LAPORAN PROJEK

Klasifikasi & Analisis Perilaku Pengunjung Mall Dengan Hierarchical Clustering Berdasarkan Karakteristik Kehidupan Untuk Perencanaan Marketing

Kelompok 5

Kelas D

Disusun Oleh:

Bintang Muhammad Ramdhan – 82200

Christopher Kenneth David - 79275

Ian Pangeswara - 55622

Valen Claudia Chuardi – 71430

INFORMATION SYSTEM STUDY PROGRAM

FACULTY OF ENGINEERING AND INFORMATICS

UNIVERSITAS MULTIMEDIA NUSANTARA

TANGERANG

2024

DAFTAR ISI

Contents

DAFTAR GAMBAR

JUDUL: Klasifikasi & Analisis Perilaku Pengunjung Mall Dengan Hierarchical Clustering Berdasarkan Karakteristik Kehidupan Untuk Perencanaan Marketing

1. Latar Belakang

Dalam membangun suatu bisnis, bisnis harus mencari tahu bagaimana cara menarik calon pelanggan untuk berbelanja di bisnis tersebut. Untuk itu, bisnis perlu menyusun strategi bisnis untuk menarik calon pelanggan dari berbagai latar belakang. Perilaku pengunjung mall merupakan kunci dalam strategi marketing yang efektif. Memahami bagaimana pengunjung berperilaku, apa yang mereka inginkan, dan apa yang mendorong mereka untuk berbelanja dapat membantu mall meningkatkan keuntungan dan kepuasan. Namun, pihak mall tidak bisa menanyakan begitu saja apa yang customer inginkan, karena akan sangat memakan waktu untuk mendapatkan kesimpulannya.

Hierarchical clustering adalah metode statistik yang digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik. Dalam konteks ini, hierarchical clustering dapat digunakan untuk mengelompokkan pengunjung mall berdasarkan karakteristik kehidupan mereka. Hal paling mudah dan bisa dipahami dari pengunjung adalah usia, jenis kelamin, seberapa sering mereka berbelanja, pekerjaan, gaji, dan apakah mereka sudah berkeluarga atau belum, karena setiap pengunjung mall memiliki karakteristik berbeda yang menentukan perilaku mereka di mall.

2. Perumusan Masalah

1) Bagaimana membangun strategi bisnis untuk menarik pelanggan dari berbagai kelompok ekonomi?
2) Bagaimana cara untuk membandingkan tiap kelompok ekonomi pelanggan

3) Bagaimana cara mengidentifikasi minat belanja kelompok pelanggan?

3. Batasan Masalah

Penelitian ini memiliki batasan yang tidak bisa diatasi sebagai berikut:

1) Terpaku pada data pelanggan yang harus diambil seiring waktu dalam bentuk tabel jika ingin analisa seiring waktu.
2) Tidak bisa setiap saat memahami perilaku pelanggan atau *real-time*. Sehingga butuh data yang sudah diambil dari waktu lawas.

4. Tujuan Penelitian

Penelitian ini memiliki beberapa tujuan sebagai berikut:

1) Melakukan klasifikasi dan analisa perilaku pengunjung mall berdasarkan karakteristik kehidupan.
2) Mengimplementasikan Agglomerative Clustering untuk beberapa feature data.
3) Melakukan analisis berdasarkan output yang dihasilkan Agglomerative Clustering.

5. Manfaat Penelitian

Apabila tujuan penelitian berhasil terpenuhi, maka manfaat yang didapat adalah sebagai berikut:

1) Dapat mengembangkan strategi marketing yang lebih terstruktur untuk setiap kelompok pengunjung.
2) Meningkatkan kepuasan dan loyalitas pengunjung.

3) Meningkatkan keuntungan perusahaan dan vendor.

6. Telaah Literatur

1. Hierarchical Clustering

Hierarchical Clustering adalah algoritma pengelompokan data yang membangun hierarki cluster secara bertahap. Algoritma ini dimulai dengan menganggap setiap data sebagai cluster individual, kemudian secara berulang menggabungkan cluster-cluster yang paling mirip hingga semua data tergabung dalam satu cluster. Ada pendekatan utama dalam Hierarchical Clustering, pendekatan yang kami lakukan menggunakan Agglomerative (Bottom-Up). Pendekatan ini dimulai dengan menganggap setiap data sebagai cluster individual dan secara iteratif menggabungkan dua cluster yang paling mirip satu sama lain hingga semua data tergabung dalam satu cluster besar. Proses penggabungan ini dilakukan berdasarkan jarak antara cluster, yang bisa diukur dengan berbagai metrik seperti jarak Euclidean, jarak Manhattan, atau koefisien korelasi. Metode penggabungan yang umum digunakan termasuk Complete Linkage, Average Linkage, Ward's Method, dan Single Linkage. Rumus Hierarchical Clustering tergantung pada metode penggabungan yang digunakan. Berikut rumus untuk beberapa metode:

1) Complete Linkage:

$$d(C_1, C_2) = \max\{d(x, y) \mid x \in C_1, y \in C_2\}$$

Dimana:

a) $d(C_1, C_2)$ adalah jarak antar cluster C_1 dan C_2

b) $d(x, y)$ adalah jarak antara data x dan y

2) Average Linkage:

$$d(C_1, C_2) = 1/|C_1||C_2| \; \Sigma\_(x \in C_1) \; \Sigma\_(y \in C_2) \; d(x, y)$$

Dimana:

a) $d(C_1, C_2)$ adalah jarak antar cluster C_1 dan C_2

b) $|C_1|$ dan $|C_2|$ adalah jumlah data dalam cluster C_1 dan C_2

c) $d(x, y)$ adalah jarak antara data x dan y

3) Ward's Method:

$$d(C_1, C_2) = \Delta SSE(C_1, C_2)$$

$$\Delta SSE(C_1, C_2) = SSE(C_1) + SSE(C_2) - SSE(C_1 \cup C_2)$$

Dimana:

a) $d(C_1, C_2)$ adalah jarak antar cluster C_1 dan C_2

b) $\Delta SSE(C_1, C_2)$ adalah perubahan total squared error (SSE) yang dihasilkan dari penggabungan cluster C_1 dan C_2

c) $SSE(C_1)$ dan $SSE(C_2)$ adalah SSE cluster C_1 dan C_2

d) $SSE(C_1 \cup C_2)$ adalah SSE cluster $C_1 \cup C_2$

4) Single Linkage:

$$d(C_1, C_2) = \min\{d(x, y) \mid x \in C_1, y \in C_2\}$$

Dimana:

a) $d(C_1, C_2)$ adalah jarak antar cluster C_1 dan C_2

b) $d(x, y)$ adalah jarak antara data x dan y

Ada Pula kelebihan dan kekurangan dalam menggunakan algoritma Hierarchical Clustering:

A. Kelebihan Hierarchical Clustering:

1) Kemampuan Menangani Data Tanpa Label: Tidak memerlukan label kelas sebelumnya dalam data.

2) Pembentukan Struktur Hierarkis: Membentuk struktur hierarkis dalam data, memudahkan pemahaman yang lebih baik tentang hubungan antar cluster.
3) Fleksibilitas dalam Menangani Struktur Cluster yang Beragam: Mampu menangani berbagai bentuk struktur cluster.

B. Kekurangan Hierarchical Clustering:
1) Keterbatasan Skalabilitas: Tidak efisien untuk data yang sangat besar karena kompleksitas perhitungan yang memakan waktu.
2) Sensitivitas terhadap Noise: Rentan terhadap noise dalam data, yang dapat mempengaruhi pembentukan cluster yang stabil.

7. Metodologi Penelitian

1) Flowchart (General)

a) Stage 1: Strategi Pemasaran Tradisional (Traditional Marketing Strategy)

Pada tahap ini, merujuk kepada metode pemasaran konvensional, seperti iklan cetak, iklan televisi, atau iklan radio.

b) Stage 2: Pengumpulan Data

Tahap ini melibatkan pengumpulan informasi tentang audiens target dan pasar. Data ini dapat dikumpulkan melalui survei, riset pasar, atau sistem manajemen hubungan pelanggan (CRM).

c) Stage 3: Analisis Karakteristik Pelanggan

Tahap ini melibatkan analisis data yang dikumpulkan pada tahap sebelumnya untuk mengidentifikasi demografi, minat, dan kebutuhan audiens target.

d) Stage 4: Penentuan Strategi Pemasaran/ Penjualan Baru

Berdasarkan analisis karakteristik pelanggan, tahap ini melibatkan pengembangan strategi pemasaran atau penjualan baru untuk

mencapai audiens target dan mencapai tujuan pemasaran yang diinginkan.

e) Stage 5: Implementasi Sistem Kecerdasan Buatan (AI System Implementation)

Tahap ini dapat merujuk pada implementasi alat kecerdasan buatan (AI) atau perangkat lunak otomatisasi pemasaran untuk mendukung strategi pemasaran baru.

f) Stage 6: Output Prediksi Sistem dan Pengelompokan Pelanggan

Tahap ini merujuk pada penggunaan AI atau perangkat lunak otomatisasi pemasaran untuk menghasilkan wawasan pelanggan dan mengelompokkan basis pelanggan ke dalam kelompok dengan karakteristik yang serupa. Ini dapat digunakan untuk personalisasi kampanye pemasaran dan menargetkan pelanggan yang tepat dengan pesan yang tepat.

2) Flowchart (System)

a) Stage 1: Data Collection

Tahap ini merupakan proses pengumpulan data dari berbagai sumber. Dalam penelitian ini, data pelanggan diambil dari dataset Kaggle.

b) Stage 2: Data Pre-Processing

Tahap ini melibatkan pembersihan dan persiapan data yang dikumpulkan untuk analisis. Langkah-langkah ini meliputi:

i) Data Pre-Processing: Drop

Tahap ini melibatkan pembersihan dan persiapan data yang dikumpulkan untuk analisis. Ini merujuk pada identifikandan penghapusan data yang tidak relevan atau tidak dapat digunakan dari dataset. Hal ini dilakukan untuk berbagai alasan, seperti meningkatkan kualitas data atau memfokuskan analisis pada aspek-aspek tertentu.

ii) Data Pre-Processing: Normalization

Tahap ini merupakan proses penskalaan atau transformasi data ke rentang yang umum. Hal ini penting untuk algoritma machine learning yang sensitif terhadap skala data.

c) Stage 3: Plot Dendrogram

Tahap ini melibatkan plotting dendogram untuk menentukan jumlah cluster yang digunakan dalam analisis clustering. Menggunakan metric 'euclidean' dan 'n_neighbors', dendogram divisualisasikan untuk membantu menentukan jumlah cluster yang optimal.

d) Stage 4: Agglomerative Clustering

Proses clustering menggunakan algoritma Agglomerative Clustering dilakukan untuk mengelompokkan data ke dalam cluster yang telah ditentukan berdasarkan dendogram.

e) Stage 5: Training for Silhoutte Score

Model yang telah dibentuk dievaluasi menggunakan silhouette score untuk menentukan seberapa baik clustering yang dilakukan. Silhouette score dihitung untuk memastikan bahwa cluster yang dihasilkan memiliki kualitas yang baik.

f) Stage 6: Clustering Visualization

Visualisasi clustering dilakukan untuk memberikan gambaran mengenai pembagian cluster dalam data. Visualisasi ini membantu memahami struktur data dan validitas cluster yang terbentuk.

g) Stage 7: KNN Algorithm

Setelah clustering, algoritma K-Nearest Neighbors (KNN) digunakan untuk klasifikasi data. KNN adalah algoritma supervised yang mengklasifikasikan titik data berdasarkan label dari tetangga terdekatnya dalam data pelatihan.

h) Stage 8: Training & Testing

Data dibagi menjadi set pelatihan dan set pengujian. Model KNN dilatih menggunakan set pelatihan dan diuji pada set pengujian untuk mengevaluasi kinerjanya.

i) Stage 9: Model Evaluation

Tahap terakhir adalah evaluasi kinerja model yang telah dilatih. Metrik yang digunakan meliputi akurasi, precision, recall, dan F1-score, yang memberikan gambaran mengenai efektivitas model dalam klasifikasi data yang baru.

8. Eksplorasi Dataset Penelitian

Penelitian ini melakukan Exploratory Data Analysis (EDA) terhadap dataset yang digunakan. Dataset yang digunakan dalam penelitian ini adalah https://www.kaggle.com/datasets/datascientistanna/customers-dataset , dataset ini diperoleh dari Kaggle.com, sebuah platform yang menyediakan berbagai dataset untuk analisis data. Awalnya, dataset terdiri dari 2000 baris dan 8 kolom, mencakup beragam atribut tentang perilaku dan karakteristik pelanggan.

[2]:

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	Male	19	15000	39	Healthcare	1	4
1	2	Male	21	35000	81	Engineer	3	3
2	3	Female	20	86000	6	Engineer	1	1
3	4	Female	23	59000	77	Lawyer	0	2
4	5	Female	31	38000	40	Entertainment	2	6
...	...	...	...	...	...	...	...	...
1995	1996	Female	71	184387	40	Artist	8	7
1996	1997	Female	91	73158	32	Doctor	7	7
1997	1998	Male	87	90961	14	Healthcare	9	2
1998	1999	Male	77	182109	4	Executive	7	2
1999	2000	Male	90	110610	52	Entertainment	5	2

2000 rows × 8 columns

Gambar 1: Dataset awal customer

Dataset dimodifikasi dalam rangka memfokuskan analisis, semua kolom kecuali 'Annual Income ($)' dan 'Spending Score (1-100)' dihapus. Dataset kemudian dinormalisasi untuk mempersiapkannya untuk analisis lebih lanjut.

Dengan demikian, dataset yang digunakan dalam penelitian ini terdiri dari 2000 baris dan 2 kolom, yang akan menjadi fokus utama dari eksplorasi dan analisis yang akan dilakukan.

[4]:

	Annual Income ($)	Spending Score (1-100)
0	0.078958	0.39
1	0.184236	0.81
2	0.452694	0.06
3	0.310569	0.77
4	0.200027	0.40
...	...	...
1995	0.970591	0.40
1996	0.385095	0.32
1997	0.478808	0.14
1998	0.958600	0.04
1999	0.582238	0.52

2000 rows × 2 columns

Gambar 2: Dataset setelah drop dan normalisasi

DAFTAR PUSTAKA

[1] Alfian, G., Octava, M. Q. H., Hilmy, F. M., Nurhaliza, R. A., Saputra, Y. M., Putri, D. G. P., Syahrian, F., Fitriyani, N. L., Atmaji, F. T. D., Farooq, U., Nguyen, D. T., & Syafrudin, M. (2023). Customer shopping behavior analysis using RFID and machine learning models. Information, 14(10), 551. https://doi.org/10.3390/info14100551

[2] Jiang, H., He, M., Xi, Y., & Zeng, J. (2021). Machine-Learning-Based user position prediction and behavior analysis for location services. Information, 12(5), 180. https://doi.org/10.3390/info12050180

[3] Pradana, M. (2021). Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering. Journal of Applied Data Sciences, 2(1). https://doi.org/10.47738/jads.v2i1.18

[4] Du H, Yu Z, Guo B, Han Q and Chen C. (2020). GroupShop: monitoring group shopping behavior in real world using mobile devices. Journal of Ambient Intelligence and Humanized Computing. 10.1007/s12652-019-01673-9. 14:5. (6367-6378). Online publication date: 1-May-2023. https://link.springer.com/10.1007/s12652-019-01673-9

[5] Machine learning-based intelligent recommendation in virtual mall. (n.d.). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/138224

[6] Werner, E., Clark, J. N., Hepburn, A., Bhamber, R. S., Ambler, M., Bourdeaux, C. P., McWilliams, C. J., & Santos-Rodriguez, R. (2023). Explainable hierarchical clustering for patient subtyping and risk prediction. Experimental Biology and Medicine. https://doi.org/10.1177/15353702231214253

[7] Li, J., Pan, S., Huang, L., & Zhu*, X. (2019). A machine learning based method for customer behavior prediction. Tehnički Vjesnik/TehničKi Vjesnik, 26(6). https://doi.org/10.17559/tv-20190603165825

[8] Setiyani, R. (2014). Artikel ilmiah PDF. Universitasjenderalsoedirman. https://www.academia.edu/5512773/Artikel_ilmiah_PDF
[9] Torrens P. (2022). Agent models of customer journeys on retail high streets. Journal of Economic Interaction and Coordination. 10.1007/s11403-022-00350-z. 18:1. (87-128). Online publication date: 1-Jan-2023. https://link.springer.com/10.1007/s11403-022-00350-z
[10] Wang P, Guo B, Wang Z and Yu Z. ShopSense:Customer Localization in Multi-Person Scenario With Passive RFID Tags. IEEE Transactions on Mobile Computing. 10.1109/TMC.2020.3029833. 21:5. (1812-1828). https://ieeexplore.ieee.org/document/9219250/
[11] Bermejo C, Chatzopoulos D and Hui P. EyeShopper. Proceedings of the 28th ACM International Conference on Multimedia. (2765-2774). https://doi.org/10.1145/3394171.3413683
[12] Guo B, Liu Y, Ouyang Y, Zheng V, Zhang D and Yu Z. Harnessing the Power of the General Public for Crowdsourced Business Intelligence: A Survey. IEEE Access. 10.1109/ACCESS.2019.2901027. 7. (26606-26630). https://ieeexplore.ieee.org/document/8649614/
[13] Guo B, Wang Z, Wang P, Xin T, Zhang D and Yu Z. DeepStore: Understanding Customer Behaviors in Unmanned Stores. IT Professional. 10.1109/MITP.2019.2928272. 22:3. (55-63). https://ieeexplore.ieee.org/document/9098002/
[14] Zhou Z, Shangguan L, Zheng X, Yang L and Liu Y. (2017). Design and Implementation of an RFID-Based Customer Shopping Behavior Mining System. IEEE/ACM Transactions on Networking. 25:4. (2405-2418). Online publication date: 1-Aug-2017. https://doi.org/10.1109/TNET.2017.2689063
[15] Jiao M, Chen X, Su Z and Chen X. (2016). Research on personalized recommendation optimization of E-commerce system based on customer trade behaviour data 2016 Chinese Control and Decision Conference (CCDC). 10.1109/CCDC.2016.7532169. 978-1-4673-9714-8. (6506-6511). http://ieeexplore.ieee.org/document/7532169/