tabianan dkk., 2022), RFM (Recency, Frequency, Monetary) Analysis yang merupakan versi tingkatan dari K-Means Clustering dalam kasus data transaksi pada ritel online (Christy dkk., 2021), berbagai macam algoritma Machine Learning dalam kasus perusahaan asuransi jiwa (Perumalsamy dkk., 2022), dan Deep Learning yang digabungkan dengan PCA (Principal Component Analysis) dalam kasus perusahaan operator telekomunikasi (Alkhayrat dkk., 2020). Agglomerative Clustering tidak biasa digunakan dalam kasus segmentasi pelanggan, namun terdapat penelitian menggunakan metode ini di bidang Biologi untuk mengelompokkan transkriptom spasial digital dengan multilayer untuk menganalisis organ atau jaringan (Moehlin dkk., 2021).

Oleh karena itu, penelitian ini bertujuan untuk membangun kecerdasan buatan dengan menggunakan algoritma Clustering yaitu Agglomerative Clustering untuk melakukan segmentasi pelanggan, dan juga melakukan evaluasi hasilnya menggunakan metrik evaluasi

Clustering seperti Silhouette Score, Calinski-Harabasz Score, dan Davies-Bouldin Score untuk mengukur performa Agglomerative Clustering. . Metode penerapan kecerdasan buatan ini diharapkan agar dapat mempermudah kompleksitas dan mempercepat analisis segmentasi pelanggan yang dapat diterapkan pada komunikasi pemasaran agar dapat melakukan strategi pemasaran yang tepat sesuai dengan kelompok pelanggan.

## METODOLOGI

Penelitian ini menggunakan pendekatan kuantitatif dengan menggunakan metode data science workflow. Pendekatan kuantitatif dalam data science karena input data dari proses data science perlu diterjemahkan dalam bentuk angka agar algoritma kecerdasan buatan dapat berfungsi. Data science tidak memiliki workflow yang tetap atau hanya satu, memiliki tahapan dengan nama berbeda, atau juga dapat berubah tergantung dari data, tujuan, atau referensi yang digunakan (Data Science PM, 2024). Walaupun begitu, masih memiliki banyak kemiripan secara garis besar. Untuk penelitian ini, data science workflow yang digunakan dapat dilihat pada Gambar 3.

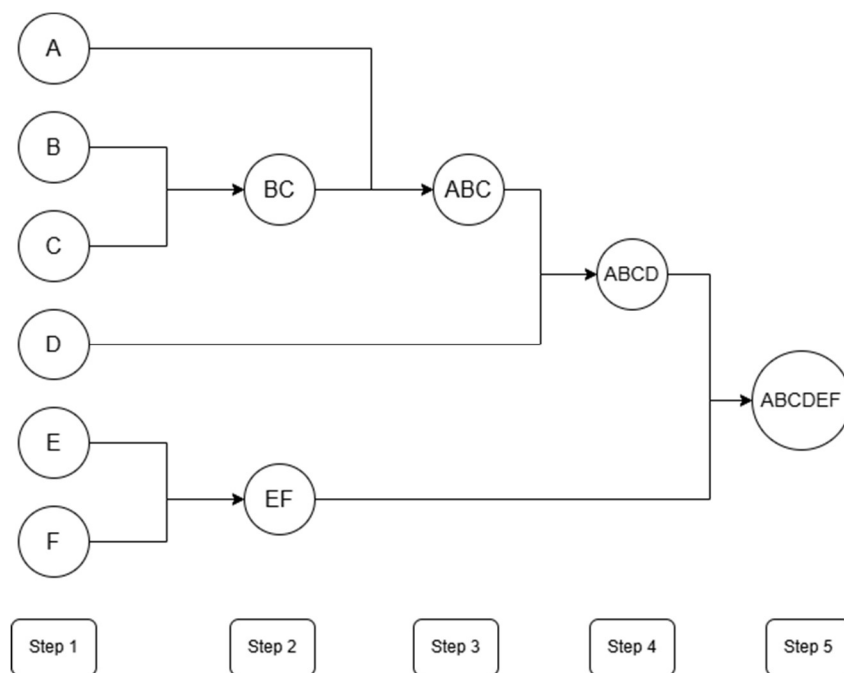


Gambar 3. Data Science workflow

Data science workflow tidak hanya terikat satu arah, namun juga bisa dua arah dan dapat dibalik atau maju-mundur tergantung dari hasil pada tahap berikutnya jika ingin melakukan uji coba lagi untuk meningkatkan performa model. Langkah pertama, **Data**, berarti mendapatkan data yang bisa berbentuk terstruktur, tidak terstruktur, atau semi-terstruktur, bisa berupa angka ataupun tidak, dan bisa berasal dari berbagai macam sumber seperti Spreadsheets (CSV or Excel), API (XML, JSON), Database (SQL), dll. Langkah kedua, **Data Exploration** atau **Data Understanding**, berarti lebih mengetahui dan mengenal data tersebut bisa dengan menggunakan atribut statistik yang sederhana (rata-rata, minimal, maksimal, simpangan baku, dll.) atau bisa juga dengan visualisasi data. Langkah ketiga, **Data Preparation** atau **Data Processing** atau **Data Preprocessing** atau **Data Cleaning** atau **Data Cleansing**, berarti memastikan kualitas data yang baik untuk proses modeling agar membuat model Machine Learning dapat bekerja dengan baik, karena jika kualitas datanya buruk maka menghasilkan luaran yang buruk juga atau disebut *Garbage In Garbage Out (GIGO)*. Langkah keempat, **Machine Learning Modeling**, berarti membangun kecerdasan buatan atau pembelajaran mesin dan juga melakukan Hyperparameter Tuning. Langkah kelima, **Evaluation**, berarti melakukan evaluasi performa model dengan menggunakan

metrik evaluasi yang sesuai dengan jenis model yang digunakan ditahap sebelumnya. Langkah keenam, **Deployment or Insight**, berarti model bisa digunakan sebagai insight (wawasan) dalam riset, bisnis, dll., dan juga dapat di-*deploy* dalam lingkup yang lebih luas seperti dalam *software* ataupun API.

Seperti yang sudah dibahas dalam Gambar 2, Agglomerative Clustering adalah bagian dari Hierarchical Clustering, yang berarti cluster dibuat dalam bentuk hierarki. Bentuk hierarki tersebut biasanya divisualisasikan dalam bentuk dendrogram. Hierarchical Clustering dibagi menjadi dua metode berdasarkan pendekatannya. Pendekatan pertama yaitu Divisive Clustering, berarti dataset pertama diasumsikan memiliki satu cluster dan kemudian melakukan pembagian atau menambah banyak cluster sampai banyak cluster sama dengan banyak data atau disebut juga dengan pendekatan top-down. Pendekatan kedua yaitu Agglomerative Clustering, berarti dataset pertama diasumsikan sebagai banyak cluster dengan jumlah yang sama dengan banyak data dan kemudian cluster-cluster berdekatan menjadi satu sampai hanya ada satu cluster atau disebut juga dengan pendekatan bottom-up (Scitovski dkk., 2021; Witten dkk., 2017). Ilustrasi tentang cara kerja Agglomerative Clustering dapat dilihat pada Gambar 4.



**Gambar 4.** Ilustrasi Agglomerative Clustering

Agglomerative Clustering memiliki banyak parameter yang bisa didefinisikan, antara lain banyak cluster ( $n$ ), metrik pengukuran jarak (distance), and linkage. Metrik distance adalah untuk mengukur jarak perbedaan antara data-data. Metrik distance biasanya contohnya menggunakan Euclidean Distance, Manhattan Distance, Cosine Distance, dll. Linkage adalah tentang cara menggabungkan dua atau lebih cluster. Metode linkage biasanya menggunakan minimum atau single-linkage, maximum atau complete-linkage, dan average-linkage (Scitovski dkk., 2021; Vijaya dkk., 2019; Witten dkk., 2017).

Untuk membangun algoritma Agglomerative Clustering dan mengukur kualitas Clustering-nya, bahasa pemrograman Python digunakan dalam penelitian ini. Alat atau *tools* ini biasanya sering digunakan untuk permasalahan di bidang data science. Environment Python yang digunakan adalah Jupyter Notebook atau bisa juga menggunakan versi *cloud* yang terkenal dari Google yang dinamakan Google Colab. Library atau package Python yang digunakan antara lain pandas, numpy, scikit-learn (sklearn), scipy, matplotlib, dan seaborn.

## HASIL DAN TEMUAN

Dataset yang digunakan berasal dari sumber data terbuka (*open source*) Kaggle dengan judul “Consumer Behavior and Shopping Habits Dataset”. Dataset ini memiliki format CSV (.csv) dengan ukuran atau dimensi 3900x18, berarti data tersebut memiliki 3900 *entries* atau baris dan 18 fitur atau kolom. Dataset yang digunakan adalah Versi 1 (Aslam & Banarjee, 2023).

### Data Exploration

Langkah pertama dari Data Exploration adalah mencari dan menemukan nilai hilang (*missing value*). Jika data memiliki *missing value*, maka akan dilakukan *missing value handling* untuk menanganinya. Metode *handling* yang biasanya digunakan antara lain seperti melakukan pembuangan (*drop*) data (bisa dari sumbu kolom maupun baris) atau melakukan imputasi (Ditjen Diktiristek, 2021). Pada dataset ini tidak memiliki nilai yang hilang, yang berarti seluruh 18 kolom masing-masing memiliki 3900 baris. Langkah berikutnya yaitu menentukan tipe data dari masing-masing kolom. Ditemukan bahwa tipe-tipe kolom ada tiga jenis yaitu float64 atau bilangan riil sebanyak 1 kolom, int64 atau bilangan bulat sebanyak 4 kolom, dan object atau string atau huruf sebanyak 13 kolom. Hasil tersebut dapat dilihat pada Gambar 5.

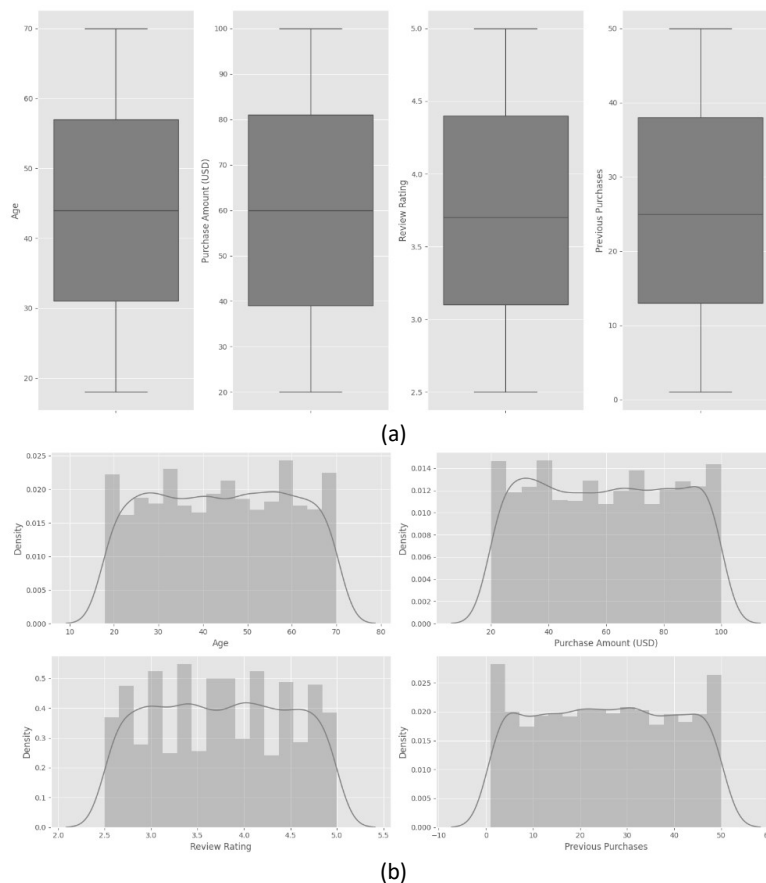
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Customer ID           3900 non-null   int64
1   Age                   3900 non-null   int64
2   Gender                3900 non-null   object
3   Item Purchased        3900 non-null   object
4   Category              3900 non-null   object
5   Purchase Amount (USD) 3900 non-null   int64
6   Location              3900 non-null   object
7   Size                  3900 non-null   object
8   Color                 3900 non-null   object
9   Season                3900 non-null   object
10  Review Rating         3900 non-null   float64
11  Subscription Status   3900 non-null   object
12  Shipping Type         3900 non-null   object
13  Discount Applied      3900 non-null   object
14  Promo Code Used       3900 non-null   object
15  Previous Purchases    3900 non-null   int64
16  Payment Method        3900 non-null   object
17  Frequency of Purchases 3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

Gambar 5. Informasi tentang dataset



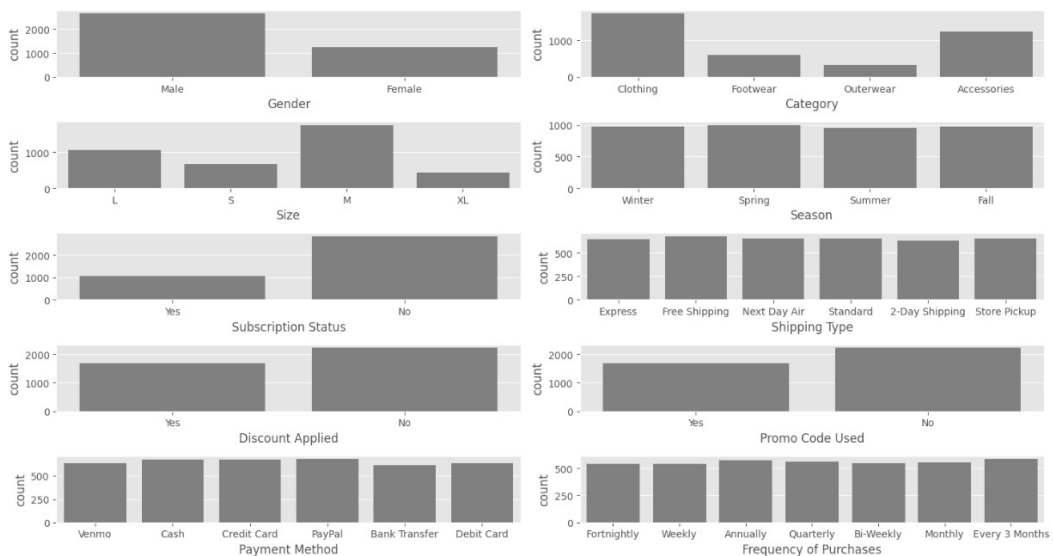
Setelah itu, kolom 'Customer ID' dibuang karena kolom tersebut tidak digunakan dalam Clustering dan juga tidak memberikan informasi berharga untuk digunakan dalam Clustering. Berdasarkan tipe data di atas, data dapat dibagi menjadi dua kategori secara umum, yaitu kolom dengan kategori angka dan non-angka. Kolom dengan kategori angka antara lain 'Age', 'Purchase Amount (USD)', 'Review Rating', dan 'Previous Purchases', sedangkan sisanya berkategori non-angka. Pemisahan ini penting, karena terdapat pendekatan yang berbeda dalam melakukan Data Exploration berdasarkan kategorinya.

Untuk kategori angka, langkah eksplorasi yang dibutuhkan adalah melakukan visualisasi data dalam bentuk *box plot* dan *distribution plot*. *Box plot* penting untuk melihat hasil visualisasi dari outlier. Outlier data yang bersifat abnormal atau anomali (bernilai terlalu tinggi ataupun terlalu rendah) dibandingkan dengan sifat data lain yang sejenis. Hal ini bisa terjadi karena terjadi kesalahan pengukuran ataupun kesalahan peng-input-an. Jika data bersifat abnormal tadi ditemukan, maka nilai data digeser ke nilai *inter-quartile range* yang sesuai. Pada dataset ini tidak terdapat outlier. *Distribution plot* diperlukan untuk melihat distribusi data dan juga kemiringan atau kecenderungan (*skewness*) dari data. Kedua visualisasi tersebut dapat dilihat pada Gambar 6.



**Gambar 6.** Visualisasi data untuk kategori angka dalam bentuk (a) *box plot* dan (b) *distribution plot*

Untuk kategori non-angka, data dibagi lagi berdasarkan banyaknya nilai unik. Pembagian pertama yaitu kolom-kolom yang memiliki nilai unik kurang dari 10 dan pembagian kedua adalah kolom-kolom dengan nilai unik sebanyak 10 atau lebih. Hal ini penting karena eksplorasi yang dilakukan yaitu visualisasi data dalam bentuk *bar plot*, dan jika jumlah nilai unik terlalu banyak maka akan sulit dilihat dalam visualisasi data. *Bar plot* untuk melihat persebaran nilai dari masing-masing non-angka dan juga banyaknya. Hasil visualisasi data *bar plot* bisa dilihat pada Gambar 7. Selain itu, untuk kategori non-angka dengan nilai unik 10 atau lebih disajikan dalam bentuk tabel yang bisa dilihat pada Tabel I.



Gambar 7. Visualisasi data *bar plot* untuk kategori non-angka dengan nilai unik di bawah 10

TABEL I	
KATEGORI NON-ANGKA DENGAN NILAI UNIK 10 ATAU LEBIH	
Nama Kolom	Banyaknya Nilai Unik
Item Purchase	25
Location	50
Color	25

### Data Preparation

Data dengan kategori non-angka perlu diterjemahkan terlebih dahulu ke dalam bentuk angka agar dapat diproses pada langkah berikutnya. Hal ini diperlukan karena dalam algoritma data science, data biasanya baru dapat diolah jika berbentuk angka. Berdasarkan hasil temuan dari kolom-kolom kategori angka maupun non-angka dapat dibagi lagi menjadi tipe butir data berdasarkan level atau skala pengukuran. Hasil pembagian tersebut dapat dilihat pada Tabel II.

TABEL II  
TIPE BUTIR DATA BERDASARKAN LEVEL PENGUKURAN

Kategori	Tipe	Banyak Kolom
Angka	Rasio	4
Non-Angka	Biner	4
	Ordinal	2
	Nominal	7

Pada kategori angka, biasanya tipe butir data adalah interval dan rasio. Tipe interval memiliki ciri-ciri seperti bisa adanya nilai negatif dan tidak memiliki titik nol mutlak. Tipe rasio memiliki ciri-ciri seperti tidak ada nilai negatif dan memiliki nilai titik nol mutlak sebagai nilai terendah. Pada kategori non-angka, tipe biner adalah tipe yang hanya memiliki dua nilai unik, tipe ordinal adalah tipe yang bisa memiliki urutan antara masing-masing nilai unik, sedangkan tipe nominal atau kategorikal tidak memiliki hubungan keturunan di antara nilai-nilai uniknya (Philippi, 2021).

Tipe biner memiliki nilai dua biasanya berupa “Ya” atau “Tidak”. Selain itu, tipe biner juga biasa dipakai pada jenis kelamin jika hanya ingin menggunakan jenis kelamin biner yaitu “Laki-laki” dan “Perempuan”. Pada dataset ini, kolom-kolom yang memiliki tipe biner antara lain adalah ‘Gender’, ‘Subscription Status’, ‘Discount Applied’, ‘Promo Code Used’. Untuk mengubah dari kategori non-angka menjadi angka maka akan dilakukan Label Encoding Data Biner. Encoding ini termasuk cukup mudah, yaitu dengan mengubah data dengan nilai “Yes” atau “Ya” sebagai 1 dan “No.” atau “Tidak” sebagai 0. Nilai 1 dalam biner merepresentasikan ada atau aktif dan nilai 0 tidak ada atau tidak aktif. Untuk jenis kelamin, tidak ada aturan khusus untuk jenis kelamin apa saja yang bernilai 1 atau 0. Dalam penelitian ini, “Female” diganti dengan nilai 1 dan “Male” diganti dengan nilai 0. Dengan demikian, keempat kolom ini sudah berhasil diterjemahkan menjadi bentuk angka.

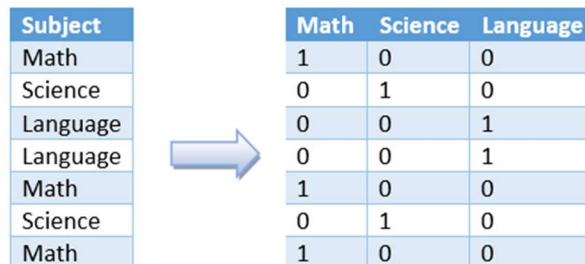
Tipe ordinal pada non-kategori biasanya diubah menjadi angka bertingkat 0,1,2, ... sesuai dengan banyaknya nilai unik. Dalam penelitian kuantitatif dengan menggunakan metode pengumpulan data survei atau kuesioner, tipe butir data ini juga bisa disebut sebagai skala Likert. Proses pengubahan dari non-kategori menjadi angka bertingkat ordinal disebut juga Ordinal Encoding. Kekurangan dari metode ini, jarak tingkat atau rank angka diasumsikan sama antara satu nilai dengan nilai yang lain. Pada dataset ini, ada dua kolom yang bisa dikategorikan sebagai ordinal yaitu ‘Size’ dan ‘Frequency of Purchases’. ‘Size’ dapat diurutkan berdasarkan ukuran kecil sampai besar, sedangkan ‘Frequency of Purchases’ dapat diurutkan dari frekuensi yang paling jarang sampai dengan yang paling sering.

Hasil Ordinal Encoding dapat diamati pada Tabel III. Hal yang perlu dilakukan pertama sebelum melakukan encoding adalah mengurutkan nilai unik sehingga akan menghasilkan hasil encoding yang sesuai. Pada ‘Size’ dilakukan pengurutan ukuran dari terkecil hingga terbesar yaitu S, M, L, XL. Pada ‘Frequency of Purchase’ diurutkan dari yang paling jarang yaitu Annually atau sekali setahun, Quarterly dan Every 3 Months memiliki nilai yang sama atau sekali 3 bulan, Monthly atau sekali sebulan, Fortnightly atau sekali dua minggu, Weekly atau sekali seminggu, dan Bi-Weekly atau dua kali seminggu.

TABEL III  
HASIL ORDINAL ENCODING

Nama Kolom	Nilai Unik Berurut	Hasil Akhir
Size	['S', 'M', 'L', 'XL']	[0, 1, 2, 3]
Frequency of Purchases	['Annually', 'Quarterly', 'Every 3 Months', 'Monthly', 'Fortnightly', 'Weekly', 'Bi-Weekly']	[0, 1, 1, 2, 3, 4, 5]

Tipe nominal atau kategorikal biasanya diterjemahkan ke bentuk angka dengan menggunakan metode One Hot Encoding. Metode ini membuat semua nilai unik menjadi kolom baru dengan nilai angka 0 atau 1 sesuai dengan nominalnya masing-masing. Untuk mempermudah memahami One Hot Encoding dapat melihat Gambar 8. Kolom-kolom dapat dilakukan One Hot Encoding antara lain 'Payment Method', 'Category', 'Season', 'Shipping Type'. Hal ini karena kolom-kolom tersebut memiliki nilai unik di bawah 10, sehingga kolom baru yang dibuat tidak bernilai banyak. Sedangkan untuk kolom-kolom pada Tabel I tidak dipakai karena memiliki nilai unik 10 atau lebih. Kekurangan dari metode One Hot Encoding ini adalah dapat mengakibatkan kenaikan ukuran dataset yang tinggi atau fenomena ini juga bisa disebut *curse of dimensionality*. Fenomena ini dapat mengakibatkan penambahan kolom atau fitur yang sebenarnya tidak terlalu penting, terlalu banyak nilai 0, data bersifat *sparse*, dan perbedaan jarak data yang tidak terlalu signifikan (Rojo-Echeburúa, 2024).



Subject	Math	Science	Language
Math	1	0	0
Science	0	1	0
Language	0	0	1
Language	0	0	1
Math	1	0	0
Science	0	1	0
Math	1	0	0

Gambar 8. Ilustrasi One Hot Encoding

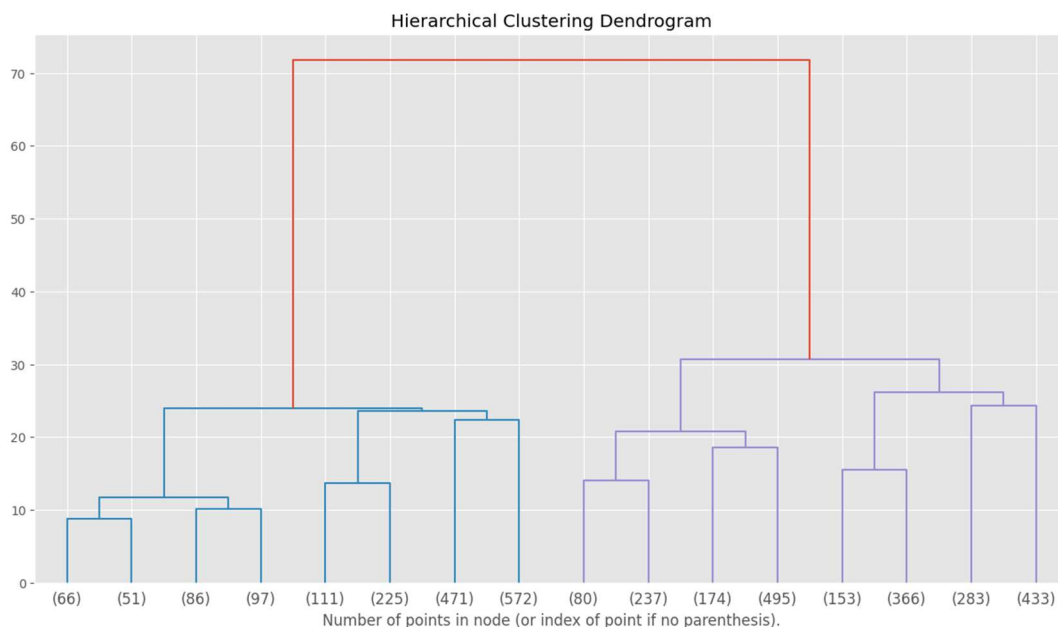
Langkah terakhir pada Data Preparation adalah melakukan Scaling. Scaling ini penting untuk Clustering ataupun juga algoritma-algoritma yang berbasis jarak (*distance*). Sebagai contoh data usia bisa memiliki nilai dari 18 sampai dengan 65 dengan rentang nilai 47. Sementara itu, pada data biner hanya memiliki nilai 0 sampai dengan 1 dengan rentang nilai 1. Perbedaan rentang nilai inilah yang membuat data dengan rentang nilai jauh jadi penentu terbesar. Untuk itu, dilakukan Scaling untuk menyamakan rentang nilai menjadi [0, 1] menggunakan Normalization Scaling atau juga bisa disebut MinMax Scaling (Ditjen Diktiristek, 2021). Rumus dari Scaling ini bisa dilihat pada Persamaan (1).

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

$x_{scaled}$  adalah hasil nilai Scaling,  $x$  adalah nilai awal,  $\min(x)$  adalah nilai terkecil dari data  $x$ , dan  $\max(x)$  adalah nilai terbesar dari data  $x$ .

### Machine Learning Modeling

Model kecerdasan buatan Agglomerative Clustering dibangun pertama-tama dengan menggunakan parameter `distance_treshold=0` dan `n_cluster=None`. Hal ini untuk membangun hierarkinya terlebih dahulu dalam bentuk dendrogram. Setelah terbentuk hierarkinya, lalu dilakukan visualisasi data berbentuk dendrogram yang selanjutnya diperlukan untuk menentukan banyaknya cluster atau grup yang diperlukan. Hasil visualisasinya dapat dilihat pada Gambar 9.



**Gambar 9.** Dendrogram dari Agglomerative Clustering

Dari gambar dendrogram tersebut, nilai atau banyaknya cluster yang baik diambil adalah 2. Hal ini karena bentuk dendrogram-nya cukup cepat untuk menyatu dengan cluster lain pada step kurang lebih di bawah 30. Cluster dengan banyak 2 cenderung lebih stabil karena masih sama dari step ke 30 sampai dengan 70 dibandingkan dari step 30 ke step 0. Oleh karena itu dipilihlah jumlah cluster (`n_cluster`) sebanyak 2 grup. Banyaknya 2 ini tidak bersifat kaku dan bisa disesuaikan juga dengan kebutuhan ataupun keinginan.

Setelah banyaknya cluster ditentukan, langkah berikutnya adalah mencari model kecerdasan buatan Agglomerative Clustering yang baik untuk melakukan prediksi dalam membuat cluster atau grup pada segmentasi pelanggan. Untuk itu, dilakukan Hyperparameter Tuning yaitu teknik dengan menyesuaikan (*tuning*) beberapa parameter yang bisa diatur pada algoritma Agglomerative Clustering. Pada kasus Clustering tidak banyak dapat dilakukan *tuning* pada umumnya dibandingkan dengan kasus Klasifikasi atau Regresi, namun model Agglomerative Clustering ini bisa dilakukan *tuning*. Tujuan dari Hyperparameter Tuning adalah untuk meningkatkan performa kecerdasan buatan agar dapat melakukan pengelompokan lebih baik lagi dibandingkan hanya dengan menggunakan pengaturan parameter *default*.

Metode paling sering dilakukan dalam Hyperparameter Tuning yaitu melakukan Grid Search. Grid Search adalah pencarian model kecerdasan buatan terbaik dengan mencoba satu demi satu seluruh kombinasi dari parameter-parameter yang diinginkan. Sebagai contoh, jika ada model kecerdasan buatan yang memiliki 3 parameter dengan masing-masing parameter memiliki nilai yang ingin dicari adalah 2,4, dan 5. Sehingga, banyaknya kemungkinan adalah  $2 \times 4 \times 5 = 40$ . Nilai 40 ini berarti model yang dibangun ada sebanyak 40 buah dan lalu untuk menemukan yang mana model terbaik bisa dilakukan pemilihan dengan mengukur performanya menggunakan metrik evaluasi kecerdasan buatan yang sesuai.

Pada penelitian ini, parameter dari model Agglomerative Clustering yang di-*tuning* adalah metric dan linkage. Metric adalah rumus jarak (*distance*) antara data-data yang digunakan dan linkage adalah aturan penyatuan cluster secara bottom-up. Masing-masing parameter memiliki banyak nilai yang ingin dicoba dicari yaitu 6 dan 3 secara berturut-turut. Sehingga, model yang dibangun adalah  $6 \times 3 = 18$  model. Untuk detail parameter dan nilai yang digunakan bisa melihat Tabel IV.

TABEL IV  
PARAMETER DAN NILAI HYPERPARAMETER TUNING

Parameter	Nilai
Metric	['euclidean', 'cosine', 'manhattan', 'minkowski', 'braycurtis', 'canberra']
Linkage	['complete', 'average', 'single']

Dari 18 model yang dibangun tersebut, untuk menentukan yang mana model dengan parameter terbaik, maka dilakukan pengukuran dengan menggunakan metrik evaluasi Clustering antara lain Silhouette Score, Calinski-Harabasz Score, dan Davies-Bouldin Score. Rentang nilai Silhouette Score adalah  $[-1,1]$  dengan semakin tinggi nilai adalah performa yang lebih baik, rentang nilai Calinski-Harabasz Score adalah  $[0,\infty)$  dengan semakin tinggi nilai adalah performa yang lebih baik, dan rentang nilai Davies-Bouldin Score adaah  $[0,\infty]$  dengan semakin rendah nilai adalah performa yang lebih baik. Ke-18 model tersebut dilakukan pe-rangking-an berdasarkan hasil dari ketiga metrik evaluasi tersebut dan model terbaik yang memiliki ranking rata-rata terkecil. Hasil dari Hyperparameter Tuning dapat dilihat pada Tabel V dengan menampilkan 5 model terbaik dari 18 model.

TABEL V  
HASIL HYPERPARAMETER TUNING

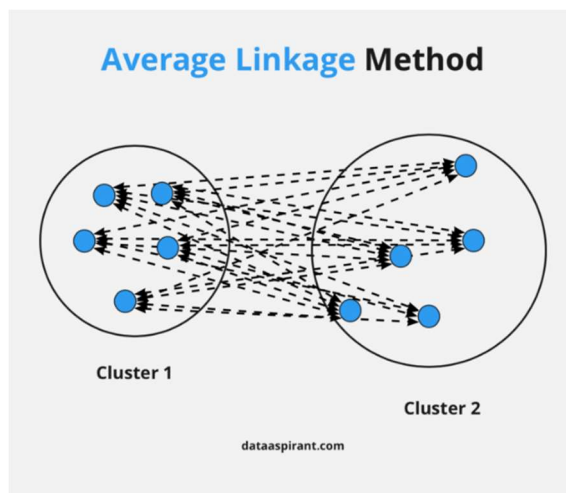
Nama Model	Metric	Linkage	Rank Silhouette	Rank Calinski	Rank Davies	Rata-rata Rank
Model 1	Manhattan	Average	1	2	6	3,00
Model 2	Braycurtis	Average	6	1	5	4,00
Model 3	Canberra	Average	3	3	7	4,33
Model 4	Cosine	Average	5	5	9	6,33
Model 5	Canbera	Complete	4	6	10	6,67

Dari hasil parameter tersebut, dapat diamati bahwa model terbaik yaitu Model 1 dengan menggunakan parameter Metric Distance yaitu Manhattan dan Linkage yaitu Average. Hal ini terbukti dengan ranking dari masing-masing pengukuran dengan rata-rata ranking yaitu 3,00. Manhattan Distance atau juga kadang disebut juga L1 Distance, Taxicab, Cityblock adalah metode pengukuran jarak antara data dengan menjumlahkan nilai mutlak dari selisih antara variabel pada data. Misalkan pada data 2 dimensi, untuk mencari Manhattan Distance dari titik  $p_1(x_1, y_1)$  ke titik  $p_2(x_2, y_2)$  maka dihitung dengan menggunakan rumus pada Persamaan (2).

$$\text{Manhattan Distance} = |x_1 - x_2| + |y_1 - y_2| \quad (2)$$

Persamaan Manhattan Distance tersebut tidak hanya berlaku untuk 2 variabel saja, tetapi juga bisa ditambah lagi sesuai dengan banyak variabel yang ada. Nama Manhattan sendiri terinspirasi dari daerah Manhattan di New York City, yaitu jika ingin pergi ke suatu tempat di Manhattan, maka tidak bisa langsung menarik satu garis lurus terpendek (cara tersebut adalah Euclidean Distance). Cara yang digunakan yaitu dengan pergi pada sumbu yang sejajar, lalu berbelok  $90^\circ$  sesuaikan arahnya sejajar dengan sumbu yang lain. Pada model ini, Manhattan Distance menjadi metrik distance terbaik karena memiliki rentang nilai yang cukup jauh, bisa bekerja baik dengan permukaan yang bersifat non-euclidean, dan bisa bekerja baik untuk data berdimensi tinggi dari hasil One Hot Encoding dari proses Data Preparation sebelumnya (Black, 2019).

Model 1 yang merupakan model terbaik memiliki Linkage yaitu Average. Selain itu, 4 model terbaik semuanya memiliki parameter Linkage Average. Metode Linkage ini memiliki kelebihan seperti bisa berfungsi baik walaupun data memiliki *noise*, permukaan data tidak harus non-linear atau non-euclidean, cocok untuk berbagai tipe data, dan lebih komprehensif karena menghitung rata-rata dari masing-masing titik data dalam cluster dibandingkan metode Linkage lain yang hanya menghitung dari satu titik saja. Ilustrasi Linkage Average dapat dilihat pada Gambar 10 (Sultana, 2020).



**Gambar 10.** Ilustrasi cara kerja Linkage Average

### Evaluation

Ketiga metrik evaluasi Clustering seperti Silhouette Score, Calinski-Harabasz Score, dan Davies-Bouldin Score selain digunakan dalam pemilihan model pada tahap sebelumnya, jika digunakan sebagai Evaluation akhir untuk menentukan seberapa bagus model kecerdasan buatan bekerja dengan baik untuk melakukan Clustering pada segmentasi pelanggan. Arti nilai yang baik dan rentang nilai dari masing-masing metrik evaluasi telah dibahas pada langkah sebelumnya.

Silhouette Score menghitung nilai  $a$  yaitu rata-rata jarak titik data dengan titik data yang lain pada cluster yang sama dan nilai  $b$  yaitu rata-rata jarak terpendek dari satu titik data ke titik data lain pada cluster yang berbeda. Persamaan Silhouette Score dapat dilihat pada Persamaan (3).

$$S = \frac{b - a}{\max(a, b)} \quad (3)$$

Calinski-Harabasz Score disebut juga Variance Ratio Criterion adalah menghitung rasio (perbandingan) jumlah dari  $B$  yaitu dispersi antara cluster berbeda (*between*) dengan  $W$  yaitu dispersi antara cluster yang sama (*within*), selain itu juga memerlukan  $k$  yaitu banyaknya cluster dan  $n$  yaitu banyaknya data. Persamaan Calinski-Harabasz dapat dilihat pada Persamaan (4).

$$CH = \frac{B}{W} \times \frac{n - k}{k - 1} \quad (4)$$

Davis-Bouldin Score menghitung rata-rata kemiripan suatu cluster dengan cluster lain yang paling mirip dengan cluster tersebut atau bisa dilambangkan  $R$  pada cluster sebanyak  $k$ . Nilai  $R$  ini adalah perbandingan *within*-cluster dengan *between*-cluster. Sehingga, hasilnya cluster yang terpisah jauh dan memiliki dispersi yang sedikit adalah cluster yang baik. Persamaan Davies-Bouldin dapat dilihat pada Persamaan (5).

$$DB = \frac{1}{k} \times \sum_{i=1}^k R_i \quad (5)$$

Dengan ketiga metrik evaluasi tersebut, maka didapatkanlah hasil performa Agglomerative Clustering yang dapat dilihat pada Tabel VI.

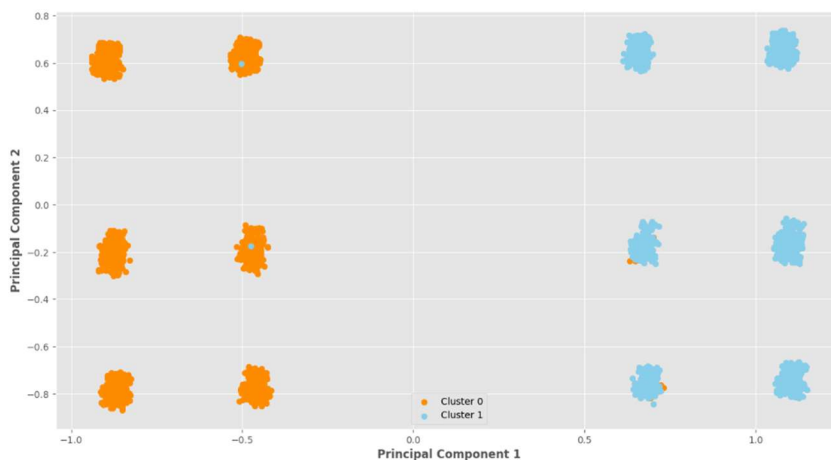
TABEL VI  
HASIL EVALUATION

Metrik Evaluasi	Nilai
Silhouette Score	0,234
Calinski-Harabasz Score	664,389
Davies-Bouldin Score	2,394



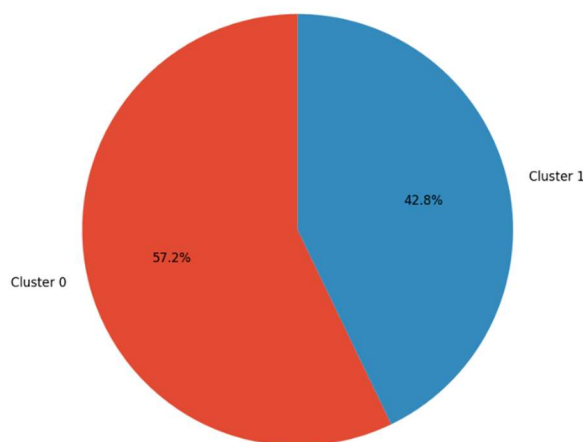
### *Deployment or Insight*

Agglomerative Clustering akhirnya bisa melakukan prediksi dengan mengelompokkan pelanggan atau segmentasi pelanggan ke dalam 2 cluster yang dinamakan Cluster 0 dan Cluster 1. Hasil dari segmentasi ini bisa digunakan langsung untuk melakukan prediksi data pelanggan yang akan ada di masa mendatang. Hasil cluster kecerdasan buatan ini dari dataset yang ada dapat dilihat pada Gambar 11. Perlu diingat, data yang digunakan untuk pelatihan memiliki dimensi atau kolom atau fitur yang cukup banyak sehingga dilakukan reduksi dimensi menggunakan PCA (Principal Component Analysis) untuk mempermudah visualisasi.



**Gambar 11.** Visualisasi hasil cluster menggunakan PCA

Dari hasil tersebut, dapat diamati bahwa Agglomerative Clustering sudah dapat melakukan segmentasi pelanggan dengan baik dengan membagi cluster menjadi 2 yang cukup jelas perbedaannya, walaupun ada sedikit cluster yang tidak sesuai dengan penempatannya, namun secara umum kecerdasan buatan ini sudah dapat berfungsi dengan baik. Lebih lanjut, banyaknya anggota masing-masing cluster dapat diamati pada Gambar 12.



**Gambar 12.** Banyaknya pelanggan pada masing-masing segmen cluster

Untuk ide deployment agar bisa digunakan dalam bisnis, bisa dari cara yang sederhana yaitu melakukan segmentasi pelanggan dengan data pelanggan baru secara berkala dari data Spreadsheet dan melakukan analisis strategi komunikasi pemasaran, melakukan segmentasi secara instan dan *real-time* dengan melakukan integrasi program Agglomerative Clustering pada *software* yang juga menggunakan bahasa pemrograman Python dengan menyambungkannya kepada database *software* (SQL), melakukan segmentasi *cross-platform* tanpa ketergantungan bahasa pemrograman Python dengan menggunakan API, ataupun bisa dengan membangun *dashboard* visualisasi data seperti dengan PCA. Namun, perlu diingat juga untuk melakukan Data Preparation yang sama sebelum melakukan segmentasi seperti Encoding dan Scaling.

#### *Perspektif Ilmu Komunikasi*

Seperti yang sudah disinggung di Gambar 1, setelah melakukan analisis pasar dan segmentasi pelanggan atau segmentasi pasar, langkah berikutnya adalah memilih strategi pemasaran yang sesuai berdasarkan segmen atau cluster yang ada. Salah satu strategi pemasaran yaitu *marketing mix* dengan memperhatikan produk, harga, komunikasi, distribusi, dan pelayanan kepada target pasar. Fokus pada bidang Ilmu Komunikasi adalah pada komunikasinya makanya disebut komunikasi pemasaran.

Komunikasi pemasaran termasuk membahas periklanan, tenaga penjualan, hubungan masyarakat, pengemasan, dan bentuk komunikasi lain yang diberikan perusahaan tentang citra dan produknya. Lebih lanjut, untuk menerapkan strategi komunikasi pemasaran yang efektif, harus bisa menjawab beberapa pertanyaan seperti *kepada siapa kita berkomunikasi?, apa dampak dari komunikasi kita kepada target pelanggan?, pesan apa yang bisa berefek yang diinginkan oleh pelanggan?, apa cara dan media yang digunakan untuk menjangkau pelanggan?, kapan sebaiknya berkomunikasi dengan target pelanggan?* (Mothersbaugh & Hawkins, 2015).

Selain itu, bisa juga memakai strategi pemasaran Segmentation-Targeting-Positioning (STC). Setelah selesai pada tahap segmentasi, maka selanjutnya tahap *targeting*. Strategi pemasaran *targeting* bisa menggunakan tiga pendekatan, yaitu strategi pemasaran tanpa perbedaan yaitu hanya dengan menggunakan satu strategi, strategi pemasaran dengan perbedaan yaitu menggunakan strategi berbeda sesuai dengan semua segmen, atau strategi pemasaran terkonsentrasi yaitu hanya fokus kepada satu atau beberapa segmen saja. Strategi ini disesuaikan dengan sumber daya perusahaan, jenis produk atau jasa, tingkat keberagaman pasar, dan strategi kompetitor. Pada era sekarang, segmentasi tidak hanya bersifat kaku namun juga bisa memiliki strategi pemasaran *targeting* yang disesuaikan dengan pribadi (*personalized*) melalui sistem rekomendasi melalui media internet ataupun mobile.

*Positioning* atau *product positioning* merupakan tahap akhir dalam strategi STC. *Positioning* membahas tentang perusahaan yang ingin memosisikan tentang bagaimana mereka ingin produk atau layanan mereka dipersepsikan oleh pelanggan, dibandingkan dengan pesaing

atau kompetitor. Komunikasi pemasaran berfokus pada pembentukan persepsi pelanggan dengan menyoroti manfaat, membedakan dari pesaing, memastikan perusahaan memiliki sumber daya dan kredibilitas yang diperlukan untuk memenuhi janji-janjinya terhadap pelanggan, dan memastikan posisi tersebut dapat dipertahankan terhadap tindakan kompetitor. *Positioning* yang efektif dapat menekankan berbagai faktor yang menjadi kelebihan, seperti layanan yang unggul atau biaya rendah, dan harus selaras dengan kebutuhan dan harapan pelanggan (Camilleri, 2018).

Perlu diketahui juga dengan melakukan segmentasi pelanggan atau segmentasi pelanggan baik menggunakan kecerdasan buatan ataupun dengan metode lain dapat meningkatkan penjualan dan keuntungan. Salah satu contoh kasusnya adalah segmentasi pasar pada toko penjualan *video game* di Chengdu, China dapat meningkatkan keuntungan sebanyak 6,95% (Sun dkk., 2019). Contoh kasus lain adalah segmentasi pelanggan pada retail ukuran menengah (UMKM) di Kuwait dengan menggunakan teknologi big data seperti Hadoop, data mining, dan algoritma Clustering yang dapat meningkatkan penjualan berkisar dari 5% sampai dengan 9% (Yoseph dkk., 2020). Penerapan kecerdasan buatan dan *machine learning* (pembelajaran mesin) secara umum juga dapat meningkatkan pendapatan dan mengurangi biaya pada perusahaan berbasis teknologi. Kecerdasan buatan yang paling banyak digunakan adalah GenAI yang termasuk dalam kategori *Natural Language Processing* (NLP) dalam bentuk chatbot. Penerapan kecerdasan buatan tersebut memiliki pengaruh terhadap total penjualan (Khalimonchuk & Pozovna, 2024).

## KESIMPULAN

Telah berhasil diterapkan kecerdasan buatan di bidang ilmu komunikasi, yaitu segmentasi pelanggan dengan menggunakan algoritma Agglomerative Clustering. Pendekatan yang digunakan adalah *multidisciplinary* dengan menggabungkan Data Science dengan Ilmu Komunikasi. Data Science workflow digunakan untuk membangun model kecerdasan buatan ini mulai dari data, Data Exploration, Data Preparation, Modeling, Evaluation, dan Deployment-Insight.

Model kecerdasan buatan yang digunakan adalah Agglomerative Clustering dengan parameter terbaik adalah banyak cluster sebanyak 2 grup, metrik jarak (*distance metric*) yang digunakan adalah Manhattan Distance, dan Linkage yang digunakan adalah Average. Parameter terbaik ini didapatkan dengan Hyperparameter Tuning melalui penilaian dari tiga metrik evaluasi Clustering yaitu Silhouette Score, Calinski-Harabasz Score, dan Davies-Bouldin Score. Selain itu, telah berhasil juga dihitung evaluasi hasil performa Clustering akhir dengan ketiga metrik tersebut dengan nilai Silhouette Score sebesar 0,234, Calinski-Harabasz Score sebesar 664,389 dan Davies-Bouldin Score sebesar 2,394.

Diharapkan dengan adanya hasil dan temuan ini, agar dapat dilanjutkan pengembangan penerapan kecerdasan buatan tersebut sampai ke tahap *deployment* atau *insight* untuk membantu menyusun strategi komunikasi pemasaran yang efektif setelah terbentuknya segmentasi pelanggan. Strategi yang bisa diterapkan antara lain bisa dengan *marketing mix* atau Segmentation-Targeting-Positioning (STP).

Batasan dari penelitian ini adalah hanya menggunakan data yang bersifat sumber terbuka (*open source*) dan menggunakan algoritma Agglomerative Clustering. Saran untuk penelitian selanjutnya adalah bisa menggunakan data dari perusahaan yang ingin dimaksimalkan strategi pemasarannya atau juga bisa mendapatkan data dari berbagai macam sumber dan format data. Saran berikutnya yaitu bisa dengan menggunakan algoritma atau metode Clustering yang berbeda seperti DBSCAN, K-Modes Clustering, ataupun sejenisnya.

## DAFTAR PUSTAKA

- Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, 7(1). <https://doi.org/10.1186/s40537-020-0286-0>
- Aslam, S., & Banarjee, S. (2023). *Consumer Behavior and Shopping Habits Dataset* (Nomor 1). Kaggle. <https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/data>
- Black, P. E. (2019, Februari 11). *Manhattan Distance*. Dictionary of Algorithms and Data Structures. <https://xlinux.nist.gov/dads/HTML/manhattanDistance.html>
- Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. Dalam *Tourism, Hospitality and Event Management* (hlm. 69–83). Springer Nature. [https://doi.org/10.1007/978-3-319-49849-2\\_4](https://doi.org/10.1007/978-3-319-49849-2_4)
- Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM Ranking – An Effective Approach to Customer Segmentation. *Journal of King Saud University - Computer and Information Sciences*, 33(10), 1251–1257. <https://doi.org/10.1016/j.jksuci.2018.09.004>
- Data Science PM. (2024, Desember 22). *What is a Data Science Workflow?* <https://www.datascience-pm.com/data-science-workflow/>
- Ditjen Diktiristek. (2021). Data Preparation 3: Mengkonstruksi Data. Dalam *Microcredential Associate Data Scientist*. Studi Independen Kampus Merdeka.
- Khalimonchuk, I., & Pozovna, I. (2024). The Role of Machine Learning and Artificial Intelligence in Optimizing Costs and Increasing Revenues of Technological Companies. *Economic Sustainability and Business Practices*, 1(1). <https://doi.org/10.21272/1817-9215.2024.3-04>
- Kovanoviene, V., Romeika, G., & Baumung, W. (2021). Creating Value for the Consumer Through Marketing Communication Tools. *Journal of Competitiveness*, 13(1), 59–75. <https://doi.org/10.7441/joc.2021.01.04>
- Moehlin, J., Mollet, B., Colombo, B. M., & Mendoza-Parra, M. A. (2021). Inferring biologically relevant molecular tissue substructures by agglomerative clustering of

- digitized spatial transcriptomes with multilayer. *Cell Systems*, 12(7), 694–705.e3. <https://doi.org/10.1016/j.cels.2021.04.008>
- Mothersbaugh, D. L., & Hawkins, D. I. (2015). *Consumer Behavior: Building Marketing Strategy* (13 ed.). McGraw-Hill Education.
- Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4), 431–448. <https://doi.org/10.1016/J.JKSUCI.2017.06.001>
- Perumalsamy, J., Krothapalli, B., & Althati, C. (2022). Machine Learning Algorithms for Customer Segmentation and Personalized Marketing in Life Insurance: A Comprehensive Analysis. *Journal of Artificial Intelligence Research By The Science Brigade (Publishing) Group 83 Journal of Artificial Intelligence Research*, 2(2), 83.
- Philippi, C. L. (2021). On measurement scales: Neither ordinal nor interval? *Philosophy of Science*, 88(5), 929–939. <https://doi.org/10.1086/714873>
- Rojo-Echeburúa, A. (2024, Juni 26). *What Is One Hot Encoding and How to Implement It in Python*. Datacamp. <https://www.datacamp.com/tutorial/one-hot-encoding-python-tutorial>
- Russel, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4 ed.). Pearson Education.
- Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021). *Cluster Analysis and Applications*. Springer. <https://doi.org/10.1007/978-3-030-74552-3>
- Shankar, V., Grewal, D., Sunder, S., Fossen, B., Peters, K., & Agarwal, A. (2022). Digital Marketing Communication in Global Marketplaces: A Review of Extant Research, Future Directions, and Potential Approaches. *International Journal of Research in Marketing*, 39(2), 541–565. <https://doi.org/10.1016/J.IJRESMAR.2021.09.005>
- Sultana, S. I. (2020, Desember 21). *How the Hierarchical Clustering Algorithm Works*. Dataaspirant. <https://dataaspirant.com/hierarchical-clustering-algorithm/>
- Sun, M., Tian, Y., Yan, Y., & Liao, Y. (2019). Improving the Profit by Using a Mixed After-sales Service as a Market Segmentation. *Nankai Business Review International*, 10(2), 233–258. <https://doi.org/10.1108/NBRI-10-2017-0057>
- Suyanto. (2021). *Artificial Intelligence: Searching, Reasoning, Planning, and Learning* (3 ed.). Informatika.
- Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability (Switzerland)*, 14(12). <https://doi.org/10.3390/su14127243>

- Vijaya, Sharma, S., & Batra, N. (2019). Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 568–573. <https://doi.org/10.1109/COMITCon.2019.8862232>
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4 ed.). Elsevier. <https://doi.org/10.1016/C2015-0-02071-8>
- Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The Impact of Big Data Market Segmentation Using Data mining and Clustering Techniques. *Journal of Intelligent and Fuzzy Systems*, 38(5), 6159–6173. <https://doi.org/10.3233/JIFS-179698>