 ISSN: 2684-8473

Pengenalan Pola Untuk Mengetahui Jumlah Target Pengunjung Mall Berdasarkan Usia, Gender, Pendapatan Pertahun, Pengeluaran, Tujuannya Untuk Mempermudah Mengetahui Target Pasar Menggunakan Metode EDA, K-Means, Hierarchial Clustering, Confusion Matrix

Daffa Setiawan Suparno

Universitas Ahmad Dahlan Yogyakarta, Address, Yogyakarta 55161, Indonesia
daffa1700018027@webmail.uad.ac.id

INFORMASI ARTIKEL

Histori Artikel
Diterima : 25 September 2020
Direvisi : 14 Mei 2021
Diterbitkan : 14 Agustus 2021

Kata Kunci:
Mall
Targetpasarmall
EDAmall
K-meansMall
Confusionmatrix

ABSTRAK

Mall merupakan jenis dari pusat perbelanjaan yang memiliki banyak pengunjung tiap harinya dengan jumlah pengunjung yang banyak maka dibutuhkan sebuah sistem untuk mengetahui target pasar dari investor investor toko yang berada di mall.Oleh karena itu pada makalah ini akan membahas tentang pembuatan sistem yang mengetahui jumlah targetpengunjung mall berdasarkan usia, gender, pendapatan pertahun pengunjung, serta pengeluaran mereka selama berbelanja. Tujuannya adalah untuk mempermudah mengetahui target pengunjung yang sering berbelanja di mall. Menggunakan metode EDA sebagai analisis awal data kemudian K-means dan *Hierarchial Clustering* sebagai pengelompokan data, dan *Confusion matrix* pengujian sistem.

I. Pendahuluan

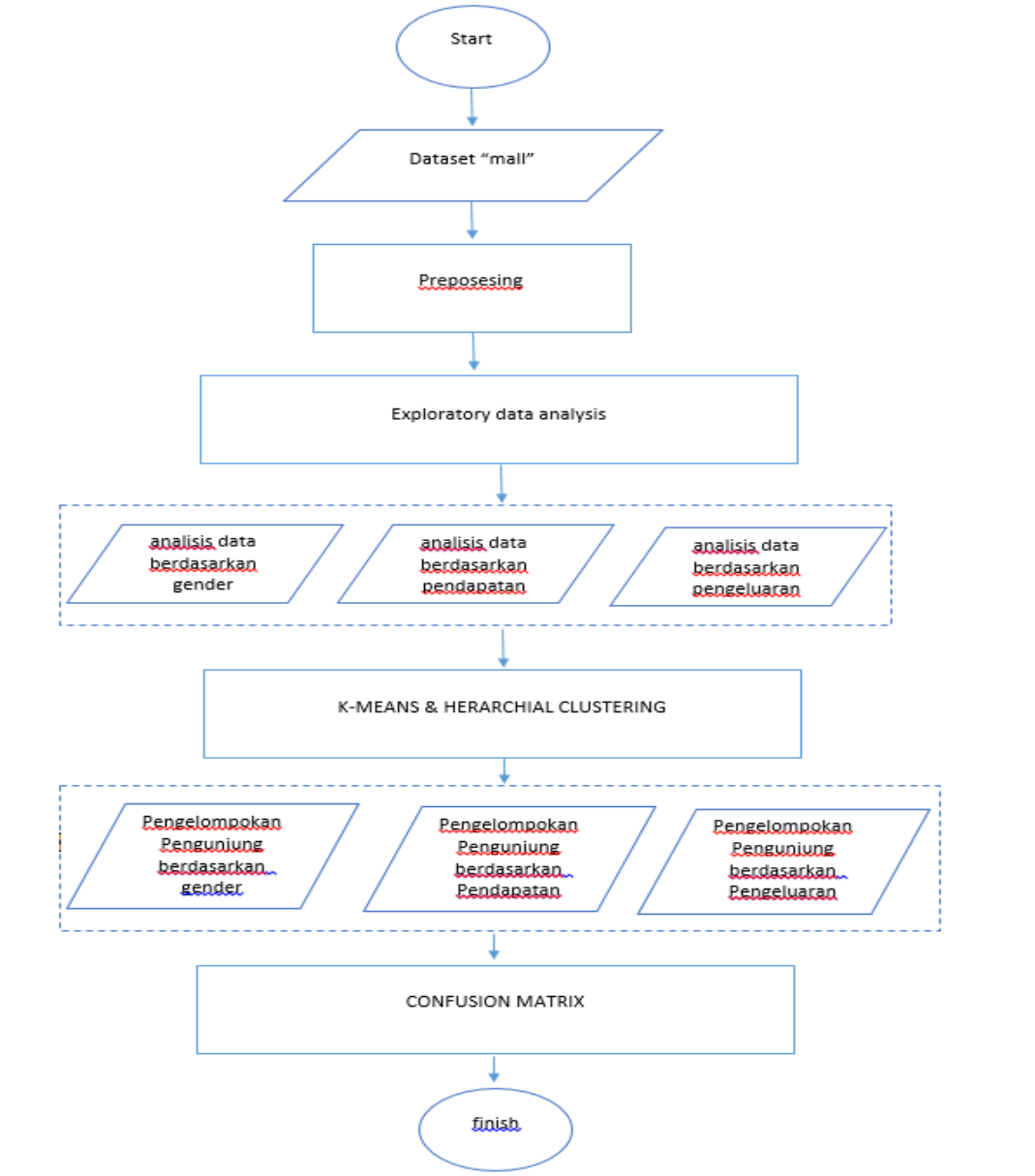
Mall merupakan suatu jenis pusat perbelanjaan yang memiliki toko toko kecil didalamnya yang memasarkan barangnya, banyak orang yang pergi ke mall bukan hanya untuk berbelanja adapula pengunjung yang sekedar hanya melihat lihat dan jalan jalan saja didalam mall, contohnya mall Ambarukmo Yogyakarta pada lebaran tahun 2018 memiliki jumlah pengunjung sebanyak 50.000 orang hal ini tentunya berdampak baik baik perekonomian infestor yang berada di mall. Permasalahan yang sering dihadapi adalah bagaimana cara seorang infestor mengetahui target pasarnya, untuk itu penelitian kali ini menyangkut tentang Pengenalan pola jumlah pengunjung mall berdasarkan gender,usia, dan juga pendapatan pertahunnya, menggunakan metode EDA, K-means, *Hierarchial Clustering*, dan juga Confusion matrix, metode metode tersebut merupakan metode yang sifatnya berkesinambungan metode EDA nantinya digunakan untuk menganalisis atau proses analisis awal data yang nantinya data tersebut akan dikelompokan menggunakan metode K-means dan *Hierarchial Clustering* dan akan diuji dengan *confusion matrix*.

Pengelompokkan pengunjung merupakan salah satu cara yang bisa digunakan untuk mengetahui segment pasar yang datang ke mall. Pengunjung mall tentunya memiliki karakteristik khusus yang dapat dimanfaatkan oleh ahli strategi penjualan/ marketing untuk mendapatkan keuntungan dari banyaknya pengunjung yang datang.. Pengunjung mall biasa datang untuk melakukan berbagai transaksi kegiatan baik itu berbelanja, makan, nongkrong, menonton, atau hanya sekedar mencari hiburan. Sistem ini dibuat untuk mengetahui seberapa banyak pengunjung yang berada di mall yang akan dikelompokan berdasarkan usia, gender, dan jumlah pendapatan, sehingga seroang infestor mall akan dapat mengetahui target pasar mereka menurut usia, gender, dan jumlah pengeluaran terbesar. Pada jurnal ini penulis menggunakan banyak referensi jurnal dalam negeri maupun luar negeri seperti jurnal berjudul "pengelompokan spesies ular berbisa menggunakan metode NLP TF-IDF" jurnal tersebut memiliki tujuan yang sama yaitu mengetahui jumlah sesuatu permasalahan yang ingin diketahui dalam kasus tersebut untuk mengetahui jumlah dan jenis spesies ular berbisa. Dan masih banyak jurnal lainnya yang penulis gunakan sebagai acuan maupun referensi dalam pelaksanaan penelitian.

II. Material dan Metode

A. *Metode Penelitian*

Metode penelitian yang digunakan adalah sistem untuk mengenali pola jumlah pengunjung pada suatu mall. Data yang penelitian gunakan merupakan data yang bersumber dari Kaggle dengan judul atau tema data adalah "mall dataset" yang berbentuk file excel, data tersebut memuat 5 class yaitu : usia, jenis kelamin, jumlah pendapatan pertahun, score pengeluaran pengunjung. Pada dataset tersebut memuat 200 data didalamnya yang akan di uji dan dilakukan pengelompokan dengan tujuan untuk mengetahui jumlah target pasar dengan menggunakan metode EDA, K-means, Hierarchial Clustering, dan nantinya akan di uji dengan metode *Confusion Matrix*. Adapun gambaran umum dari sistem yang akan dibuat seperti berikut ini



Gambar 1. Tahapan Pelaksanaan Penelitian

Mula-mula sistem akan menerima inputan berupa sebuah dataset yang berformat csv atau sejenisnya. Data ini adalah sebuah data yang mengandung beberapa class didalamnya seperti :

a. No
b. Gender
c. Age
d. Annual income
e. Spending score

Kemudian dari data tersebut kita lakukan pemrosesan awal yaitu dengan metode EDA pembacaan data, EDA berguna untuk melakukan pemrosesan data awal untuk menganalisis data sesuai dengan keinginan kita, dalam kasus ini EDA digunakan untuk mengelompokan data awal menjadi beberapa bagian. Kemudaian akan masuk kebagian pengelompokan data menggunakan metode K-means dan Herarchial Clustering, setelah itu data akan diuji menggunakan Confusion matrix.

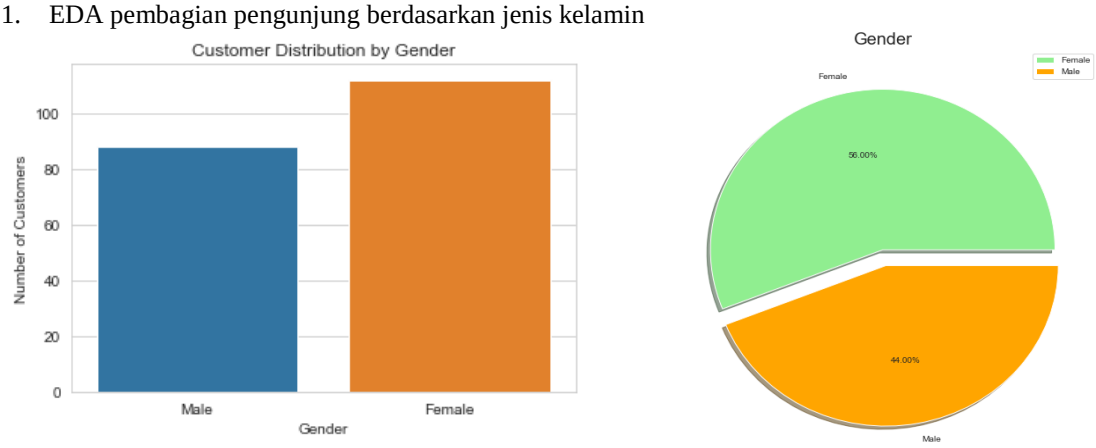
B. *Metode Exploratory Data Analysis (EDA)*

Setelah mendapatkan sebuah data maka kita akan mengelompokan suatu data tersebut sehingga data tersebut dapat kita analisis dengan metode EDA, EDA atau eksplorasi data analisis adalah suatu pendekatan yang digunakan untuk menganalisis data menggunakan berbagai teknik khususnya secara grafis. Tujuannya adalah untuk memaksimalkan wawasan dari data yang ada, mendeteksi outlier dan anomaly data, mengenali struktur data dasar, menguji asumsi data. Sehingga data tersebut dapat kita analisis berdasarkan sebuah plot data, histogram ataupun boxplot, berikut ini merupakan sistem kerja EDA.

Masalah ⟶ Data ⟶ Analisis ⟶ Model ⟶ Kesimpulan

Gambar 2. Tahapan EDA

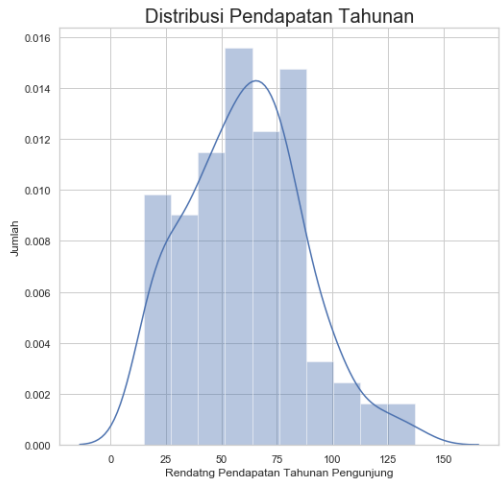
1. EDA pembagian pengunjung berdasarkan jenis kelamin



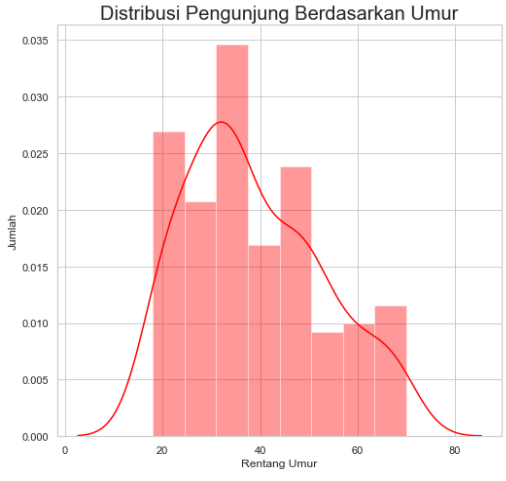
Gambar 3. Pengunjung Berdasarkan Jenis Kelamin

Berdasarkan hasil penambilan data pada gambar 3. menunjukan bahwa pengunjung mall berjenis kelamin wanita lebih mendominasi kunjungan ke mall dengan jumlah pengunjung sebanyak 56% pengunjung wanita dan 44% pengunjung pria

2. EDA distribusi pendapatan tahunan pengunjung dan distribusi pengunjung berdasarkan umur



Gambar 4. Distribusi Pendapatan Tahunan

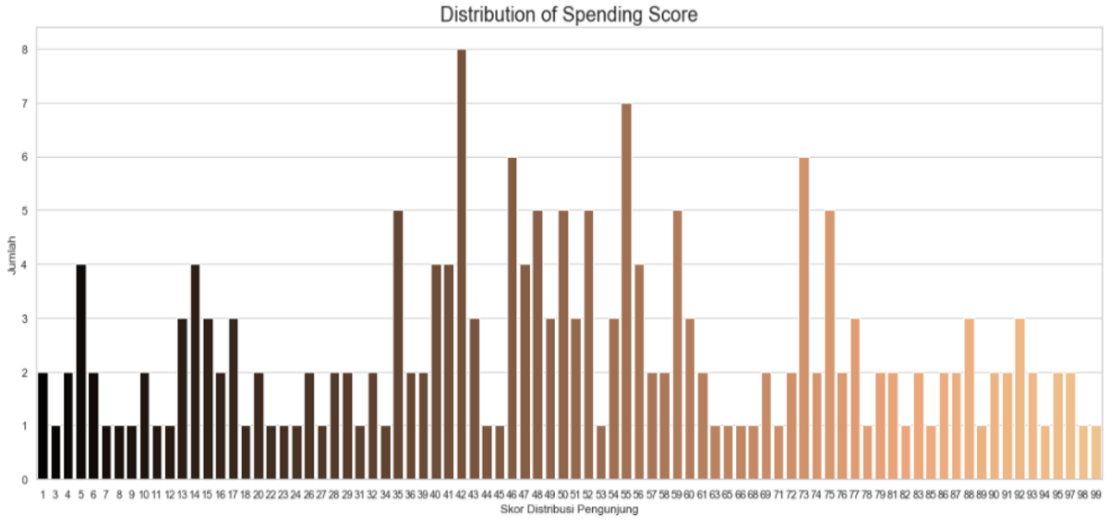


Gambar 5. Distribusi Berdasarkan Umur

Pada gambar 4 dapat menyimpulkan satu hal bahwa ada beberapa orang yang menghasilkan lebih dari 100 Dolar AS. Sebagian besar orang memiliki penghasilan sekitar 50-75 dolar AS, kita dapat menyimpulkan bahwa penghasilan terendah sekitar 20 dolar AS. Pada gambar 5, pelanggan paling sering ke mall berusia sekitar 30-35 tahun. Sedangkan kelompok umur yang paling jarang ke mall adalah pelanggan anak anak dan paruh baya.

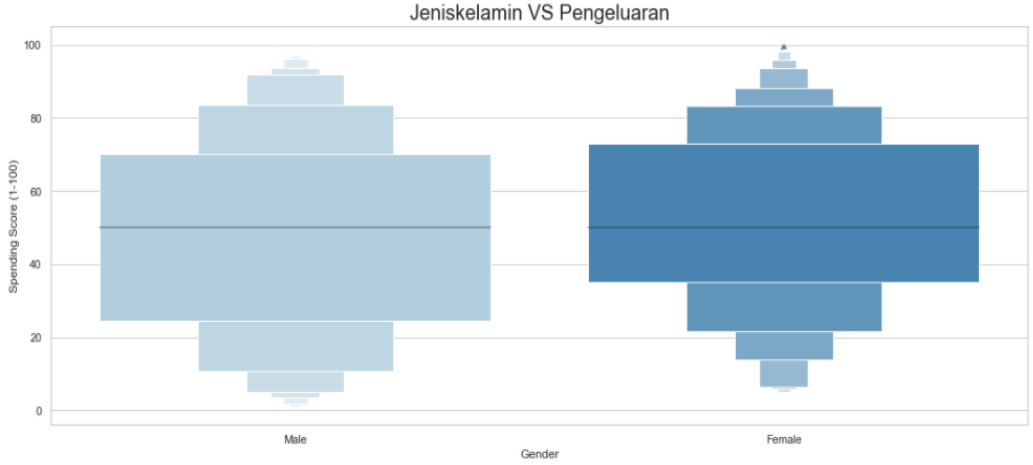
3. EDA distribusi pengeluaran pengunjung

EDA distribusi pengeluaran pengunjung dapat mengetahui seberapa sering pengeluaran pengunjung dan skor pengeluaran tersering berada di angka 35 dan 60 adapula pelanggan yang memiliki skor pengeluaran sebanyak 100 disini menunjukan bahwa pengunjung mall memiliki kalangan dengan jumlah pengeluaran yang bervariatif sesaui kebutuhan pelanggan. Untuk lebih jelasnya distribusi pengeluaran pengunjung dapat dilihat pada gambar 6.



Gambar 6.Distribusi Pengeluaran Pengunjung

4. EDA distribusi jenis kelamin vs pengeluaran



Gambar 7. Distribusi Jenis Kelamin dibandingkan dengan Pengeluaran

Gambar 7 dengan menggunakan boxplot antara gender pengunjung dan juga jumlah pengeluaran per gendernya, terlihat jelas bahwa angka pengeluaran pria hanya berkisar antara 25rb dolar As hingga 70rb dolar As, sedangkan wanita lebih memiliki angka pengeluaran sebesar 35rb hingga 75rb dolar, hal ini membuktikan bahwa target pasar terbesar merupakan kaum wanita.

C. K-Means Clustering

K-means clustering adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervise (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi, untuk memproses data mall diatas data dimuali dengan kelompok pertama centroid yang dipilih secara acak, sehingga digunakan sebagai titik awal untuk setiap clusternya. Adapun tahapan algoritma ini adalah sebagai berikut.

1. Penentuan berapa banyak jumlah cluster
2. Secara acak tentukan record menjadi lokasi pusat cluster
3. Temukan pusat cluster terdekat untuk setiap record. Adapun persamaan yang sering digunakan dalam pemecahan masalah dalam menentukan jarak terdekat adalah persamaan Euclidean berikut :

$$d_{Euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Dimana x=x1,x2,x3……xm dan y=y1,y2,y3…ym, sementara m menyatakan banyaknya nilai atribut dari 2 buah record.

4. Tentukan cluster terdekat untuk setiap data dengan membandingkan nilai jarak terdekat, lalu perbaharui nilai pusat clusternya

$$ClusterCenter = \sum \frac{a_i}{n}$$

III. Hasil dan Pembahasan

A. K-Means Clustering

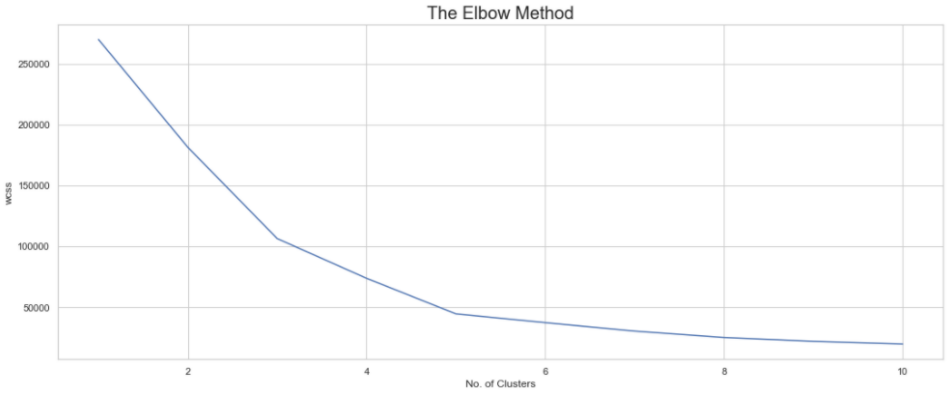
Data pengunjung mall yang telah dikumpulkan kemudian diolah untuk dijadikan informasi dapat dilihat pada tabel 1.

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

Tabel 1. Data Pengunjung Mall

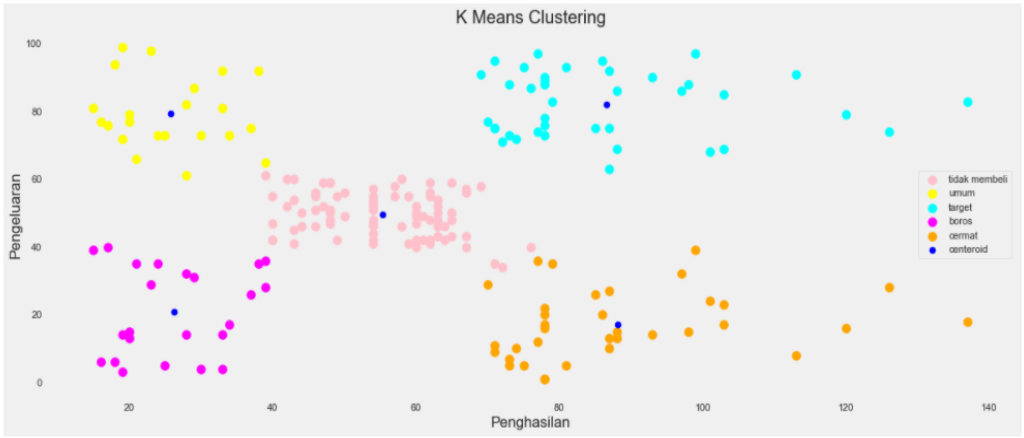
1. Data Penghasilan dan Pengeluaran

Kemudian kita akan membuat *elbow method*, *elbow method* ini digunakan agar kita dapat mengetahui dan dapat memilih cluster mana yang akan kita lakukan pengelompokan data pengunjung mall. Pengelompokan ini dapat dilakukan berdasarkan penghasilan dan juga pengeluaran untuk lebih jelasnya dapat dilihat pada gambar 8.



Gambar 8. Cluster Pengunjung berdasarkan *The Elbow Method*

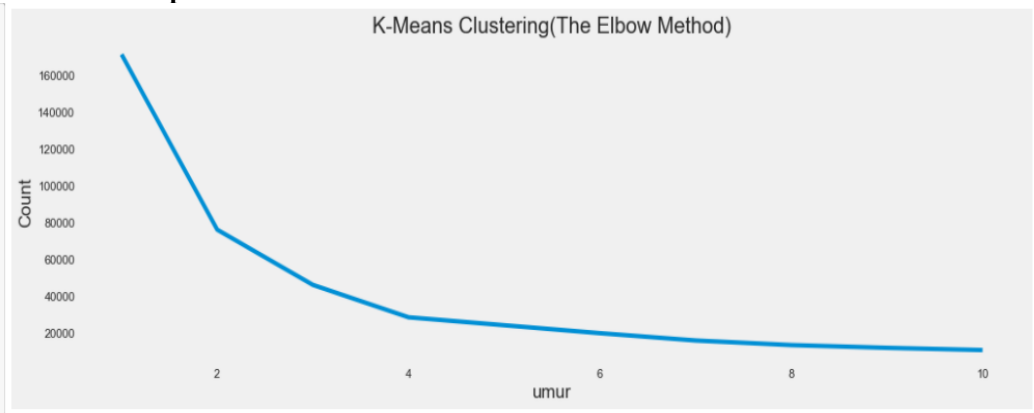
Pada *elbow method* tersebut kita dapat mengetahui cluster berapa yang akan kita kelompokan, dan pada elbow method diatas kita mendapatkan cluster 5 sehingga kita akan membuat K-means clustering untuk cluster 5 tersebut.



Gambar 9. *K-Means Clustering* Berdasarkan pengeluaran

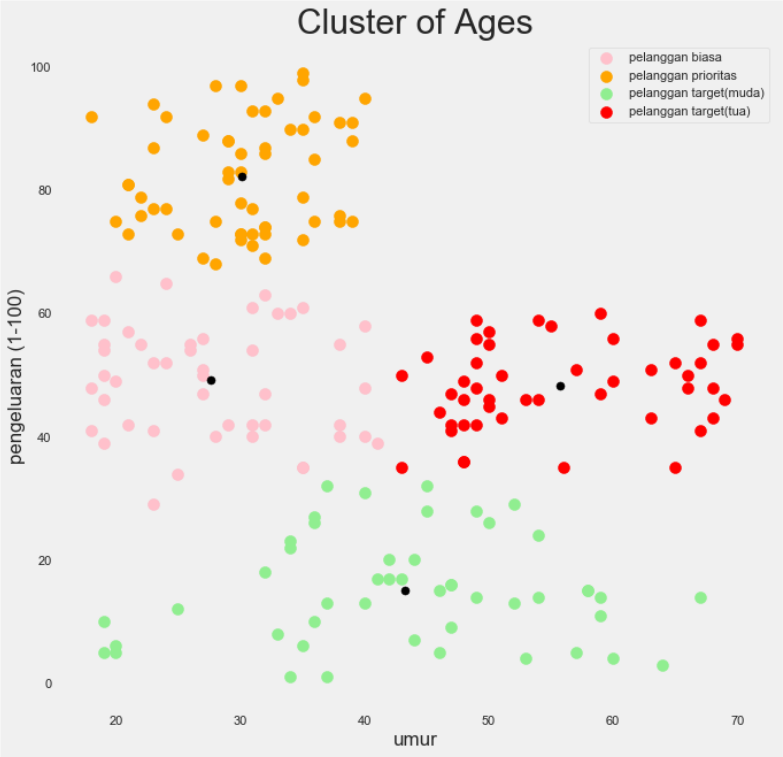
Berdasarkan gambar 9 dari k-means clustering dapat disimpulkan bahwa terdapat berbagai jenis pelanggan yaitu pelanggan yang tidak membeli apapun, pelanggan umum, pelanggan target, pelanggan boros, serta pelannggan yang cermat.

2. Data Kelompok Berdasarkan Usia



Gambar 10. *K-Means Clustering* Berdasarkan Usia

Berdasrakan gambar 10. elbow method tersebut kita dapat mengetahui cluster berapa yang akan kita kelompokan, dan pada elbow method diatas kita mendapatkan cluster 4 sehingga kita akan membuat K-means clustering untuk cluster 4 tersebut.

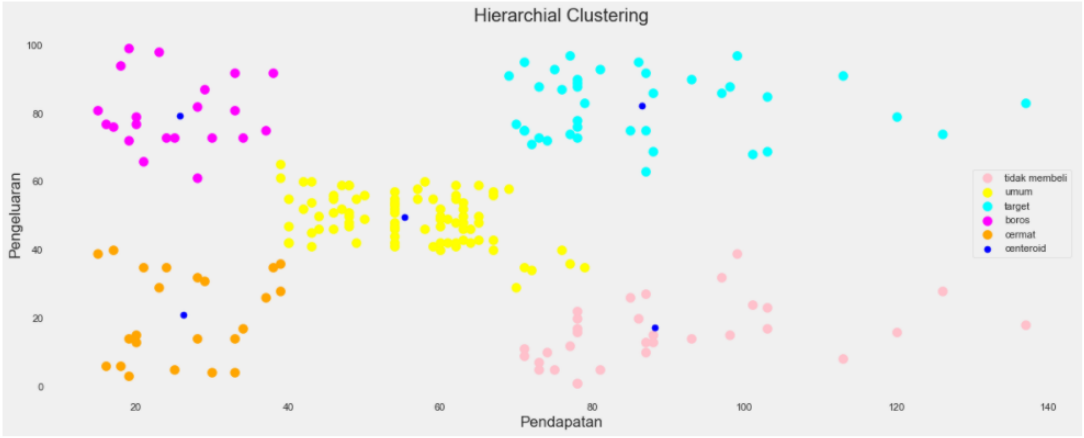


Gambar 11. Cluster pengunjung berdasarkan Usia dan Pengeluaran

Berdasarkan gambar 11 cluster dapat dilihat bahwa terdapat 4 kategori pelanggan yaitu pelanggan biasa, pelanggan prioritas, pelanggan target(muda), pelanggan target (tua). Pelanggan target muda masih mendominasi pengunjung mall.

B. Hierarchial Clustering

Hierarchical clustering, juga dikenal sebagai analisis cluster hirarki, merupakan algortma yang mengelompokan objek yang mirip kedalam kelompok yang disebut sebagai cluster. Titik akhirnya merupakan satuset cluster, dimana setiap cluster berbeda satu sama lainnya, dan objek dalam cluster cera luas mirip satu dengan lainnya.



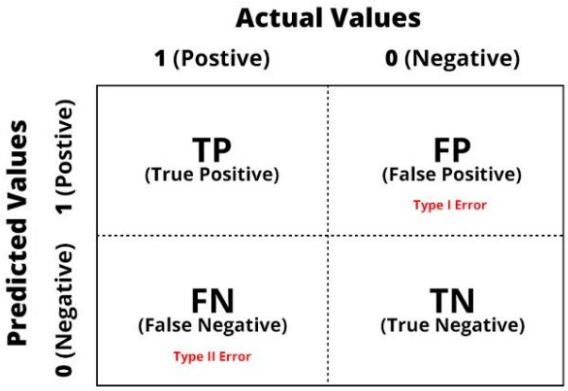
Gambar 12. *Hirarchial Clustering* pengunjung berdasarkan Pengeluaran dan Pendapatan

Dari Gambar 12. *Hierarchial clustering* diatas dapat disimpulkan bahwa terdapat berbagai jenis pelanggan yaitu pelanggan yang tidak membeli apapun, pelanggan umum, pelanggan target, pelanggan boros, serta

pelannggan yang cermat. Data cluster k-means dan hierarchial memiliki data yang sama akan tetapi mempunyai pengelompokan data yang berbeda

C. *Confusion Matrix*

Confusion matix juga sering disebut error matrix. Pada dasarnya *confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan sistem dengan hasil klasifikasi sebenarnya. *Confusion matric* berbentuk table matrix yang menggambarkan kinerja model klasifikasi pada serangkaian data yang diuji yang nilai sebenarnya diketahui.



Gambar 13. *Confusion Matric*

Terdapat 4 istliah sebagai representasi hasil proses klasifikasi pada *confusion matrix*. Keempat istilah tersebut adalah *true positif*(TP), *true negative* (TN), *false positif*(FP), dan *false negative*(FN). Berikut ini penerapannya pada sistem

	precision	recall	f1-score	support
Female	0.58	0.71	0.64	45
Male	0.48	0.34	0.40	35
accuracy			0.55	80
macro avg	0.53	0.53	0.52	80
weighted avg	0.54	0.55	0.53	80

Gambar 14. Penerapan klasifikasi *Confusion Matrix*

Pada gambar 14. menjelaskan bahwa akurasi dari site mini adalah 0.55 yaitu 55 % tentunya hal ini masih jauh dari target yang diharapkan. Tingkat akurasi yang rendah dipengaruhi oleh jumlah data dan juga data yang akan diolah.

IV. Kesimpulan

Hasil yang diperoleh dari pengujian ini, dapat ditarik kesimpulan bahwa sistem mampu melakukan pengenalan pola dan melakukan pengelompokan data dengan metode EDA, K-means, *Hierarchial Clustering* dan *Confusion matrix* sesuai dengan apa yang diharapkan. Hal ini dibuktikan dengan keandalan sistem melakukan plot data dan juga analisis data dari dataset mall yang tersedia, sehingga didapatkannya data target pasar yaitu pengunjung dengan usia muda hal itu dapat dibuktikan dengan jumlah pengeluaran mereka, di no dua target ada pengunjung dewasa dengan jumlah pengeluaran sedang mereka.Kendala yang dihadapi dalam pembuatan dan penelitian kali ini adalah terbatasnya jumlah dataset pengunjung sehingga hasil akurasi kurang memuaskan. Meskipun hasil akurasi masih belum memuaskan akan tetapi penulis tetap akan selalu melakukan pengujian ulang dengan data data lain yang tersedia terutama bagi dataset berjumlah banyak sehingga didapatkan tingkat akurasi yang baik dari data yang tersedia.

Daftar Pustaka

[1] Nur Liyana Izzati Rusli, Amiza Amir, Nik Adilah Hanin Zahri, R. Badlishah Ahmad," (2019) Snake species identification by using natural language processing" Vol. 13, No. 3, March 2019, pp. 999~1006 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i3.pp999-1006

[2] Eleonora Maria Aiello, Chiara Toffanin, Mirko Messori, Claudio Cobelli and Lalo Magni "Postprandial Glucose Regulation via KNN meal classification in Type 1 Diabetes" Citation information: DOI 10.1109/LCSYS.2018.2844179, IEEE Control Systems Letters

[3] Jayme Garcia arnal barbed, Luciano Vieira koeningkan, thiago Teixeira santos "identifying multiple plant diseases using digital image processing"

[4] Mitika Chaudhary, Vinay Prakash and Neeraj Kumari "Identification Vehicle Movement Detection in Forest Area using MFCC and KNN" Proceedings of the SMART–2018, IEEE Conference ID: 44078

[5] Khin Nyein Nyein Hlaing and Anilkumar Kothalil Gopalakrishnan "Myanmar Paper Currency Recognition Using GLCM and k-NN" IEEE 10.1109/ACDT.2016.7437645

[6] Santosh Kumar, Sanjay Kumar Singh1,Ravi Shankar Singh,Amit Kumar Singh, Shrikant Tiwari "Real-time recognition of cattle using animal biometric" Received: 14 June 2016/Accepted: 3 October 2016, DOI 10.1007/s11554-016-0645-4

[7] Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar S. K¨uhl, and Joachim Denzler "Chimpanzee Faces in the Wild: Log-Euclidean CNNs for Predicting Identities and Attributes of Primates" GCPR 2016, LNCS 9796, pp. 51–63, 2016, DOI: 10.1007/978-3-319-45886-1 5

[8] Harry Rubin-Falconea,b, Francesca Zanderigoa, Binod Thapa-Chhetryb, Martin Lana, Jeffrey M. Millera, M. Elizabeth Sublettea, Maria A. Oquendoc, David J. Hellersteina, Patrick J. McGratha, Johnathan W. Stewarta, J. John Manna "Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder" Received 6 September 2017; Received in revised form 6 November 2017; Accepted 11 November 2017, https://doi.org/10.1016/j.jad.2017.11.043

[9] Ronald Julian O'Brien, Juan Manuel Fontana, Nicolás Ponso, Leonardo Molisani "A pattern recognition system based on acoustic signals for fault detection on composite materials" Volume64, July–August 2017,Pages 1-10, https://doi.org/10.1016/j.euromechsol.2017.01.007