Jurnal Komputer Multidisipliner

Vol.7 No.10, Oktober 2024
ISSN: 24559633

ANALISIS PENGELOMPOKAN KEBIASAAN PENGGUNAAN MEDIA SOSIAL MENGGUNAKAN METODE *HIERARCHICAL CLUSTERING* : PENDEKATAN *DIVISIVE*

Ronasip Heppy Ria Sibarani[1], Dewi Lestari[2], Deswita Ananda[3], Isaac Paul Hamonangan Marbun[4], Arnita[5]

Universitas Negeri Medan

E-mail: ronasibsibarani@gmail.com[1], dl6011667@gmail.com[2], deswitaananda0322@gmail.com[3], isaacpaulmarbun@mhs.unimed.ac.id[4], arnita@unimed.ac.id[5]

Abstrak

Menganalisis pola penggunaan media sosial di kalangan pengguna muda dengan fokus pada platform yang digunakan, durasi penggunaan harian, dan tingkat keterlibatan pengguna. Penelitian ini bertujuan untuk mengidentifikasi segmen-segmen pengguna yang memiliki perilaku penggunaan media sosial yang berbeda dengan menggunakan metode Divisive Hierarchical Clustering. Wawasan yang diperoleh akan digunakan untuk menyusun strategi pemasaran yang lebih personal dan efektif, dengan menargetkan platform dan pola penggunaan spesifik yang relevan bagi setiap segmen pengguna. Metode yang digunakan dalam penelitian ini yaitu metode *Hierarchical Clustering*, khususnya *Divisive Hierarchical Clustering*, digunakan untuk mengelompokkan kebiasaan pengguna. Sebelum analisis dilakukan, data terlebih dahulu diproses melalui tahapan pembersihan, penskalaan, dan *encoding*. Selanjutnya, *Principal Component Analysis* (PCA) digunakan untuk mereduksi dimensi data dan memvisualisasikan hasil clustering, sementara kualitas hasil clustering dinilai menggunakan *Silhouette Score*. Hasil dari analisis ini diharapkan dapat memberikan panduan untuk pengembangan strategi pemasaran yang lebih personal dan efektif. Penelitian ini menggunakan data yang diperoleh dari survei online yang melibatkan 29 responden. Data yang dikumpulkan meliputi informasi demografi, platform media sosial yang digunakan, jumlah waktu yang dihabiskan, dan pendapatan bulanan. Hasil penelitian menunjukkan bahwa pengelompokan dua cluster memberikan hasil yang paling optimal, dengan nilai *Silhouette Score* tertinggi sebesar 0.08. Pengelompokan ini berhasil mengidentifikasi dua kelompok pengguna yang memiliki kebiasaan penggunaan yang berbeda, terutama dalam hal waktu yang dihabiskan dan tingkat keterlibatan pada media sosial. Selain itu, platform media sosial yang paling sering digunakan adalah Instagram, YouTube, Twitter, dan TikTok.Penelitian ini memberikan kontribusi baru dalam pemahaman pola penggunaan media sosial di kalangan pengguna muda dengan menggunakan metode *Divisive Hierarchical Clustering*. Temuan ini dapat dimanfaatkan oleh perusahaan untuk menyusun strategi pemasaran yang lebih personal, serta meningkatkan pengalaman pengguna melalui personalisasi layanan yang lebih tepat sasaran. Belum ada penelitian yang secara spesifik melakukan Analisis Pengelompokan Kebiasaan Penggunaan Media Sosial Menggunakan Metode Hierarchical Clustering: Pendekatan Divisive. Kebanyakan studi sebelumnya menggunakan metode Pendekatan agglomerative lebih sering digunakan dibandingkan pendekatan divisive karena cenderung lebih mudah diimplementasikan dan dimengerti.

Kata Kunci — Hierarchical Clustering, Divisive Clustering, Media Sosial, Analisis Pengelompokan, Pola Penggunaan, Principal Component Analysis (Pca), Silhouette Score, Personalization, Strategi Pemasaran.

1. PENDAHULUAN

Media sosial sudah menjadi hal yang cukup penting di era sekarang. Kebutuhan tentang sebuah informasi yang cepat dan tepat untuk kepentingan maupun kepuasan pribadi yang terkadang melupakan satu hal yaitu durasi yang mereka gunakan untuk mengakses media sosial tersebut. Kehadiran media sosial tentu membawa dampak positif maupun dampak negatif. Dampak positif yang diperoleh salah satunya adalah Analisis klaster merupakan salah satu teknik analisis multivariat yang digunakan untuk mengelompokkan objek berdasarkan kriteria tertentu sehingga diperoleh grup-grup yang lebih kecil dimana dalam satu grup memiliki objek yang memiliki kemiripan satu sama lain, tetapi grup satu pertukaran informasi yang cepat dan tepat. Sedangkan dampak negatif yang diperoleh adalah tentang durasi penggunaan media sosial yang berlebihan. Dampak negatif yang dimaksud yaitu gangguan tidur, tingkat konsentrasi menurun, merusak mata, gangguan kecemasan dan kecanduan media sosial [1].

Psikoterapis dari California School of Professional Psychology, Philip Cushman, menyarankan agar penggunaan media sosial dibatasi antara setengah jam hingga maksimal dua jam per hari. Rekomendasi ini sejalan dengan studi dari University of Pittsburgh, yang menunjukkan bahwa orang yang menghabiskan terlalu banyak waktu di media sosial setiap hari memiliki risiko depresi hingga tiga kali lipat dibandingkan dengan mereka yang jarang menggunakan media sosial. Berdasarkan temuan tersebut, penggunaan media sosial yang berlebihan dapat berdampak buruk pada kesehatan mental, dengan peningkatan risiko depresi bagi pengguna aktif [2].

Dengan pertumbuhan pesat pengguna media sosial, analisis kebiasaan pengguna menjadi penting untuk memahami pola interaksi yang terjadi. Dalam konteks analisis kebiasaan penggunaan media sosial, clustering berfungsi untuk mengidentifikasi pola-pola perilaku pengguna yang serupa, misalnya berdasarkan frekuensi penggunaan, jenis konten yang diakses, atau preferensi interaksi sosial. Salah satu metode yang efektif dalam mengelompokkan pola penggunaan adalah Hierarchical Clustering.

Ada dua metode dalam analisis cluster, yaitu metode hierarki dan non-hierarki. Metode hierarki adalah metode yang dirancang untuk dekomposisi berhierarki (tingkatan) dari kumpulan data menurut karakteristik pada data tersebut. Sedangkan metode non-hierarki adalah metode yang digunakan untuk mengelompokkan data dan jumlah cluster yang akan dibuat dapat ditetapkan sebelumnya [1]. Hasil cluster dikatakan baik ketika mempunyai homogenitas yang besar antar objek dalam satu cluster dan heterogenitas yang besar pula antar cluster yang satu dengan cluster lainnya [4].

Hierarchical clustering merupakan teknik pengelompokan data yang membentuk hirarki, di mana objek data secara bertahap dikelompokkan ke dalam cluster berdasarkan jarak atau kesamaan antar objek.[5]. Penelitian ini menggunakan Analisis Divisive Hierarchical, yang merupakan metode pengelompokan hierarki yang memecah data dengan strategi top-down. Ia dimulai dengan menempatkan semua objek dalam satu kelompok besar, yang berfungsi sebagai akar hierarki. Kemudian, ia membagi kelompok akar menjadi kelompok sub yang lebih kecil, dan proses ini diulangi, memecah kelompok menjadi yang lebih kecil. Proses pemartisian berlanjut hingga setiap kelompok pada tingkat terendah hanya berisi satu objek, atau objek dalam sebuah kelompok cukup mirip satu sama lain, membentuk yang bisa disebut sebagai kelompok-kelompok yang lebih kecil [6].

Divisive hierarchical clustering memiliki kelebihan dan kekurangan yang perlu dipertimbangkan dalam penggunaannya. Analisis Divisive memiliki beberapa kelebihan, di antaranya adalah kemampuannya memulai dari keseluruhan data, sehingga lebih efisien dalam menangani dataset besar. Metode ini juga fleksibel dalam memisahkan data

berdasarkan jarak antar elemen, dan sangat baik untuk mendeteksi cluster kecil yang tersembunyi. Namun, juga memiliki beberapa kekurangan yaitu prosesnya membutuhkan komputasi yang intensif, terutama untuk dataset besar, serta penentuan pemisahan awal yang optimal seringkali sulit dilakukan. Selain itu, metode ini kurang baik dalam mempertahankan struktur hierarki keseluruhan, dan kurang cocok digunakan jika jumlah cluster tidak diketahui [7].

Penelitian ini bertujuan untuk menganalisis dan mengelompokkan kebiasaan pengguna media sosial menggunakan metode Hierarchical Clustering, yang diharapkan dapat memberikan wawasan mendalam tentang pola-pola penggunaan yang berbeda di kalangan pengguna.

Penelitian ini juga memanfaatkan Principal Component Analysis (PCA) untuk mereduksi dimensi data sehingga memudahkan visualisasi dan interpretasi hasil clustering. Metode Principal Component Analysis (PCA) dapat digunakan untuk menyelesaikan proses-proses citra digital sebagai salah satu bentuk data. PCA mampu menyederhanakan suatu data dengan cara mentransformasi linier sehingga terbentuk sistem koordinat baru dengan varian maksimum [8]. Selain itu, kualitas clustering akan dievaluasi menggunakan Silhouette Score, yang memberikan gambaran seberapa baik data terkelompok dalam cluster tertentu. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam mengembangkan pendekatan yang lebih efektif dalam memahami kebiasaan penggunaan media sosial dan menerapkan strategi personalisasi yang lebih sesuai dengan kebutuhan pengguna.

2. METODE PENELITIAN

Metode yang digunakan meliputi pengumpulan data melalui survei, preprocessing data, dan analisis clustering menggunakan algoritma tertentu. Langkah-langkah rinci yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Sumber Data

Data yang digunakan dalam penelitian ini berasal dari survei online yang diambil dari Kaggle pada tahun 2022, yang terdiri dari 19 variabel dan 29 data. Data ini diisi oleh 29 responden, dengan masing-masing responden memberikan informasi terkait demografi, yang mencakup ID, waktu mulai (Start time), waktu selesai (Completion time), alamat email (Email), nama (Name), rentang usia (Age_Range), jenis kelamin (Sex), bidang studi atau pekerjaan (Field), tingkat pendidikan (Education), status kerja (Employability), serta pendapatan bulanan (Income). Selain itu, informasi mengenai penggunaan media sosial juga disertakan, seperti platform yang digunakan (Daily_used), jumlah akun media sosial yang dimiliki (Accounts), jumlah jam yang dihabiskan per hari (Hrs), waktu penggunaan (Time), jumlah pengikut (Followers), jumlah postingan yang telah diunggah (Posted), frekuensi menunda posting (Posted_later), serta keyakinan terkait dampak media sosial terhadap kehidupan mereka (Believed).

Survei ini bertujuan untuk memberikan gambaran umum mengenai kebiasaan penggunaan media sosial di kalangan muda. Para responden dipilih secara acak, mencakup kelompok usia yang mayoritas masih dalam masa pendidikan atau awal karier. Hasil survei menunjukkan bahwa kelompok usia muda cenderung menghabiskan banyak waktu di media sosial, dengan dominasi penggunaan platform populer seperti Instagram, Twitter, dan TikTok. Namun, meskipun mereka sering terlibat dalam aktivitas konsumsi konten, hanya sebagian kecil yang secara aktif memproduksi atau memposting konten mereka sendiri, yang tercermin dari variabel Posted dan Posted_later.

Kondisi ekonomi responden juga bervariasi, namun sebagian besar di antaranya belum memiliki penghasilan yang stabil, yang sejalan dengan status mereka sebagai mahasiswa atau individu yang baru memasuki dunia kerja. Penelitian ini mengungkap pola yang konsisten terkait hubungan antara demografi, pendidikan, dan perilaku penggunaan media sosial, di mana individu yang lebih muda cenderung lebih sering berinteraksi

dengan media sosial dibandingkan kelompok usia yang lebih tua.

2. Preprocessing Data

Sebelum dilakukan analisis lebih lanjut, data yang terkumpul melalui survei harus diproses untuk memastikan kelayakannya dalam analisis statistik. Praproses merupakan sebuah tahap awal yang harus dilakukan pada data mining. Tujuan praproses dalam data mining adalah untuk mempersiapkan data mentah sebelum dilakukan proses lain. Praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Praproses juga dilakukan untuk mendapatkan hasil yang lebih akurat, pengurangan waktu perhitungan untuk large scale problem, dan membuat nilai data menjadi lebih kecil tanpa merubah informasi yang didalamnya. Praproses data dapat berupa data cleaning, data integration, data reduction, dan data transformation [9]. Tahapan preprocessing data meliputi:

a. Pembersihan Data: Beberapa kolom yang tidak relevan dengan tujuan penelitian, seperti ID, Nama, Email, Start Time, dan Completion Time, dihapus. Kolom-kolom ini tidak memiliki kontribusi langsung terhadap analisis kebiasaan penggunaan media sosial. Hal ini dilakukan untuk memfokuskan analisis pada variabel-variabel yang lebih bermakna dan menghindari noise yang dapat mengganggu hasil analisis.
b. Penanganan Data Hilang: Sebelum penskalaan dan analisis, data diperiksa untuk memastikan tidak ada nilai yang hilang (missing values). Baris-baris yang memiliki nilai yang hilang (missing values) dihapus dari dataset untuk menghindari bias atau error dalam proses analisis. terdapat beberapa kolom yang memiliki data hilang (missing values). Penanganan missing values sangat penting untuk memastikan analisis tidak terganggu oleh data yang tidak lengkap. Dalam dataset ini, kolom Income ditemukan memiliki missing values, sedangkan kolom lain seperti Age_Range, Sex, dan Field tidak memiliki missing values.
c. Pengecekan Outlier: Sebelum analisis dilakukan, data diperiksa untuk mengidentifikasi outliers. Outlier adalah data yang memiliki nilai ekstrim atau jauh berbeda dari sebagian besar data lainnya. Outlier dapat mempengaruhi hasil analisis, terutama pada algoritma berbasis jarak seperti clustering. Pengecekan outlier dilakukan menggunakan metode seperti Z-Score atau IQR (Interquartile Range) [10]. Pada dataset yang akan diteliti dalam penelitian ini, tidak ditemukan adanya outlier.
d. Penskalaan Data: Penskalaan data adalah proses mengubah nilai variabel numerik dalam dataset agar berada dalam rentang yang sama atau dalam skala tertentu. Langkah ini bertujuan untuk menormalkan data dan memastikan bahwa nilai-nilai yang berbeda berada dalam skala yang seragam[11]. Data outlier merupakan suatu kejadian yang wajar terjadi dan sering kali muncul dalam analisis data, termasuk data time series. Pengaruh dari adanya data outlier bisa menyebabkan bias atau salah prediksi pada model data time series tersebut [12]. Outlier penting untuk dilakukan sebelum melakukan analisis menggunakan algoritma berbasis jarak, seperti clustering dimana variabel dengan skala besar bisa mendominasi hasil analisis. Data numerik seperti pendapatan, jumlah jam penggunaan media sosial per hari, dan jumlah pengikut di media sosial di-skala menggunakan metode Standardization. Metode ini digunakan untuk memastikan bahwa semua variabel memiliki skala yang seragam dan tidak ada variabel yang mendominasi analisis karena perbedaan rentang nilai.
e. Encoding Variabel Kategorikal: Encoding adalah proses mengubah data dari suatu bentuk atau format ke bentuk atau format lain, terutama saat bekerja dengan data kategorikal. One-hot encoding adalah sebuah metode untuk mengkonversi variabel kategorikal menjadi bentuk yang dapat digunakan untuk pemodelan, khususnya dalam konteks machine learning [13]. Variabel-variabel kategorikal seperti Age_Range, Sex, Field, Education, Employability, dan platform media sosial yang digunakan (Daily_used, Accounts) dikonversi menjadi format numerik menggunakan Label Encoding. Nah, variabel-variable diatas masih mengantung variabel kategorikal, yang artinya perlu dilakukan encoding untuk mengubahnya menjadi bentuk numerik agar dapat digunakan dalam algoritma pembelajaran mesin [14].

Konversi ini dilakukan agar algoritma clustering dapat memproses variabel-variabel tersebut secara numerik. Dengan menerapkan encoding variabel kategorikal dengan benar, kita dapat memastikan bahwa algoritma machine learning dapat memproses dan memanfaatkan semua variabel dengan baik, tanpa menyebabkan bias atau kesalahan interpretasi yang disebabkan oleh perbedaan jenis data.

3. Metode Analisis

Untuk menganalisis pola penggunaan media sosial, digunakan dua teknik utama Divisive Hierarchical Clustering (DHC) dan Principal Component Analysis (PCA). Kedua teknik ini dipilih karena kemampuan mereka dalam mengidentifikasi pola pengelompokan dalam data dan memvisualisasikannya secara lebih mudah.

a. Divisive Hierarchical Clustering atau top-down merupakan metode memecah data dimulai dengan satu cluster yang berisi semua data kemudian dipecah menjadi beberapa cluster-cluster lainnya [15]. Dalam penelitian ini, dilakukan eksperimen untuk membagi data menjadi 2 hingga 10 cluster, dan hasil yang dihasilkan dinilai menggunakan Silhouette Score. Skor ini mengukur seberapa baik data dikelompokkan, di mana nilai 1 menunjukkan pengelompokan yang sempurna dan nilai mendekati -1 menunjukkan pengelompokan yang buruk.
b. Principal Component Analysis (PCA) adalah teknik yang digunakan untuk menyederhanakan suatu data dengan cara mentransformasi linier sehingga terbentuk sistem koordinat baru dengan varians maksimum. PCA dapat digunakan untuk mereduksi dimensi suatu data tanpa mengurangi karakteristik data tersebut secara signifikan [16]. Metode PCA digunakan pada penelitian ini dikarenakan tujuan dari penelitian ini adalah berusaha meringkas data dengan jumlah variabel yang lebih kecil PCA digunakan untuk mereduksi dimensi data yang kompleks agar lebih mudah divisualisasikan. Teknik ini mereduksi data ke dalam dua komponen utama yang menangkap sebagian besar variasi dalam dataset. Dengan PCA, hasil clustering dapat divisualisasikan dalam dua dimensi untuk membantu dalam interpretasi pengelompokan yang dihasilkan.

4. Evaluasi Kualitas Clustering

Setelah clustering dilakukan, kualitas pengelompokan dinilai menggunakan Silhouette Score. Silhouette Score adalah metrik evaluasi yang digunakan untuk mengukur seberapa baik suatu objek cocok dengan cluster yang telah ditetapkan. Metrik ini memberikan ukuran seberapa dekat setiap titik data dalam cluster tertentu dibandingkan dengan cluster lainnya. [16]. Skor ini digunakan untuk menentukan jumlah cluster yang optimal yang memberikan hasil clustering terbaik. Skor ini berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan objek tersebut terpisah dan lebih terorganisir dalam kluster yang benar. Dalam penelitian ini, cluster dengan jumlah lima memberikan hasil yang paling baik berdasarkan Silhouette Score. Silhouette Score tertinggi menunjukkan bahwa K=5 adalah jumlah cluster yang paling sesuai dengan struktur data. Hal ini berarti data yang di-cluster menjadi lima kelompok menunjukkan kekompakan internal yang baik, di mana objek dalam cluster lebih mirip satu sama lain daripada dengan objek di cluster lain.

Jika jumlah cluster lebih dari lima atau kurang dari lima, nilai Silhouette Score cenderung menurun, yang menunjukkan bahwa clustering tidak optimal.

5. Alat dan Perangkat Lunak

Analisis dilakukan menggunakan berbagai pustaka dan alat pemrograman berbasis Python, termasuk:

a. Pandas: Untuk pengolahan dan manipulasi data, seperti pembersihan data, pengecekan missing values, dan penghapusan baris yang tidak relevan.
b. Scikit-learn: Untuk penerapan metode Divisive Hierarchical Clustering (DHC), Principal Component Analysis (PCA), dan evaluasi clustering menggunakan Silhouette Score.
c. Matplotlib: Untuk visualisasi hasil clustering dan PCA dalam bentuk grafis, seperti scatter plots dan dendrograms, yang membantu dalam memahami hasil clustering dan

distribusi data

d. NumPy: Untuk perhitungan matematis dasar, seperti standardization pada data numerik yang digunakan dalam penskalaan dan transformasi data.

Data diolah dan dianalisis dalam lingkungan pemrograman Python, menggunakan platform Google Colab sebagai media komputasi.

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan berisi informasi dari 29 responden terkait kebiasaan penggunaan media sosial mereka, informasi demografis seperti jenis kelamin, rentang usia, tingkat pendidikan, dan status kerja. Data juga mencakup jumlah waktu yang dihabiskan di media sosial per hari (dalam jam), jenis platform media sosial yang digunakan, jumlah pengikut, dan pendapatan bulanan responden.

Dataset yang digunakan dalam penelitian ini yaitu memahami kebiasaan penggunaan dalam platform media sosial yang mana banyak digunakan para remaja dari rentang usia 18 hingga 24 tahun. Berdasarkan variabel Daily_used dan Accounts, diketahui bahwa platform seperti Instagram, YouTube, Twitter, dan TikTok menjadi media sosial yang paling sering digunakan. Selain itu, banyak responden memiliki akun di berbagai platform dan menghabiskan rata-rata 2 hingga 4 jam per hari untuk berselancar di media sosial.

Di dalam variabel Income (Pendapatan), Hrs (Jumlah Jam Kerja), serta Followers (Pengikut di Media Sosial) dilakukan normalisasi dengan menggunakan teknik Standard Scaler, yang mengubah nilai-nilai variabel tersebut hingga memiliki rata-rata 0 dan standar deviasi 1.

Tabel 1. Data Awal Standarisasi

Income	Income_Standarisasi	Hrs	Hrs_Standarisasi	Followers	Followers_Standarisasi
0.0	-0.63245553	0.5	-1.6184384	0	-0.52899986
0.0	-0.63245553	2.5	-0.22409147	6000	4.30257437
0.0	-0.63245553	4.5	1.17025546	2000	1.08152488
0.0	-0.63245553	4.5	1.17025546	3	-0.52658407
0.0	-0.63245553	2.5	-0.22409147	10	-0.52094724

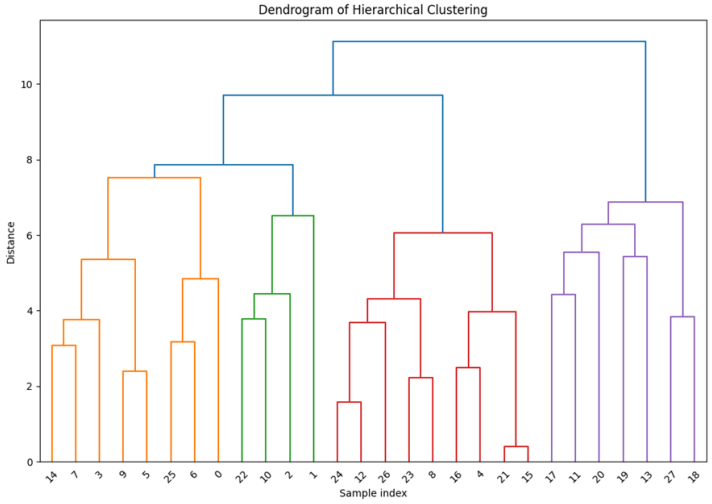
Sebelum melakukan clustering, sering kali dilakukan standarisasi atau normalisasi pada dataset agar fitur-fitur yang memiliki skala berbeda tidak mendominasi dalam pembentukan cluster. Jika tidak, fitur dengan skala besar (misalnya pendapatan) dapat memengaruhi jarak antar titik data lebih daripada fitur dengan skala kecil (misalnya usia).

Standarisasi ini memastikan bahwa semua fitur memiliki skala yang sama. Setelah standarisasi, semua fitur numerik akan memiliki rata-rata 0 dan deviasi standarisasi 1. Dengan data yang terstandarisasi, algoritma pemodelan seperti K-Means clustering, PCA, dan lainnya dapat bekerja lebih efektif tanpa terpengaruh oleh skala fitur yang berbeda.

Clasterisasi dengan Divisive Hierarchical

Hierarchical clustering digunakan untuk mengelompokkan responden berdasarkan pola penggunaan media sosial mereka. Berdasarkan dendrogram dan analisis silhouette score, ditemukan adanya beberapa cluster dengan variasi kebiasaan penggunaan yang berbeda. Dendrogram adalah bentuk dari pohon biner yang biasanya digunakan untuk memvisualisasikan hubungan dari hierarki dalam data seperti pada clustering. Sebuah dendrogram sering diintegrasikan dengan matriks yang dapat diurutkan ulang, yang disebut sebagai tampilan dendrogram-matriks, tetapi skala untuk menentukan

lokasi jarak antar cluster sangat sulit dilakukan [17]. Cluster ini memungkinkan kita mengidentifikasi kelompok pengguna dengan pola unik, seperti pengguna yang menghabiskan lebih banyak waktu di media sosial pada waktu tertentu atau memiliki keterlibatan yang lebih tinggi dalam aktivitas posting.



Gambar 1. Dendogram of Hierarchical Clustering

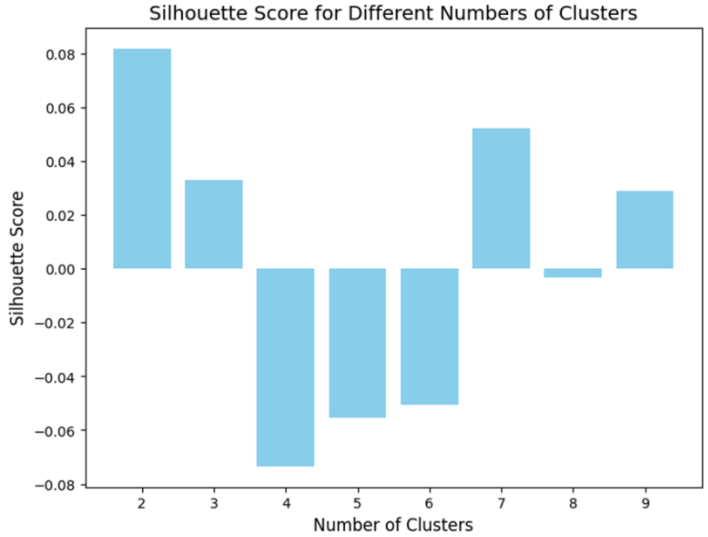
Sumbu X (Sample Index): Representasi dari setiap titik data atau kluster.

Sumbu Y (Distance): Representasi jarak atau dissimilarity saat kluster digabungkan.

Setiap percabangan dalam dendogram menunjukkan dua kluster yang digabungkan pada tahap tertentu. Semakin tinggi posisi percabangan di dendogram, semakin besar jarak antara kluster yang digabungkan. Jika dua titik data atau kluster digabungkan di bagian bawah dendogram (jarak Y rendah), ini berarti keduanya sangat mirip atau dekat satu sama lain. Jika penggabungan terjadi di bagian atas dendogram (jarak Y tinggi), berarti keduanya jauh berbeda dan baru digabungkan pada tahap yang lebih tinggi.

Silhouette Score

Dilakukan pengujian terhadap berbagai jumlah cluster, mulai dari 2 hingga 10 cluster. Skor Silhouette, yang mengukur konsistensi dan kualitas cluster (dengan nilai 1 sebagai hasil clustering sempurna dan -1 menunjukkan pembagian yang buruk), dihitung untuk setiap skenario.

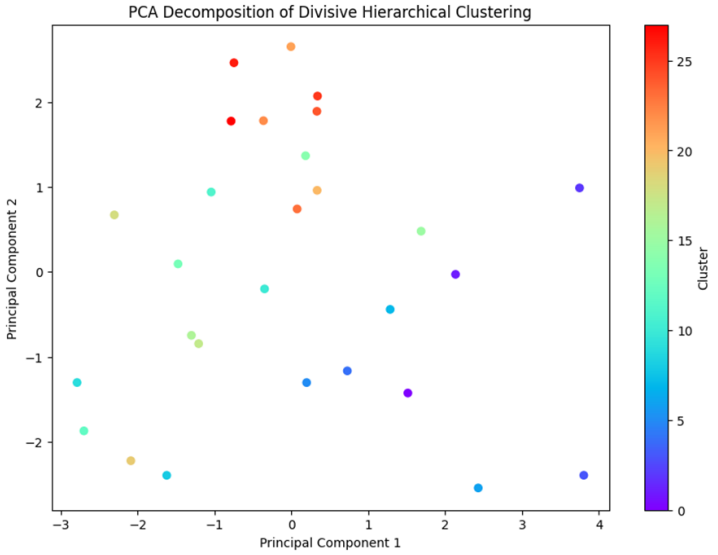


Gambar 2. Silhouette Score

Dari grafik silhouette score ini, K=2 menghasilkan nilai Silhouette Score tertinggi sekitar 0.08, yang menunjukkan bahwa pengelompokan data menjadi dua cluster adalah yang paling sesuai dengan struktur alami data. K=3 menghasilkan nilai yang lebih rendah sekitar 0.03, menunjukkan bahwa data masih dapat di-cluster menjadi tiga grup, tetapi kualitasnya mulai menurun. K=4 hingga K=6 menghasilkan nilai negatif, yang berarti penambahan jumlah cluster ini membuat kualitas pengelompokan semakin buruk. Data yang di-cluster dalam rentang ini berada lebih dekat dengan cluster yang salah. K=7 menunjukkan sedikit perbaikan, tetapi tidak cukup signifikan dibandingkan dengan dua cluster. K=8 hingga K=9 menghasilkan Silhouette Score yang sangat rendah atau negatif, yang menunjukkan bahwa semakin banyak cluster tidak meningkatkan kualitas pengelompokan.

Secara keseluruhan, dua cluster (K=2) memberikan hasil clustering yang paling optimal berdasarkan nilai Silhouette Score yang tertinggi. Penambahan cluster melebihi dua hanya memperburuk pengelompokan, dengan data yang semakin banyak ditempatkan secara tidak tepat. Oleh karena itu, pengelompokan data menjadi dua cluster merupakan representasi terbaik dari struktur data.

Principal Component Analiysis (PCA)



Gambar 3. Principal Component Analysis (PCA)

Principal Component 1 dan Principal Component 2 mewakili kombinasi fitur asli dalam bentuk dua dimensi yang paling signifikan. Sumbu X (PC1) dan sumbu Y (PC2) berfungsi untuk memproyeksikan data dalam ruang dua dimensi sehingga variabilitas data yang maksimum bisa diobservasi. Setiap titik mewakili satu objek atau sampel dalam dataset. Warna menunjukkan cluster yang berbeda. Skala warna di sebelah kanan menunjukkan berbagai cluster, di mana angka-angka merepresentasikan ID cluster. Titik yang warnanya lebih mirip berada dalam cluster yang sama atau dekat.

Warna merah dan oranye di bagian atas grafik menunjukkan sampel yang termasuk dalam cluster dengan ID cluster yang lebih tinggi (misalnya cluster 20-25). Warna biru dan ungu di bagian bawah grafik menandakan sampel yang termasuk dalam cluster dengan ID yang lebih rendah (misalnya cluster 0-5).

Dari visualisasi ini, dapat dilihat bahwa cluster dengan ID lebih tinggi (berwarna merah/oranye) sebagian besar terkelompok di bagian atas gambar, sementara cluster dengan ID lebih rendah (berwarna biru/ungu) berada di bagian bawah.

Beberapa cluster, seperti yang ditunjukkan oleh warna merah/oranye dan biru/ungu, tampak terpisah dengan baik dalam visualisasi dua dimensi ini. Ini menunjukkan bahwa

PCA berhasil memvisualisasikan hasil clustering divisive dengan baik. Di beberapa area (misalnya antara warna hijau muda dan biru terang), tampak ada sedikit tumpang tindih antara beberapa cluster. Ini mungkin menandakan bahwa beberapa cluster tidak terlalu jelas dipisahkan satu sama lain, atau mereka memiliki karakteristik yang mirip dalam ruang dua dimensi yang dihasilkan oleh PCA. Beberapa cluster memiliki titik yang lebih terkonsentrasi (seperti cluster merah), yang menunjukkan bahwa sampel dalam cluster tersebut sangat mirip. Sebaliknya, cluster lain tampak lebih tersebar (seperti cluster hijau muda), menandakan adanya variasi yang lebih besar di dalam cluster itu.

4. KESIMPULAN

Penelitian ini mengkaji kebiasaan penggunaan media sosial pada responden yang didominasi oleh remaja berusia 18 hingga 24 tahun. Dari data yang dikumpulkan, platform media sosial yang paling sering digunakan adalah Instagram, YouTube, Twitter, dan TikTok, dengan rata-rata penggunaan harian sebesar 2 hingga 4 jam. Responden juga cenderung memiliki akun di berbagai platform dan menunjukkan variasi jumlah pengikut serta pendapatan bulanan.

Untuk memastikan semua fitur numerik berada dalam skala yang sama, dilakukan normalisasi menggunakan teknik Standard Scaler, yang membuat nilai rata-rata setiap fitur menjadi 0 dengan standar deviasi 1. Langkah ini penting untuk mencegah fitur dengan skala besar seperti pendapatan mendominasi proses pengelompokan dalam analisis selanjutnya.

Dalam penelitian ini, metode Hierarchical Clustering digunakan untuk mengelompokkan responden berdasarkan pola penggunaan media sosial mereka. Berdasarkan analisis dendrogram dan Silhouette Score, ditemukan bahwa pembagian data menjadi dua cluster (K=2) memberikan hasil yang paling optimal. Skor Silhouette tertinggi sekitar 0.08 mendukung kesimpulan bahwa dua cluster memberikan representasi terbaik dari struktur alami data. Penambahan jumlah cluster lebih dari dua hanya memperburuk kualitas pengelompokan.

Penggunaan Principal Component Analysis (PCA) juga berhasil memvisualisasikan hasil pengelompokan dalam dua dimensi, memperlihatkan bahwa pengguna dalam setiap cluster memiliki karakteristik yang serupa. Temuan ini diharapkan dapat digunakan untuk menyusun strategi pemasaran yang lebih personal dan efektif, serta meningkatkan pengalaman pengguna melalui layanan yang lebih sesuai dengan kebutuhan individu.

Secara keseluruhan, penelitian ini berhasil mengidentifikasi dua kelompok utama responden berdasarkan kebiasaan penggunaan media sosial. Cluster ini mencerminkan pola perilaku yang berbeda, seperti waktu yang dihabiskan di media sosial, keterlibatan di berbagai platform, dan interaksi dengan pengikut. Hasil penelitian ini dapat menjadi dasar bagi penelitian lebih lanjut dalam memahami perilaku digital remaja dan juga untuk pengembangan strategi pemasaran yang lebih terarah di platform media sosial.

DAFTAR PUSTAKA

Zulkifli Alamtaha, Ismail Djakaria, Nisky Imansyah Yahya, "Imolementasi Algoritma Hierarchical Clustering dan No-Hierarchical Clustering untuk Pengelompokkan Penggunaan Media Sosial," Journal of Statistics and Its Application, Vols. 4, No.1, pp. 33-43, Januari, 2023.

Mulia Sosiady, Nasrullah Djamil, Ermansyah, "The Effects of Internet Addiction Disorder on Students' Learning Motivation," Al-Ishlah: Jurnal Pendidikan , Vols. 14, 3, pp. 3449-3460, September, 2022.

F. B. Putra, A. T. R. Dani and S. Wigantono, "PENERAPAN ALGORITMA HIERARCHICAL CLUSTERING DALAM PENGELOMPOKKAN KABUPATEN/KOTA DI PAPUA BERDASARKAN INDIKATOR KEMISKINAN," Putra, F. B., Dani, A. T. R., &

Wigantono, S. (2023). PENERAPAN ALGORITMA HIERARCHICAL CLUSTERING DALAM PENGELOMPOMAp (Mathematics and Applications) Journal, vol. Journal 5(2), pp. 118 - 126, Dec 2023.

A. T. R. Dani, S. Wahyuningsih and N. A. Rizki, "Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu," Jambura Journal of Mathematics, vol. 1 Nomor 2, pp. 64 - 78, Juli 2019.

D. I. Yulianti , T. I. Hermanto and M. Defriani, "Analisis Clustering Donor Darah dengan Metode Agglomerative Hierarchical Clustering," RESOLUSI: Rekayasa Teknik Informatika dan Informasi , Vols. 3, No 6, pp. 303 - 308, Juli 2023.

A. Y. N. Harahap, R. E. Sari, H. Gunawan and A. B. Nasution, "Evaluation of Product Sales Data Using Clustering Method and Hierarchical Divisive Clustering at PT.AYN," Indonesian Journal of Interdisciplinary Research in Science and Technology (MARCOPOLO), Vols. 2, No. 7, pp. 1145 - 1158, 2024.

S. Wijuniamurti, S. Nugroho and R. Rachmawati , "Agglomerative Nesting (AGNES) Method and Divisive Analysis (DIANA) Method For Hierarchical Clustering On Some Distance Measurement Concepts," JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE, Vols. 1, No 1, pp. 8 - 11, March 2022.

R. I. Borman and B. Priyopradono, "Implementasi Penerjemah Bahasa Isyarat Pada Bahasa Isyarat Indonesia (BISINDO) Dengan Metode Principal Component Analysis (PCA)," Jurnal Informatika: Jurnal Pengembangan IT (JPIT), Vols. 03, No.1, pp. 103 - 108, Januari 2018.

D. A. Nasution, H. H. Khotimah and N. Chamidah, "PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN," CESS (Journal of Computer Engineering System and Science), vol. 4 No. 1, pp. 78 - 82, Januari 2019.

Emi, D. Kusnandar and N. N. Debataraja, "PERBANDINGAN METODE COEFFICIENT OF DETERMINATION RATIO DAN REGRESI DIAGNOSTIK DALAM MENDETEKSI OUTLIER PADA ANALISIS REGRESI LINIER BERGANDA," Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster), Vols. 04, No. 3, pp. 153 - 158, 2015.

Fahrani, Farih Sartika; Qoiriah, Anita;, "Klasterisasi Musik Menggunakan K-Means," JINACS: (Journal of Informatics and Computer Science), vol. 06 Nomor 02, pp. 566 - 572, 2024.

Aulia, Rya Sofia; Atok, Raden Mohamad;, "Penentuan Panjang Optimal Data Deret Waktu Bebas Outlier dengan Menggunakan Metode Window Time," JURNAL SAINS DAN SENI ITS, Vols. 0, No. 1, pp. 137 - 143, 2017.

Simanjuntak , Lina Rahel; Ramadhan, Rizki gung; Siagian, Theresia Romauli; Anggriani, Vina;, "Komparasi Algoritma KNN dan SVM dalam Memprediksi Penyakit Stroke," Jurnal Teknik Mesin, Elektro dan Ilmu Komputer(TEKNIK), Vols. 3, No.3, pp. 60 - 74, November 2023.

A. A. A. Daniswara and K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," JINACS: (Journal of Informatics and Computer Science), Vols. 05, Nomor 01, pp. 97 - 100, 2023.

Setyawan, Dimas Ari; Fathicah, Chastine;, "PENGEMBANGAN METODE DECISION TREE DENGAN DISKRITISASI DATA DAN SPLITTING ATRIBUT MENGGUNAKAN HIERARCHICAL CLUSTERING DAN DISPERSION RATIO," in (Doctoral dissertation, Institut Teknologi Sepuluh Nopember)..

Alinafiah, Andika Muhammad; Octariadi, Barry Caesar; Sucipto;, "IMPLEMENTAS DATA MINING DALAM PENGELOLAAN STOK OBAT MENGGUNAKAN METODE K-MEANS CLUSTERING DAN ASSOCIATION ION," JIP (Jurnal Informatika Polinema), vol. 10(4), pp. 551 - 558, 2024.

Kusumastuti, Rajnaparmitha; Bayunanda, Elfandary; Rifa'i, Anggi Muhammad; Asgar, Muhammad Ryandy Ghonim; Ilmawati, Fahman Inti; Kusriani;, "Hot Spot Clustering Using Agglomerative Hierarchical Clustering (AHC) Algorithm," CogITo Smart Journal, vol. 8(2), pp. 501-513, 2022.