Jurnal Ilmu Komputer dan Sistem Informasi (JIKSI)

Vol. 10, No. 11, November 2024, hlm. 10-15

PENGELOMPOKAN DATA PENJUALAN PRODUK MONITOR DI AMAZON MENGGUNAKAN ALGORITMA GAUSSIAN MIXTURE MODEL

Ade Rizal Effendi Saragih[1], Arnita[2], Mhd Zulfikar Pinem[3], Dewi Putri Sagita Suhendra[4], Novi Anggraini Siregar[5]

rizalsaragih.4222550011@mhs.unimed.ac.id[1], arnita@unimed.ac.id[2], zulfikar@mhs.unimed.ac.id[3], dewiputri.4221250002@mhs.unimed.ac.id[4], novisiregar.4221250018@mhs.unimed.ac.id[5]

Universitas Negeri Medan

Abstrak

Penelitian ini bertujuan menganalisis dan mengelompokkan data penjualan produk monitor di Amazon menggunakan algoritma Gaussian Mixture Model (GMM) untuk mengidentifikasi pola-pola penjualan yang tidak terlihat melalui metode konvensional. Dataset dari Kaggle, berisi atribut seperti harga, merek, ukuran layar, resolusi, rasio aspek, rating, dan ulasan, digunakan setelah pra-pemrosesan untuk memastikan kualitas data optimal. Clustering dilakukan dengan GMM yang lebih fleksibel dalam mengidentifikasi struktur data kompleks dibandingkan algoritma K-Means, serta Principal Component Analysis (PCA) diterapkan untuk mereduksi dimensi data dan mempermudah visualisasi hasil clustering. Hasil penelitian menunjukkan bahwa model GMM menghasilkan pengelompokan dengan rata-rata Silhouette Coefficient sebesar 0.409, yang mengindikasikan kualitas pengelompokan cukup baik, dengan kluster yang menggambarkan segmen pasar berbeda berdasarkan variabel seperti harga, merek, dan fitur produk. Penelitian ini memberikan wawasan baru dalam segmentasi pasar monitor di Amazon serta mendukung pengambilan keputusan strategis dalam bisnis, terutama strategi pemasaran yang lebih terfokus, dan menunjukkan efektivitas GMM dalam menganalisis data penjualan kompleks, yang berpotensi diterapkan pada sektor bisnis lainnya.

Kata Kunci: Data Clustering, Gaussian Mixture Model, Penjualan E-Commerce.

Abstract

This research aims to analyze and cluster the sales data of monitor products on Amazon using the Gaussian Mixture Model (GMM) algorithm to uncover sales patterns that may not be detected through conventional methods. The dataset, sourced from Kaggle, contains attributes such as price, brand, screen size, resolution, aspect ratio, ratings, and reviews, which were preprocessed to ensure data quality before analysis. Clustering was performed using GMM, which is more flexible in identifying complex data structures compared to algorithms like K-Means. Additionally, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data and facilitate the visualization of clustering results. The study's findings show that the GMM model achieved a clustering performance with an average Silhouette Coefficient of 0.409, indicating fairly good clustering quality. The resulting clusters represent different market segments based on variables such as price, brand, and product features. This research offers new insights into monitor market segmentation on Amazon and supports strategic decision-making in business, particularly in focused marketing strategies. It also demonstrates the effectiveness of GMM in analyzing complex sales data, which could be applied to other business sectors.

Keywords: *Data Clustering, Gaussian Mixture Model, E-Commerce Sales.*

1. PENDAHULUAN

Dalam era globalisasi dan perkembangan teknologi yang pesat, penjualan produk melalui platform E-Commerce seperti Amazon terus mengalami pertumbuhan yang signifikan. Monitor,

sebagai salah satu produk yang banyak dicari di Amazon, memiliki variasi yang sangat luas baik dari segi spesifikasi teknis, harga, maupun merek. Konsumen yang memiliki kebutuhan berbeda sering kali mengalami kesulitan dalam memilih monitor yang tepat, sementara penjual juga dihadapkan pada tantangan dalam memahami preferensi konsumen untuk menyusun strategi pemasaran yang efektif. Oleh karena itu, diperlukan analisis yang tepat untuk mengelompokkan produk berdasarkan pola penjualan, guna memudahkan konsumen dalam memilih dan membantu penjual dalam menargetkan pasar yang sesuai.

Dataset yang digunakan dalam penelitian ini adalah Amazon Products Sales Monitor Dataset dari Kaggle, yang menyediakan informasi komprehensif tentang penjualan monitor di Amazon. Dataset ini mencakup fitur seperti harga, rating, jumlah ulasan, dan kategori produk, yang memungkinkan analisis mendalam terhadap pola penjualan. Pendekatan untuk menangani data dengan distribusi yang tidak selalu linier dan pola yang tidak teratur. Di sinilah peran algoritma yang mampu menangkap kompleksitas data tersebut menjadi sangat penting. Sebelum menganalisis data, dilakukan proses pembersihan atau preprocessing data. Data preprocessing merupakan serangkaian tahap yang diterapkan pada data mentah sebelum digunakan untuk analisis lanjutan atau pengembangan model [1]. Langkah ini bertujuan untuk meningkatkan kualitas data, menjamin keakuratan hasil analisis, serta mengatasi berbagai permasalahan yang mungkin terdapat pada data mentah [2].

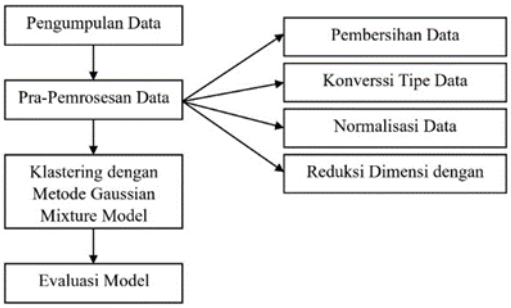
Clustering adalah proses mengelompokkan data atau objek ke dalam kelas berdasarkan kemiripan atribut. Ini merupakan metode dalam data mining, di mana pengelompokan yang baik menghasilkan objek yang memiliki kesamaan tinggi dalam satu cluster, tetapi kesamaan rendah dengan objek di cluster lain [3]. Gaussian Mixture Model (GMM) adalah metode clustering non-hierarki yang memodelkan data dengan distribusi Gaussian untuk memodelkan data dengan parameter mean dan variansi tertentu. Model ini membentuk cluster berdasarkan probabilitas dari fungsi kepadatan, sehingga memungkinkan pembagian kelompok data secara lebih fleksibel dan akurat [4]. Dalam melakukan clustering, Algoritma Gaussian Mixture Model adalah salah satu jenis soft clustering dimana satu data point bisa berada pada dua atau lebih cluster. Algoritma Gaussian Mixture Model sendiri merupakan model statistik yang sangat populer dan umum digunakan. Selain itu, algoritma ini relatif efisien untuk pengolahan objek dalam jumlah besar [5]. Berdasarkan hasil penelitian [5], dengan GMM, setiap titik data dapat memiliki probabilitas untuk termasuk ke dalam beberapa kluster, sehingga hasil pengelompokan lebih kaya informasi dan dapat menangkap variabilitas yang lebih luas di dalam data. Keunggulan utama GMM terletak pada kemampuannya dalam menangani data dengan distribusi yang beragam [6]. Dalam kasus penjualan monitor di Amazon, pola penjualan bisa jadi dipengaruhi oleh berbagai faktor seperti musim, promosi, atau preferensi regional yang menyebabkan variasi data yang tidak selalu mengikuti pola yang sederhana. GMM mampu menangkap variasi ini dengan lebih baik karena dapat mengidentifikasi kluster-kluster dengan bentuk yang lebih kompleks, bahkan kluster yang saling tumpang tindih [7].

Oleh karena itu, implementasi algoritma Gaussian Mixture Model (GMM) dalam pengelompokan data penjualan monitor di Amazon diharapkan mampu memberikan solusi yang efektif dalam mengidentifikasi pola penjualan yang kompleks. Metode ini mendukung analisis yang lebih mendalam dan komprehensif, sehingga dapat meningkatkan efektivitas

strategi pemasaran dan pengambilan keputusan di platform e-commerce.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk mengelompokkan data penjualan produk monitor di platform Amazon menggunakan Algoritma Gaussian Mixture Model (GMM). Data yang digunakan mencakup berbagai informasi tentang monitor yang tersedia di Amazon, termasuk spesifikasi produk, harga, dan ulasan pelanggan. Berikut ini adalah tahapan metodologi yang dijalankan dalam penelitian ini:



Gambar 1. Alur Penelitian

3. HASIL DAN PEMBAHASAN

Pada penelitian ini, penulis melakukan data clustering terhadap data penjualan monitor di E-Commerce Amazon menggunakan metode Gaussian Mixture Model. Dataset penjualan monitor yang digunakan terdiri dari 7 kolom, yaitu Title, Brand, Screen Size, Resolution, Aspect Ratio, Rating, dan Price.

	Title	Brand	Screen Size	Resolution	Aspect Ratio	Rating	Price
0	acer SB240Y G0bi 23.8" IPS Full HD Ultra-Slim ...	acer	23.8 Inches	FHD 1080p	16:9	4.4	94.99
1	acer Nitro 31.5" FHD 1920 x 1080 1500R Curved ...	acer	31.5 Inches	FHD 1080p	16:9	4.6	259.99
2	Acer SB272 EBI 27" Full HD (1920 x 1080) IPS Z...	acer	27 Inches	FHD 1080p	16:9	4.5	99.99
3	Sceptre 30-inch Curved Gaming Monitor 21:9 256...	Sceptre	30 Inches	FHD 1080p Ultra Wide	21:9	4.5	199.97
4	SAMSUNG 32" UJ59 Series 4K UHD (3840x2160) Com...	SAMSUNG	31.5 Inches	4K UHD 2160p	16:9	4.3	279.99

Gambar 2. Lima Data Teratas

1. Deskripsi Data

	Title	Brand	Screen Size
count	947	947	947
unique	296	62	57
top	ASUS 31.5" 1080P Monitor (VA329HE) - Full HD, ...	acer	27 Inches
freq	84	532	260

Gambar 3. Deskripsi Data

Gambar di atas menunjukkan deskripsi dari dataset yang digunakan. Terdapat 947 entri data monitor, dengan 296 judul monitor yang unik. Judul monitor yang paling sering muncul adalah "ASUS 31.5' 1080P Monitor (VA329HE)" dengan frekuensi kemunculan sebanyak 84 kali. Terdapat 62 merek monitor berbeda. Merek yang paling sering muncul adalah "acer" sebanyak 532 kali. Ukuran layar monitor memiliki 57 ukuran yang berbeda. Ukuran layar yang paling sering muncul adalah 27 inci sebanyak 260 kali. Terdapat 41 resolusi berbeda. Resolusi yang paling sering muncul adalah "FHD 1080p" dengan 566 kali kemunculan. Ada 16 rasio aspek berbeda. Rasio aspek yang paling sering muncul adalah 16:9 sebanyak 833 kali. Data berisi 25 nilai rating yang berbeda. Rating tertinggi adalah 4.6 dan muncul sebanyak 368 kali. Ada 166 harga yang berbeda dengan harga yang paling sering muncul adalah $199.99 sebanyak 122 kali.

2. Konversi Tipe Data, Mengatasi Missing Value, dan Normalisasi Data

```
Screen Size setelah konversi:
0    0.643079
1    0.569997
2    0.138943
3    0.333683
4    0.569997
Name: Screen Size, dtype: float64

Rating setelah konversi:
0   -0.177481
1    0.532443
2    0.177481
3    0.177481
4   -0.532443
Name: Rating, dtype: float64

Price setelah konversi:
0   -0.841962
1    0.446504
2   -0.802917
3   -0.022185
4    0.602682
Name: Price, dtype: float64
```

```
Data setelah mengisi nilai yang hilang:
   Screen Size    Rating     Price
0     0.643079 -0.177481 -0.841962
1     0.569997  0.532443  0.446504
2     0.138943  0.177481 -0.802917
3     0.333683  0.177481 -0.022185
4     0.569997 -0.532443  0.602682
```

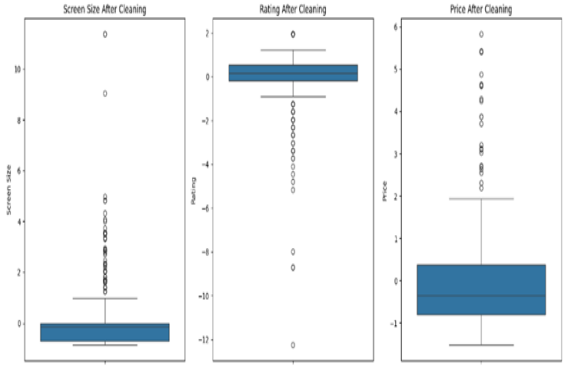
```
Data setelah normalisasi:
   Screen Size    Rating     Price
0    -0.029848 -0.177481 -0.841962
1    -0.127744  0.532443  0.446504
2    -0.705160  0.177481 -0.802917
3    -0.444297  0.177481 -0.022185
4    -0.127744 -0.532443  0.602682
```

Gambar 4. Konversi Tipe Data, Mengisi Missing Value, dan Normalisasi Data

Selanjutnya, dilakukan beberapa hal seperti melakukan konversi tipe data dari kolom screen size, rating, dan price yang sebelumnya bertipe data object menjadi float. Kemudian dilakukan pengecekan data hilang dan jika ada maka akan diisi menggunakan SimpleImputer dengan strategi rata-rata. Setelah itu, dilakukan normalisasi data

menggunakan StandardScaler dari module scikit-learn.

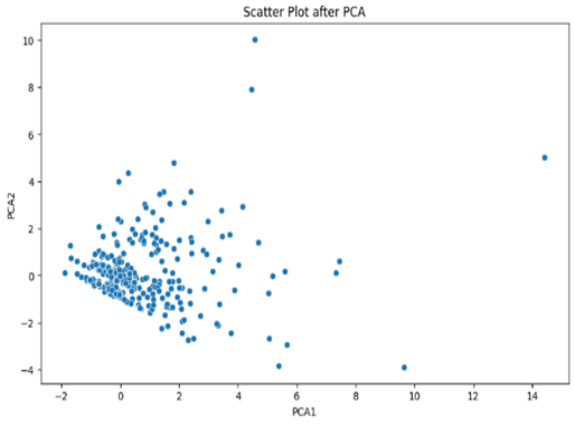
3. Boxplot Setelah Data Cleaning



Gambar 5. Boxplot Setelah Data Cleaning

Dari boxplot tersebut, didapati beberapa informasi seperti ada variasi ukuran layar yang cukup signifikan, dengan beberapa perangkat memiliki ukuran layar yang jauh lebih besar dari rata-rata. Rata-rata rating cenderung rendah, dengan banyak perangkat mendapatkan rating yang sangat rendah. Kemungkinan ada beberapa faktor yang menyebabkan hal ini, seperti kualitas perangkat setelah cleaning, atau mungkin ekspektasi pengguna yang terlalu tinggi. Rata-rata harga juga cenderung rendah, dengan beberapa perangkat memiliki harga yang sangat rendah. Ini bisa mengindikasikan adanya promo atau diskon yang signifikan, atau mungkin kualitas perangkat yang lebih rendah.

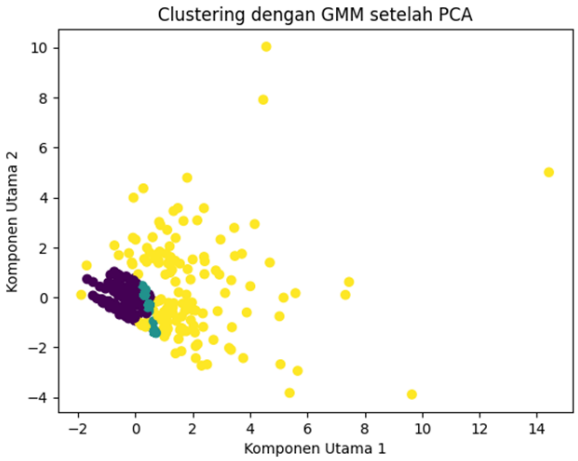
4. Reduksi Dimensi



Gambar 6. Reduksi Dimensi

Setelah melakukan preprocessing (seperti konversi tipe data, imputasi, dan normalisasi), langkah selanjutnya dalam mereduksi data dapat dilakukan dengan menggunakan teknik Dimensionality Reduction seperti Principal Component Analysis (PCA). Teknik ini akan membantu mengurangi jumlah fitur sambil tetap mempertahankan sebanyak mungkin informasi dari data asli.

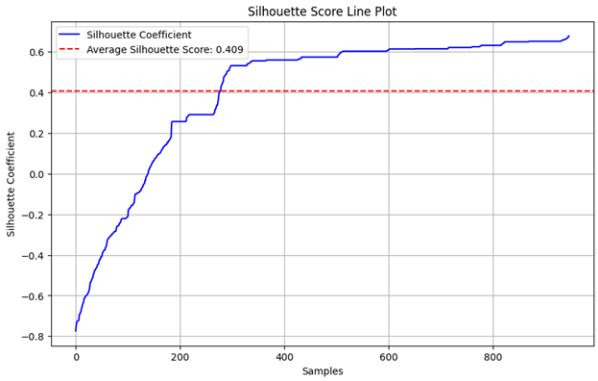
5. Data Clustering



Gambar 7. Scatter Plot Data

Scatter plot di atas merupakan hasil dari metode Gaussian Mixture Model dengan jumlah komponen yang digunakan yaitu 3. Sebagian besar data terkelompok rapat di area antara (-2, -2) hingga (4, 2), namun terdapat beberapa outlier atau data yang terletak jauh dari pusat kluster (ditandai dengan titik kuning yang tersebar).

6. Sillhouette Score



Gambar 8. Sillhouette Score

Pada grafik ini, nilai rata-rata Silhouette Coefficient adalah 0.409. Nilai ini

menunjukkan bahwa secara keseluruhan, kualitas pengelompokan cukup baik. Data point cenderung lebih cocok dengan clusternya dibandingkan dengan cluster lainnya. Fluktuasi nilai Silhouette Coefficient menunjukkan bahwa ada beberapa data point yang memiliki kualitas pengelompokan yang lebih baik dibandingkan dengan data point lainnya. Ini bisa disebabkan oleh beberapa faktor, seperti kerapatan data di sekitar titik tersebut atau bentuk cluster yang tidak terlalu kompak.

7. Evaluasi Model

Untuk mengevaluasi hasil clustering yang dihasilkan oleh Gaussian Mixture Model (GMM), ada beberapa metrik dan metode yang bisa digunakan. Evaluasi ini membantu menilai kualitas clustering dan apakah model GMM sesuai dengan data. Berikut adalah beberapa metode yang umum digunakan.

1) Log-Likelihood

```
[ ] log_likelihood = gmm.score(df_pca)
    print(f'Log-Likelihood: {log_likelihood:.3f}')

    Log-Likelihood: -1.691
```

Gambar 9. Nilai Log-Likelihood

GMM mengoptimalkan log-likelihood untuk menemukan parameter terbaik yang menjelaskan data. Nilai log-likelihood yang lebih tinggi menunjukkan model yang lebih baik dalam memodelkan data [11].

2) Akaike Information Criterion (AIC) & Bayesian Information Criterion (BIC)

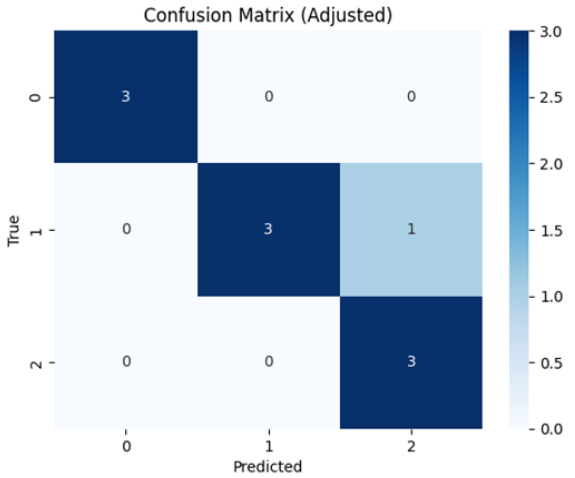
```
[ ] aic = gmm.aic(df_pca)
    bic = gmm.bic(df_pca)
    print(f'AIC: {aic:.3f}')
    print(f'BIC: {bic:.3f}')

    AIC: 3236.749
    BIC: 3319.255
```

Gambar 10. Nilai AIC dan BIC

AIC dan BIC adalah kriteria berbasis informasi yang digunakan untuk memilih model terbaik di antara beberapa kandidat [8]. Keduanya mempertimbangkan trade-off antara kecocokan model (dengan log-likelihood) dan kompleksitas model (dengan jumlah parameter).

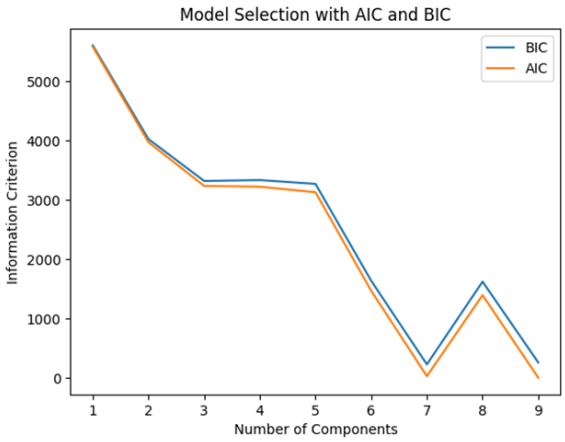
3) Confusion Matrix (Jika Ground Truth Tersedia)



Gambar 11. Confusion Matrix

Jika datanya memiliki label ground truth, Kita bisa membuat confusion matrix untuk membandingkan hasil klastering dengan label asli. Jumlah true positive untuk kelas 0 dan 2 cukup tinggi, menunjukkan bahwa model mampu mengklasifikasikan data yang benar-benar berasal dari kelas tersebut dengan baik. Terdapat beberapa kasus di mana data yang sebenarnya berkelas 1 diprediksi sebagai kelas 2. Ini mengindikasikan bahwa model masih kesulitan dalam membedakan antara kelas 1 dan 2.

4) Perbandingan Antar Model GMM dengan Jumlah Komponen Berbeda



Gambar 12. Grafik perbandingan Model GMM dengan Komponen yang Berbeda

Berdasarkan grafik tersebut, model dengan jumlah komponen tertentu (misalnya, 3 atau 4) memiliki nilai BIC dan AIC yang lebih rendah dibandingkan dengan jumlah komponen lainnya. Ini mengindikasikan bahwa model dengan jumlah komponen tersebut adalah model yang lebih baik. Dalam kasus ini, baik BIC maupun AIC memberikan hasil yang serupa. Namun, dalam kasus lain, keduanya mungkin memberikan hasil yang berbeda.

4. KESIMPULAN

Penelitian ini berhasil melakukan clustering data penjualan monitor di Amazon menggunakan metode Gaussian Mixture Model (GMM). Hasil analisis menunjukkan bahwa GMM efektif dalam mengidentifikasi pola penjualan dan segmen pasar yang berbeda. Dengan menggunakan GMM, penelitian ini memberikan wawasan mendalam tentang karakteristik unik dari setiap segmen pasar monitor di Amazon. Hasil ini diharapkan dapat membantu dalam pengambilan keputusan bisnis dan strategi pemasaran yang lebih baik

5. DAFTAR PUSTAKA

A. Milson, D. E. Herwindiati, and N. J. Perdana, “Penerapan Klasifikasi Suara Sebagai Autentikasi Keamanan Sistem Login Menggunakan Gaussian Mixture Models,” Computatio: Journal of Computer Science and Information Systems, vol. 8, no. 1, pp. 104–109, 2024.

E. Patel and D. S. Kushwaha, “Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model,” Procedia Comput Sci, vol. 171, pp. 158–167, 2020.

F. Marisa, A. L. Maukar, and T. M. Akhriza, Data Mining Konsep Dan Penerapannya. Yogyakarta: Deepublish Publisher, 2021.

I. Pii, N. Suarna, and N. Rahaningsih, “Penerapan Data Mining Pada Penjualan Produk Pakaian Dameyra Fashion Menggunakan Metode K-Means Clustering,” JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 1, pp. 423–430, Feb. 2023.

Muttaqin et al., Pengenalan Data Mining. Yayasan Kita Menulis, 2023.