

PENGELOMPOKAN PENJUALAN PRODUK DENGAN MENGUNAKAN *K-MEANS CLUSTERING* : STUDI KASUS ANALISIS PENJUALAN *COFFEE SHOP* OLEH *KAGGLE.COM*

Sifa Rismawati^{1(*)}, Shofa Shofiah Hilabi², Bayu Priyatna³, Agustia Hananto⁴

¹Universitas Buana Perjuangan Karawang

²Universitas Buana Perjuangan Karawang

³Universitas Buana Perjuangan Karawang

⁴Universitas Buana Perjuangan Karawang

Abstract

The purpose of this study is to categorize coffee shop items sold by analyzing sales records with the K-Means clustering method. The Kaggle dataset consists of three main features: Product_ID, number of items sold (Transaction_Qty), and price per item (unit_price). This method was chosen because of its ability to identify sales patterns in grouping products into three clusters, namely cluster 0 has 56 products with the highest sales, cluster 1 has 23 products with medium sales, cluster 2 has 1 product with low sales. This study involved collecting data, cleaning it, standardizing it, finding out the best number of clusters, and then checking the results with a silhouette score of 0.67. The findings show that the K-Means method helps create products that are useful for business decisions, and also improves data-driven stock management and marketing plans.

Kata Kunci: *K-Means Clustering, Data Mining, Product Sales, Data Analysis, Silhouette Score*

Informasi Artikel:

Dikirim : 14 Mei 2025

Ditelaah : 15 Mei 2025

Diterima : 20 Mei 2025

Publikasi : 25 Juni 2025

Januari – Juni 2025, Vol 6 (1) : hlm 1-10

©2025 Institut Teknologi dan Bisnis Ahmad Dahlan.

All rights reserved.

PENDAHULUAN

Kemajuan teknologi informasi telah berkembang dan mendorong sebuah pertumbuhan data dalam jumlah yang sangat besar diberbagai *industry* seperti bisnis retail yaitu pada sebuah *coffee shop*. Sejumlah data besar pada penjualan dapat diperiksa untuk mendapatkan pemahaman terperinci untuk meningkatkan strategi berdasarkan informasi. Teknik yang digunakan ialah *Data mining* sebuah analisis umum dalam pengelompokan *K-means Clustering*, (Priyatman, 2019) karena dapat mengkategorikan informasi dengan kemiripan fitur spesifik dalam pengelompokan datanya. (Deden, 2024)

Beberapa penelitian sebelumnya telah membuktikan efektivitas metode *K-Means* dalam mengelompokkan data penjualan. (Prasetiani & Rochmawati, 2022) Menerapkan metode *K-Means* untuk mengkategorikan produk pilihan di sebuah *coffee shop* secara efektif menentukan selera pelanggan dari pembelian. Selain itu, (Awalina & Rahayu, 2023) Menerapkan pengelompokan *K-Means* untuk mengkategorikan pembelian *marketplace* dan mencapai pengelompokan yang meningkatkan taktik pemasaran untuk mempermudah pemilik bisnis dalam mengambil keputusan. (Hananto, 2022) Sementara itu, *K-Means* mengurutkan angka populasi, menunjukkan kemampuan beradaptasi dari metode ini untuk berbagai jenis informasi *numerik*. Dengan begitu metode *K-Means* diterapkan untuk mengelompokkan produk suatu data berdasarkan volume penjualan dan harga. (Nur Afidah, 2023)

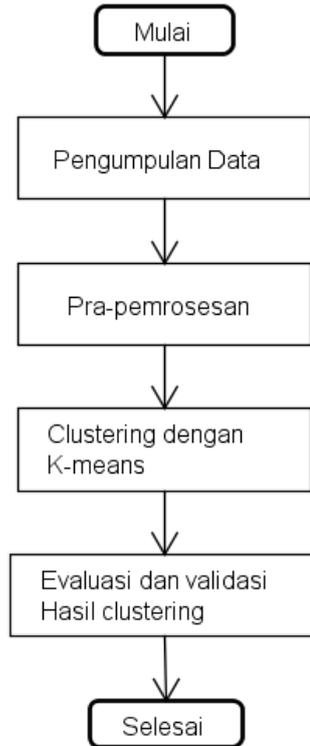
Berbeda-beda dari penelitian sebelumnya, Penelitian ini Menggunakan *dataset* studi kasus dari *kaggle.com* yang terdapat transaksi-transaksi penjualan pada *coffee shop* dalam mengelompokkan produk berdasarkan frekuensi transaksi dan biaya per unit. Keunikan dalam penelitian ini terdapat pada dua kombinasi variabel dalam tahap proses *clustering* adalah *total_sales*, dan *average_price* yang kemudian terdapat evaluasi dari *silhouette score* yaitu dengan hasil diperoleh nilai sebesar 0,67. Dalam mengukur pengelompokan validasi hasilnya data yang diperoleh (Siahaan, 2024).

Adapun tujuan dari penelitian ini adalah pengelompokan dalam mengumpulkan data maupun dalam menganalisis data penjualan produk berdasarkan peningkatan kemajuan suatu bisnis *coffee shop* dengan menggunakan *algoritma K-means clustering*, (Hilabi, 2017) Selain itu mendapatkan jumlah *cluster* yang lebih optimal berdasarkan evaluasi *silhouette score* dengan memberikan rekomendasi berbasis teknologi data, (Priyatna, 2019) Serta memberikan kemudahan bagi para pemilik bisnis. (Ayu pangestu, 2021).

Pada penelitian-penelitian di atas menunjukkan sebuah gambaran *dataset* berdasarkan data kategori produk, harga per unit, dan waktu pembelian yang dimana menggunakan sebuah algoritma yang sama yaitu *K-Means* tetapi dengan cara yang berbeda, (Lia Hananto, 2021) Maupun karakteristik dalam suatu kelompok data. (Azzahra & Amru Yasir, 2024).

METODE

Dalam penelitian ini, teknik yang digunakan adalah *k-means clustering* untuk memeriksa catatan penjualan sebelumnya. Data fitur atribut seperti *product_id*, jumlah transaksi, dan harga per unit. Penelitian ini berupaya mengungkap pola dan tren penting yang dapat membantu membentuk strategi bisnis. Berikut adalah tahapan-tahapan pada penelitiannya dibawah ini. :



Gambar 1. Flowchart prosedur metode penelitian

Gambar 1, menggambarkan perkembangan penelitian yang berfokus pada pengkategorikan penjualan produk melalui penerapan pengelompokan *K-Means*. Studi ini bertujuan untuk meningkatkan kategorisasi produk, dan temuan ini akan berfungsi sebagai panduan bagi pemilik bisnis.

1. Pengumpulan Data

Sistem pada penelitian ini mengumpulkan data penjualan, yang berisi detail seperti *Id product*, jumlah total penjualan, dan biaya untuk setiap *item*. Informasi dalam penelitiannya berasal dari bulan Januari hingga Desember 2024 (Prasetyo, 2024).

2. *Pra-pemrosesan*

Pada titik ini, informasi yang dikumpulkan ditinjau untuk mengkonfirmasi tidak ada nilai yang hilang atau titik data yang tidak biasa. Data yang tidak lengkap atau salah dapat dikoreksi atau dihitung menggunakan teknik yang tepat. Untuk meningkatkan keakuratan pengelompokan, Karna penting untuk menormalkan data sehingga semua fitur berada pada skala yang sebanding. (Prasetyani & Rochmawati, 2022)

3. *Clustering dengan K-Means*

Untuk mengetahui jumlah kelompok terbaik, pada penelitian ini terdapat tiga *cluster* yang pertama *cluster* tinggi dengan penjualan terbanyak, *cluster* penjualan sedang dan *cluster* penjualan yang rendah. Untuk mendapatkan hasil terbaik, formula digunakan selama *fase* pengelompokan dengan metode *K-Means*, yang melibatkan dua langkah utama mengukur jarak ke titik tengah terdekat untuk mengatur data. Langkah pertama adalah menghitung *centroid* dari setiap *cluster*. (Hidayat & Kusniyati, 2022)

Dalam memperoleh hasil yang lebih maksimal terdapat rumus yang akan diterapkan dalam proses *clustering* dengan *algoritma k-means*. Ada dua perhitungan untuk proses penelitiannya yaitu memakai rumus perhitungan jarak *Euclidean* untuk mengelompokkan data ke *centroid* terdekat. Dalam hal ini, dalam menggunakan *k-means clustering* rumus yang digunakan adalah sebagai berikut.

a. Rumus jarak *Euclidean*

Rumus ini digunakan jarak antara data ke-*i* dengan *centroid* ke-*j* :

$$d(x, c) = \sqrt{\sum(x_i - c_i)^2}$$

Keterangan :

x = vektor data (misalnya *total_sales* dan *avg_price*)

c = vektor *centroid*

n = jumlah atribut

d(x,c) = jarak *Euclidean* antara data dan *centroid*

b. Rumus Pembaruan *centroid*

Pada rumus *centroid* ini akan diperbarui untuk menghitung rata-rata semua data.

$$C_j = (1 / |K_j|) * \sum(x \in K_j)$$

Keterangan:

- *C_j* = *centroid* baru dari kluster ke-*j*

- *K_j* = himpunan data dalam kluster ke-*j*

- *x* = data pada kluster tersebut

Pada rumus diatas merupakan rumus yang akurat dalam pengelompokan data penjualan produk *coffee shop*, sehingga menghasilkan 3 kluster utama yang mewakili produk dengan penjualan tinggi, sedang, dan rendah. (Tri Cahaya, 2024)

4. Evaluasi dan Validasi *Hasil Clustering*

Setelah jumlah *klaster* ditentukan, *algoritma K-Means* diterapkan pada data yang telah di *pra-pemrosesan*. *Algoritma* ini bekerja dengan menginisialisasi *centroid* secara acak, kemudian mengiterasi proses pengelompokan data ke *centroid* terdekat dan memperbarui posisi *centroid* hingga konvergensi tercapai. Proses ini memastikan bahwa data dikelompokkan ke dalam kluster dengan karakteristik yang serupa. (Kamalludin & Budiarto, 2023).

HASIL DAN PEMBAHASAN

Pada bab akhir ini menjelaskan sebuah hasil dan proses analisis data penjualan pada sebuah *coffee shop* dengan menggunakan metode *clustering K-Means*. Pada hasil ini terdapat

pengelompokan produk dengan tujuan dapat membantu pihak pemilik bisnis dalam memahami tren penjualan, mengoptimalkan stok, dan merancang strategi pemasaran yang lebih efektif, Berikut hasil tampilan dalam menganalisis data penjualan *coffee shop*.

1. Pengumpulan data

Data:

	transaction_id	transaction_date	transaction_time	transaction_qty	store_id	store_location	product_id	unit_price	product_category	product_type	product_detail
0	1	1/1/2023	7:06:11	2	5	Lower Manhattan	32	3.00	Coffee	Gourmet brewed coffee	Ethiopia Rg
1	2	1/1/2023	7:08:56	2	5	Lower Manhattan	57	3.10	Tea	Brewed Chai tea	Spicy Eye Opener Chai Lg
2	3	1/1/2023	7:14:04	2	5	Lower Manhattan	59	4.50	Drinking Chocolate	Hot chocolate	Dark chocolate Lg
3	4	1/1/2023	7:20:24	1	5	Lower Manhattan	22	2.00	Coffee	Drip coffee	Our Old Time Diner Blend Sm
4	5	1/1/2023	7:22:41	2	5	Lower Manhattan	57	3.10	Tea	Brewed Chai tea	Spicy Eye Opener Chai Lg
149111	149452	6/30/2023	20:18:41	2	8	Hell's Kitchen	44	2.50	Tea	Brewed herbal tea	Peppermint Rg
149112	149453	6/30/2023	20:25:10	2	8	Hell's Kitchen	49	3.00	Tea	Brewed Black tea	English Breakfast Lg
149113	149454	6/30/2023	20:31:34	1	8	Hell's Kitchen	45	3.00	Tea	Brewed herbal tea	Peppermint Lg
149114	149455	6/30/2023	20:57:19	1	8	Hell's Kitchen	40	3.75	Coffee	Barista Espresso	Cappuccino
149115	149456	6/30/2023	20:57:19	2	8	Hell's Kitchen	64	0.80	Flavours	Regular syrup	Hazelnut syrup

Jumlah baris: 149116

Gambar 2. Variabel *dataset*

Pada Gambar 2, *Dataset* berisi kolom-kolom seperti *product_id*, *transaction_qty*, dan *unit_price*. Dari data tersebut merupakan transaksi penjualan dari *coffee shop* yang dikumpulkan dalam bentuk file *Excel* atau *CSV*.

a. *Dataset*

Pada tahap ini *dataset* penelitian akan di proses menggunakan tahap pendahuluan dan *pra-pemrosesan* dalam menampilkan data.

```
product_id      0
total_sales     0
avg_price       0
dtype: int64
Data:
```

	product_id	total_sales	avg_price
0	1	214	18.000000
1	2	183	18.000000
2	3	169	14.750000
3	4	150	20.450000
4	5	148	15.000000
75	81	221	27.886878
76	82	240	12.247748
77	83	315	14.325658
78	84	2658	0.800000
79	87	3262	2.696220

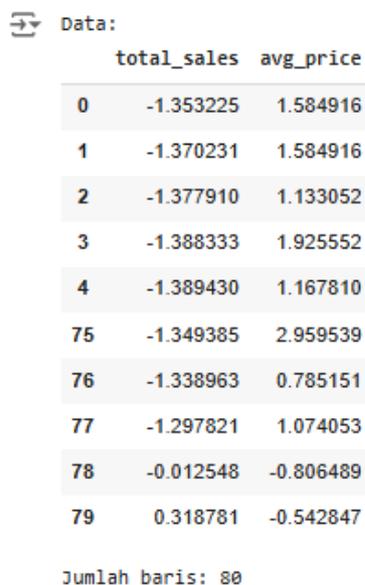
Jumlah baris: 80

Gambar 3. *pra-pemrosesan* data

Pada Gambar 3, struktur *dataset* diperiksa sebelumnya, dengan memeriksa keberadaan kolom, tipe data, dan volume data. Mendeteksi informasi yang hilang atau data yang salah. Identifikasi fitur utama untuk pemeriksaan (seperti *Product_ID*, jumlah transaksi, dan harga per unit). Sebelum menganalisis data, kelompokkan informasi penjualan dengan *Product_id* untuk mengetahui total penjualan (*Total_Sales*) dan harga jual rata-rata (*Avg_Price*).

b. Normalisasi Data

Pada Gambar 4, Data pertama kali dinormalisasi menggunakan skala standar untuk memastikan semua fitur berada pada skala yang sama sebelum pengelompokan. Di dalamnya terdapat fitur seperti penjualan total dan harga rata-rata, yang ditampilkan dalam 80 baris.



The image shows a data table with the following structure:

	total_sales	avg_price
0	-1.353225	1.584916
1	-1.370231	1.584916
2	-1.377910	1.133052
3	-1.388333	1.925552
4	-1.389430	1.167810
75	-1.349385	2.959539
76	-1.338963	0.785151
77	-1.297821	1.074053
78	-0.012548	-0.806489
79	0.318781	-0.542847

Jumlah baris: 80

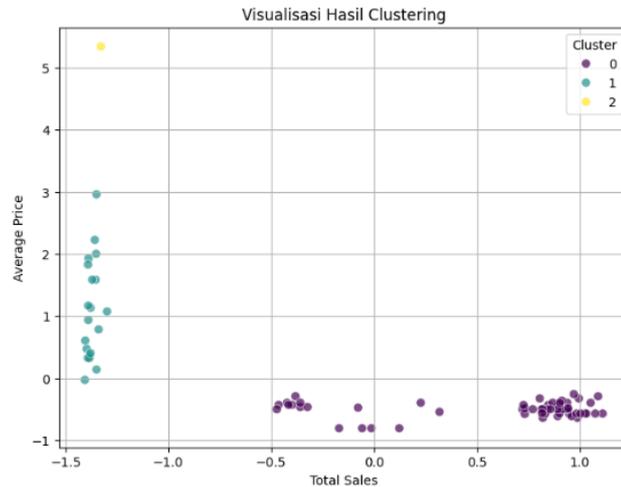
Gambar 4. Normalisasi data

Gambar 4, Normalisasi fitur (*Total_sales* dan *Avg Price*) untuk memastikan bahwa skala data tidak mempengaruhi algoritma pengelompokan dalam menormalkan data.

2. *Clustering* dengan *K-means*

a. Visualisasi hasil *clustering*

Hasil analisis pada tahap ini dapat dilihat pada Gambar 5, yang menunjukkan bahwa jumlah optimal *cluster* berada di divisi ke -3 dari *cluster*, pada tahap pengelompokan kemudian divisualisasikan menggunakan *plot pencar*, yang menunjukkan bahwa produk dalam kategori tertentu cenderung memiliki pola penjualan yang konsisten. *Silhouette Score* yang mengukur seberapa baik produk dikelompokkan berdasarkan tren penjualan, memiliki hasil 0. (Awalina & Rahayu, 2023)



Gambar 5. Visualisasi Hasil Clustering

Pada gambar 5, Hasil yang di tampilkan dengan perhitungan pada algoritma *k-means clustering*, Dengan total penjualan dan harga rata-rata warna yang berbeda menunjukkan *cluster* yang berbeda.

b. Hasil *clustering*

Langkah selanjutnya adalah menentukan nilai *centroid* dan mengelompokkan total penjualan ke dalam *cluster* yang sesuai.yang sesuai.

```

total_sales  avg_price  cluster
0      -1.353225  1.584916  1
1      -1.370231  1.584916  1
2      -1.377910  1.133052  1
3      -1.388333  1.925552  1
4      -1.389430  1.167810  1
..      ...      ...      ...
75     -1.349385  2.959539  1
76     -1.338963  0.785151  1
77     -1.297821  1.074053  1
78     -0.012548  -0.806489  0
79      0.318781  -0.542847  0

```

[80 rows x 3 columns]

Gambar 6. Hasil Clustering

Setelah menerapkan perhitungan pada algoritma *k-means clustering*, Diperoleh pengelompokan seperti pada gambar 6.

3. Evaluasi dan Validasi hasil *Clustering*

a. Evaluasi

Silhouette Score digunakan untuk mengevaluasi efektivitas pengelompokan, dan *plot pencar* dibuat untuk memvisualisasikan distribusi kelompok dalam data penjualan.

```
→ Silhouette Score: 0.6748114993369614
```

Gambar 7. *Silhouette Score*

Pada Gambar 7, hasil evaluasi menunjukkan bahwa *Silhouette Score* adalah 0,67. Dalam nilai ini hasil pengelompokan data memiliki kualitas yang baik, hasil yang diperoleh dari model pengelompokan yang digunakan layak sebagai dasar untuk analisis lebih lanjut sebagai pengelompokan data dan segmentasi produk.

b. Validasi hasil *clustering*

Hasil analisis yang diperoleh *cluster* optimal adalah 3 menggunakan *cluster* yang terkandung dalam gambar tabel di bawah ini.

```
cluster
0      56
1      23
2       1
dtype: int64
```

Gambar 8. Hasil *Cluster k-means clustering*

Gambar 8, menunjukkan hasil pengelompokan yang dicapai dengan fitur *Google Colab*. Algoritma *K-Means* memiliki *item* ke dalam 3 kategori diantaranya *Cluster 0* berisi 56 *item*, *cluster 1* memiliki 23 *item*, dan *cluster 2* hanya memiliki 1 *item*.

KESIMPULAN

Menggunakan metode pengelompokan *K-Means* pada data transaksi *Kaggle*, studi ini secara efektif mengkategorikan produk menjadi tiga kelompok penjualan tinggi, sedang, dan rendah. Hasil dari metode pengelompokan mengungkapkan bahwa *cluster 0* memiliki 56 *item*, *cluster 1* memiliki 23 *item*, dan *cluster 2* hanya memiliki 1 *item*, menunjukkan bahwa *item* dalam *cluster 2* memiliki evaluasi paling unik atau *ekstrem* menggunakan *Silhouette Score* 0,67 sehingga metode *K-Means* terbukti dapat membantu pemilik bisnis memahami kinerja produk, mengoptimalkan stok, dan mengembangkan strategi pemasaran yang lebih efektif.

DAFTAR PUSTAKA

Awalina, E. F. L., & Rahayu, W. I. (2023). **Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada**

- Transaksi Online Retail.** *Jurnal Teknologi Dan Informasi*, 13(2), 122–137.
<https://doi.org/10.34010/jati.v13i2.10090>
- Azzahra, L., & Amru Yasir. (2024). **Metode K-Means Clustering Dalam Pengelompokan Penjualan Produk Frozen Food.** *Jurnal Ilmu Komputer Dan Sistem Informasi*, 3(1), 1–10. <https://doi.org/10.70340/jirsi.v3i1.88>
- Deden, I, S., Shofia Hilabi, S., Huda, B., & Hananto, A. L. (2024). Website: <https://j-innovative.org/index.php/Innovative> **Implementasi Algoritma K-Means Untuk Klasterisasi Data Stunting.** *INNOVATIVE: Journal Of Social Science Research*, 4, 363–373. <https://j-innovative.org/index.php/Innovative>
- Hananto, A., Pramono, E., & Huda, B. (2022). **Application Of Recapitulation And Staff Performance Assessment Using Standard Working Method.** *Buana Information Technology and Computer Sciences (BIT and CS)*, 3(1), 5–10. <https://doi.org/10.36805/bit-cs.v3i1.2047>
- Hidayat, R., & Kusniyati, H. (2022). **Analisis Clustering Dalam Pengelompokan Penjualan Menggunakan Algoritma K-Means Pada Cafe 47°Coffee Clustering Analysis in Sales Grouping Using The K-Means Algorithm at Cafe 47°Coffee.** 7(2), 420–434. www.jurnal.unimed.ac.id
- Hilabi, S. S. (2017). **Rancang Bangun Situs Responsif Di Universitas Buana Perjuangan Karawang Dengan Menggunakan Metode Perpaduan Grid System Dan Css Media Query.** *Techno Xplore: Jurnal Ilmu Komputer Dan Teknologi Informasi*, 2(1). <https://doi.org/10.36805/technoexplo.v2i1.220>
- Kamalludin, A., & Budianto, et al. (2023). **Penerapan Algoritme Klasterisasi K-Means Kinerja Produk Dengan Analisis Recency Frequency Monetary Pada Cafe Xyz.** *Prosiding Seminar ...*, 2(April), 258–266. <http://senafti.budiluhur.ac.id/index.php/senafti/article/view/631%0A>
- Lia Hananto, A., Assiroj, P., Priyatna, B., Nurhayati, Fauzi, A., Yuniar Rahman, A., & Shofiah Hilabi, S. (2021). **Analysis of Drug Data Mining with Clustering Technique Using K-Means Algorithm.** *Journal of Physics: Conference Series*, 1908(1). <https://doi.org/10.1088/1742-6596/1908/1/012024>
- Nur Afidah, N. (2023). **Penerapan Metode Clustering dengan Algoritma K-means untuk Pengelompokan Data Migrasi Penduduk Tiap Kecamatan di Kabupaten Rembang.** *PRISMA, Prosiding Seminar Nasional Matematika*, 6, 729–738. <https://journal.unnes.ac.id/sju/index.php/prisma/>
- Prasetyani, S. D., & Rochmawati, N. (2022). **Penerapan Data Mining Untuk Clustering Menu Favorit Menggunakan Algoritma K-Means (Studi Kasus Kedai Expo).** *Journal of Informatics and Computer Science (JINACS)*, 3(03), 278–286. <https://doi.org/10.26740/jinacs.v3n03.p278-286>
- Prasetyo, D., Lestati, W., Atina, V., Bangsa, U. D., Surakarta, K., Informasi, J. S., Ilmu, F., Universitas, K., & Bangsa, D. (2024). **Penerapan Clustering Dengan K-Means Untuk Pemilihan Menu Favorit Di Tetra Coffeeshop.** 11(3).
- Priyatman, H., Sajid, F., & Haldivany, D. (2019). **Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa.** *Jurnal Edukasi*

Dan Penelitian Informatika (JEPIN), 5(1), 62. <https://doi.org/10.26418/jp.v5i1.29611>

Priyatna, B. (2019). *Accounting Information System*. 17–30.

Siahaan, A. S., Saragih, R., Simanjuntak, M., & Binjai, S. K. (2024). ***Penerapan Metode K-Means Clustering untuk Pengelompokan Minat Konsumen terhadap Pengguna Jasa Layanan pada Kantor Pos Binjai***. 2(5).

Tri Cahaya, D., Puspita, D., & Syahri, R. (2024). ***Penerapan Metode K-Means Clustering Untuk Pengelompokan Potensi Padi Di Kota Pagar Alam***. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(2), 2187–2193. <https://doi.org/10.36040/jati.v8i2.9432>